JB Accepts, published online ahead of print on 25 May 2012
J. Bacteriol. doi:10.1128/JB.00266-12
Copyright © 2012, American Society for Microbiology. All Rights Reserved.

JB Accepts published online ahead of print

May 9, 2012 sbs

Caldicellulosiruptor core and pan genomes reveal determinants for non-cellulosomal thermophilic deconstruction of plant biomass

Sara E. Blumer-Schuette (seblumer@ncsu.edu)[1, 4], Richard J. Giannone (giannonerj@ornl.gov)[2, 4], Jeffrey V. Zurawski (jvzuraws@ncsu.edu)[1, 4], Inci Ozdemir (iozdemi@ncsu.edu)[1, 4], Qin Ma (maqin2001@gmail.com)[3, 4], Yanbin Yin (yanbin.yin@gmail.com)[3, 4], Ying Xu (xyn@bmb.uga.edu)[3, 4], Irina Kataeva (kataeva@bmb.uga.edu)[3, 4], Farris L. Poole II (fpoole@bmb.uga.edu)[3, 4], Michael W. W. Adams (adams@bmb.uga.edu)[3, 4], Scott D. Hamilton-Brehm (hamiltonbres@ornl.gov)[2, 4], James G. Elkins (elkinsjg@ornl.gov)[2, 4], Frank W. Larimer (larimerfw@ornl.gov)[2], Miriam L. Land (landml@ornl.gov)[2], Loren Hauser (hauserlj@ornl.gov)[2], Robert W. Cottingham (cottinghamrw@ornl.gov)[2, 4], Robert L. Hettich (hettichrl@ornl.gov)[2, 4], and Robert M. Kelly (rmkelly@eos.ncsu.edu)[1, 4*]

[1]Dept. of Chem. and Biomolec. Engr., North Carolina State University, Raleigh, NC, 27695

[2] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831

[3]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, 30602

[4] BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, TN, 37831

Submitted to: ***Journal of Bacteriology (February, 2012)***

Revision submitted to JB00266-12 (May, 2012)

Running title: Cellulose degradation by extremely thermophilic *Caldicellulosiruptor spp.*

Keywords: *Caldicellulosiruptor*, extreme thermophile, plant biomass, glycoside hydrolases, pan-genome, core-genome

*Address correspondence to: **Robert M. Kelly**
Department 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@eos.ncsu.edu

JB Accepts published online ahead of print

ABSTRACT

Extremely thermophilic bacteria of the genus *Caldicellulosiruptor* utilize carbohydrate components of plant cell walls, including cellulose and hemicellulose, facilitated by a diverse set of glycoside hydrolases (GHs). From a biofuels perspective, this capability is crucial for deconstruction of plant biomass into fermentable sugars. While all species from the genus grow on xylan and acid pre-treated switchgrass, growth on crystalline cellulose is variable. The basis for this variability was examined using microbiological, genomic and proteomic analysis of eight globally-diverse *Caldicellulosiruptor* species. The open *Caldicellulosiruptor* pan-genome (4,009 ORFs) encodes 106 GHs, representing 43 GH families, but only 26 GHs from 17 families are included in the core (non-cellulosic) genome (1,543 ORFs). Differentiating the strongly cellulolytic *Caldicellulosiruptor* species from the others is a specific genomic locus that encodes multi-domain cellulases from GH families 9 and 48, associated with cellulose binding modules. This locus also encodes a novel adhesin associated with type IV pili, which was identified in the exo-proteome bound to crystalline cellulose. Taking into account the core, pan and individual genomes, the ancestral *Caldicellulosiruptor* was likely cellulolytic, and evolved, in some cases, into species that lost the ability to degrade crystalline cellulose, while maintaining the capacity to hydrolyze amorphous cellulose and hemicellulose.

JB Accepts published online ahead of print

INTRODUCTION

Interest in cellulosic biofuels (30) has sparked efforts to isolate microorganisms capable of both hydrolysis and fermentation of plant biomass, a process referred to as "consolidated bioprocessing" (CBP) (50, 51). Since plant biomass deconstruction could be accelerated at elevated temperatures, thermophilic microorganisms have been considered as catalysts for CBP (8). Of particular note in this regard are members of the extremely thermophilic genus *Caldicellulosiruptor* that inhabit globally diverse, terrestrial hot springs (12, 28, 57, 58, 62, 70, 81, 98) and thermally-heated mud flats (32). *Caldicellulosiruptor* species are gram-positive bacteria and typically associate with plant debris; consequently, all isolates characterized to date hydrolyze certain complex carbohydrates characteristic of plant cell walls (8, 97). As such, members of the genus *Caldicellulosiruptor* are excellent genetic reservoirs of enzymes for plant biomass degradation and, pending the development of functional genetics systems, potential metabolic hosts for CBP (9).

Currently, there are two main paradigms described for microbial degradation of crystalline cellulose: cellulosomal (3) and non-cellulosomal (49, 55). Enzymatically, both systems require the concerted efforts of cellobiohydrolases, endo-cellulases and β-glucosidases (50). Crystalline cellulose deconstruction via cell membrane-bound cellulosomes was first described in the thermophile *Clostridium thermocellum*, and has since been described in other mesophillic Firmicutes, such as *Clostridium cellulolyticum*, *Acetivibrio cellulolyticus* and *Ruminococcus flavefaciens* (3). Analysis of genome sequence data from biomass-degrading microorganisms has helped to identify non-cellulosomal bacteria that also lack identifiable cellobiohydrolases, such as *Cytophaga hutchinsonii* (96) and *Fibrobacter succinogenes* (78), both which require close attachment to cellulose for efficient hydrolysis and *Sacharophagus degradans* (95) which uses processive endo-cellulases (94) indicating that there is a great diversity in strategies used for crystalline cellulose hydrolysis. As members of the phylum

JB Accepts published online ahead of print

Firmicutes, *Caldicellulosiruptor* species are distinct from the thermophilic, anaerobic clostridia in that they secrete free and S-layer bound cellulases and hemicellulases (9, 24, 25, 44, 45, 59, 61, 64, 76, 85, 89, 90) that are not assembled into cellulosomes (85, 89). In this respect, their strategy for crystalline cellulose deconstruction is similar to non-cellulosomal biomass-degrading aerobic fungi, such as *Trichoderma reesei*, (55) thermophillic fungi *Myceliophthora thermophila* and *Thielavia terrestris* (7) or the thermophilic aerobe *Thermobifida fusca* (49).

The non-cellulosomal strategy used by the genus *Caldicellulosiruptor* for plant biomass deconstruction involves novel, multi-domain, carbohydrate-active enzymes (24, 25, 59, 61, 76, 85, 89). However, while all *Caldicellulosiruptor* species hydrolyze hemicellulose, not all can degrade crystalline cellulose. This gives rise to significant disparity across the genus with respect to the capacity to deconstruct plant cell walls. To date, only limited information is available on the genus *Caldicellulosiruptor*, especially with respect to the diversity within the genus and the characteristics of individual species. Given the variability within the genus for cellulose deconstruction, insights into this important environmental and biotechnological capability could be obtained by comparative examination of weakly to strongly cellulolytic *Caldicellulosiruptor* species. To this end, here we examine the core and pan genome of eight members of this genus, in conjunction with exo-proteomics analysis, to identify common and distinctive determinants that drive plant biomass deconstruction.

MATERIALS AND METHODS

Cultivation of *Caldicellulosiruptor* spp. Seven *Caldicellulosiruptor* species were revived from freeze-dried cultures provided by the German Collection of Microorganisms and Cell Cultures (DSMZ [http://www.dsmz.de]) in the recommended culture medium, after which they were transferred to modified DSMZ medium 640 (trypticase, resazurin, cysteine-HCl, and

JB Accepts published online ahead of print

$FeCl_3$ x 6 H_2O were not added, the reducing agent: 10% [wt/vol] Na_2S x 9 H_2O was added to a final concentration of 0.5% in prepared medium)(9). The eighth species examined in this study, *C. obsidiansis*, was isolated from the Obsidian Pool, Yellowstone National Park (28). Complex substrates used as carbon sources for growth included microcrystalline cellulose (Avicel, PH-101, FMC), birchwood xylan (Sigma), and acid pre-treated switchgrass (*Panicum virgatum* (71)), all added to growth medium at 5 g/L, and in the case of biomass 5 g/L wet weight. Twenty four-hour cell density measurements were the average of two biological replicates in 50 mL cultures. Enumeration of cell densities were conducted under epi-fluorescence microscopy using acridine orange (Kodak) as a fluorescent dye (31). Qualitative rating of cellulose hydrolysis ability was by the organism's ability to shred Whatman No.1 filter paper while grown in Hungate tubes, as described previously (9). Those species capable of shredding filter paper were designated as cellulolytic. Those that were noted to grow on microcrystalline cellulose (Avicel), but not shred filter paper, were designated as weakly cellulolytic.

Genomic DNA isolation and quality control. High-molecular-weight genomic DNA (gDNA) for five *Caldicellulosiruptor* species was harvested as described before (9). Overall, the cultures were grown to early stationary on DSMZ culture medium as recommended by DSMZ (without resazurin): DSMZ medium 671 with cellobiose (*C. hydrothermalis*, *C. kristjanssonii*, *C. kronotskyensis* and *C. lactoaceticus*) and DSMZ medium 144 with glucose (*C. owensensis*). Cultures were harvested by centrifugation at 5000 rpm for 15 minutes and gDNA was isolated, as previously described (21), but with an additional step requiring lysozyme (100 mg/ml) and the final precipitation of gDNA in isopropanol, collected by a flamed glass hook and then gently washed in 70% [vol/vol] ethanol. Dried gDNA was re-suspended in TE buffer to roughly 1 µg/µl and checked for quality on a 1% [wt/vol] agarose gel in 1x TAE buffer. Molecular weight standards and the protocol for assessing gDNA quality using agarose gel electrophoresis were

JB Accepts published online ahead of print

both provided by the DOE Joint Genome Institute (JGI) (http://my.jgi.doe.gov/general/protocols/20100809-Genomic-DNA-QC.doc).

Genome sequencing. The finished genome sequences of *C. bescii* (61, 98), *C. obsidiansis* (18), and *C. saccharolyticus* (89) were completed prior to this project. For *C. hydrothermalis*, *C. kristjanssonii*, *C. kronotskyensis*, *C. lactoaceticus* and *C. owensensis*, a combination of 454 Titanium (52) and Illumina (5) technologies was used (10), similar to the sequencing strategy of *C. obsidiansis*. Detailed protocols explaining these methods can be found at http://www.jgi.doe.gov/.

Genome assembly and annotation. For the five genome sequences that were completed for this project, assembly has been previously described (10). Genes were identified using Prodigal (34) as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using the JGI GenePRIMP pipeline (65). The predicted open reading frames (ORFs) were translated and used to search the National Center for Biotechnology Information (NCBI) nr database (6), UniProt (15), TIGRFAM (27), Pfam (69), PRIAM (14), KEGG (35), COG (84), and InterPro (100) databases. These data sources were combined to assert a product description for each predicted protein. Non-coding genes and miscellaneous features were predicted using tRNAscan-SE (47), RNAmmer (40), Rfam (26), TMHMM (37), and SignalP (v3.0, 4). The *C. saccharolyticus* annotation was updated using the same pipeline except that manual curation was done without GenePRIMP. Further annotation of selected proteins included molecular weight and isoelectric point (pI) prediction (20), signal peptide (SignalP v4.0, 68) and transmembrane (TMHMM, 37) prediction.

Phylogenetic analysis. All three copies of 16S ribosomal RNA genes were used in construction of a phylogenetic tree. ClustalW (87) was used to align 16S sequences from all sequenced *Caldicellulosiruptor* species, plus one copy of a 16S rRNA gene from three distantly

JB Accepts published online ahead of print

related species. Pairwise distance calculations were done using the MEGA 4 software package (83), with the Tajima-Nei substitution model. These distance calculations were then used to construct dendrograms, based on neighbor joining and assessed with 1000 bootstraps. Average nucleotide identity was used to assess relatedness of species taking their whole genome sequence into consideration. All eight sequenced *Caldicellulosiruptor* species and the same three outliers mentioned above were uploaded into the Jspecies package using the ANIb BLASTn option (72). ANI, reported as percent identity, was represented using the cellplot feature of JMP (v9, SAS) to create a heat plot. ANIb percentages can be found in Table S1.

Prediction of orthologous and functional groups of proteins. Using all eight finished genomes, orthologous groups of proteins were predicted by OrthoMCL (43). Parameters selected were a p-value of $1xe^{-5}$, a percent identity cutoff of 0, a percent match cutoff of 0, MCL inflation of 1.5 and weight of 316. OrthoMCL output (Dataset S1), based on protein-protein homology, was used to then compute the core- and pan-genome according to Tettelin (86). Top ranked similarity searches against genomes in the KEGG database (35) used BLASTp (1). Functional classification of proteins was determined based on searched against databases from NCBI (COG) (84), CAZy (13), integrated microbial genomes (IMG) (54) and InterProScan sequence search (100). Prediction of carbohydrate transporters were done as previously described (91) and also utilized the 'Find Functions' database of the IMG portal (54).

Fractionation of substrate-bound, extra-, and intra-cellular proteins. Seven *Caldicellulosiruptor* species, four cellulolytic and three weakly cellulolytic were selected for proteomic analysis. Samples were transferred on Avicel PH-101 three times prior to inoculation of two independent 500 mL cultures each in 1000-ml 45-mm-diameter screw-top pyrex bottles. A starting inoculum of 1×10^6 cells/mL was used for all cultures, and growth proceeded in batch for 24 hours. After 24 hours, biological repeats were combined for processing. Spent Avicel with substrate-bound (SB) proteins was isolated by decanting the supernatant (SN) and cells (WC),

JB Accepts published online ahead of print

and washing the SB fraction twice with ice-cold DSMZ medium 640, following which the medium was decanted and combined with the SN plus WC fraction. Further washing of the SB fraction was done, as described previously (73), with TBS-Ca-T buffer (25 mM Tris-HCl, pH 7.0, 150 mM NaCl, 1 mM $CaCl_2$, and 0.05% [vol/vol] Tween-20). Cell-free SN fraction was obtained by centrifugation at 4°C and 5,000 rpm for 15 min, followed by bottle-top filtration through a 0.22-µm-pore-size filter (Millipore). The resulting WC pellet after centrifugation was washed once with ice-cold DSMZ medium 640 and collected by centrifugation, as described above.

Proteomic measurements of Avicel-induced protein fractions. Each fraction for proteomic analysis (WC, SN, and SB) was independently prepared for bottom-up, two-dimensional liquid chromatography tandem mass spectrometry (2D-LC-MS/MS) in order to retain fractional protein localization. Proteins in each fraction were first isolated and denatured by one of several related methodologies: **A** - Cells in the WC fraction were lysed by a combination of boiling and sonication (Branson Sonifier) in SDS lysis buffer (4% SDS, 100 mM Tris-HCl, 50 mM DTT). Released proteins (2 mg of crude lysate, as measured by BCA assay) were then isolated via 20% TCA and re-suspended in 250 µl 8 M urea, as previously described (22). **B -** SN proteins were concentrated to 1 ml by centrifugal membrane filtration (Vivaspin 20 PES, 5 kDa MW cutoff, GE Healthcare), TCA-precipitated, acetone-washed and re-suspended in 250 µl of 8 M urea. **C -** Proteins bound to Avicel (SB) were first stripped from the spent substrate (~10 ml) with 10 ml SDS lysis buffer plus boiling and sonication. Samples were then centrifuged at 4500 x g and supernatant collected. Proteins in this crude SB fraction were then concentrated, precipitated, washed, and re-suspended, as described in **B** above. Following isolation and denaturation, proteins obtained from each fraction, now in 250 µl of 8 M urea, were reduced (dithiothreitol), alkylated (iodoacetamide), digested (two additions of trypsin), and prepared for 2D-LC-MS/MS analysis, as previously described (22). Peptide concentrations were measured by BCA assay.

JB Accepts published online ahead of print

206 To reveal the protein complement of each fraction, 25-, 50-, or 100-µg peptides (SB, SN,
207 WC respectively) were bomb-loaded onto a biphasic MudPIT back column (56, 93) packed with
208 ~5 cm strong cation exchange (SCX) resin, followed by ~3 cm reversed phase (RP) resin (Luna
209 and Aqua resins, respectively; Phenomenex). Loaded peptides were then washed/desalted
210 offline, placed in-line with an in-house pulled nanospray emitter packed with 15 cm RP resin and
211 analyzed via MudPIT 2D-LC-MS/MS analysis, as previously described (22). Briefly, for WC
212 analysis, a full 24-hr MudPIT was employed (eleven salt cuts at 5, 7.5, 10, 12.5, 15, 17.5, 20,
213 25, 35, 50, 100% of 500 mM ammonium acetate followed by 100-min organic gradient). For
214 both SN and SB peptide fractions, a mini-MudPIT was utilized (four salt cuts at 10, 20, 35, 100%
215 of 500 mM ammonium acetate followed by 100-min organic gradient). Peptide fragmentation
216 data was collected via a hybrid LTQ XL-Orbitrap mass spectrometer (Thermo Fisher Scientific)
217 operating in data-dependent mode. Full MS1 scans (2 µscans; 5 MS/MS per MS1) were
218 obtained using the Orbitrap mass analyzer set to 15K resolution, while MS/MS scans (2 µscans)
219 were obtained/performed in the LTQ.

220 Resultant peptide fragmentation data (MS/MS) obtained from each fraction/organism
221 were scored against their respective annotated proteomes downloaded from NCBI (Table 1)
222 using the SEQUEST database searching algorithm (19). Peptide-sequenced MS/MS spectra
223 were filtered (XCorr: +1= 1.8, +2 = 2.5, +3= 3.5, DeltCN 0.08) and assembled into protein loci by
224 DTASelect (82). Peptide spectral counts (SpC) resulting from intraspecies, non-unique peptides
225 were balanced across their shared protein source (22) to prevent overestimation of protein
226 abundance that could occur between proteins with high degrees of homology, i.e. glycoside
227 hydrolases. Once balanced, SpC for each fraction (SB, SN, and WC) were converted to
228 normalized spectral abundance factors (NSAF) (101) applying a fractional SpC shift (0.33) to all
229 proteins as described in (44). Normalized data from each species and fraction were grouped
230 together based on OrthoMCL in order to identify trending by orthologous proteins. Using the

JB Accepts published online ahead of print

NSAF values, enrichment scores for both SB (SBE=$NSAF_{SB}$/[$NSAF_{WC}$+$NSAF_{SN}$]) and EC (ECE=[$NSAF_{SB}$+$NSAF_{SN}$]/$NSAF_{WC}$) fractions were calculated. Subcellular and extracellular partitioning was calculated (50% equals equal partitioning) to visualize in Excel how the NSAF was split between fractions.

RESULTS AND DISCUSSION

General genome characteristics. Eight closed *Caldicellulosiurptor* genome sequences were examined to seek out determinants for the capacity to degrade plant biomass, defined by the ability to hydrolyze the various polysaccharide components of biomass, including crystalline cellulose (Table 1). These species represent globally diverse, terrestrial isolation sites (North America, Iceland, Russia, and New Zealand) (Fig.1A). Genome sizes for the *Caldicellulosiruptor* species range from 2.43 to 2.97 Mb, with an average genome size of 2.74 Mb and average G+C content of 35.5% across the genus (Table 1). Previous work has demonstrated a range in cellulolytic capacity for this closely-related genus (9). No one feature of the *Caldicellulosiruptor* genomes appears to correlate with location or phenotype; however, the two North American strains have the smallest genomes, both by nucleotide number and ORF count (Table 1). Phylogenetic analysis based on the three 16S rRNA loci found in each genome (Fig.1B) confirms previous reports that the genera are closely related to each other, with *C. saccharolyticus*, an isolate from New Zealand, the most divergent among this group. Building a dendrogram based on 16S ribosomal RNA (rRNA) phylogeny, species sharing common biogeography form location-specific clades, regardless of phenotype, such as the isolates from North America, Iceland and Russia (Fig.1A and 1B). Using members from the orders Clostridiales, Thermoanaerobacterales and Dictyoglomales as outgroups, *C. saccharolyticus*

JB Accepts published online ahead of print

appears to be the oldest member of its genus, due to greater divergence from the other species, having branched off earliest in the *Caldicellulosiruptor* clade (Fig.1A).

The ancestral nature of *C. saccharolyticus* is reinforced by considering the whole genome using average nucleotide identity (ANI, Fig.1C) (72). Considering location-specific clades formed using 16S rRNA sequences, we explored whether or not this same trend would hold true once entire genomes were considered. This proved to be the case, with the Icelandic species, which are highly related (shared ~98% identity) and the North American species (shared ~92% identity) (Fig.1C and Table S1). Interestingly, one species isolated from Russia, *C. hydrothermalis*, is slightly more related to an Icelandic species, *C. lactoaceticus*, when ANI is considered (Fig.1C and Table S1). Furthermore, when the closed genome sequences are aligned, based on geographical location, areas of macrosynteny are apparent, again regardless of cellulolytic phenotype (see Fig.S1), these areas of macrosynteny are not apparent when all eight genomes are aligned, due to genetic rearrangement during evolution of the genus (data not shown). While 16S phylogeny and ANI are widely used for taxonomic classification of species, they are not appropriate metrics to assign phenotype within the genus *Caldicellulosiruptor*, especially with respect to cellulolytic capability.

Growth on plant biomass and complex carbohydrates differentiates between weakly and strongly cellulolytic *Caldicellulosiruptor* species. To explore the relationship between genome content and growth on complex carbohydrates, the eight *Caldicellulosiruptor* species were cultured on crystalline cellulose (Avicel), xylan, acid-treated switchgrass and filter paper (Fig.2). While all species grew well on acid-pretreated switchgrass (SWG), which contains hemicellullose, cellulose and pectin (71), more variability was noted for growth on Avicel (Fig.2A). All species also grew well on xylan, as expected, based on the core genome (Fig.2). However, growth on Avicel (Fig.2A) and filter paper (Fig.2B) differentiated the strongly from weakly celluloytic species across the genus. In general, *C. bescii*, *C. kronotskyensis*, *C.*

JB Accepts published online ahead of print

279 *saccharolyticus*, and *C. obsidiansis* grew best on Avicel and filter paper, with *C. lactoaceticus*
280 growth at a somewhat lower level. *C hydrothermalis*, *C. kristjanssonii*, and *C. owensensis*,
281 however, grew less on these substrates, and did not break down filter paper to any visible
282 extent (Fig.2B). These growth experiments provided a perspective for comparative genomic
283 analysis with respect to crystalline cellulose hydrolyzing capability.

284 ***Caldicellulosiruptor* core- and pan-genomes.** In order to identify specific determinants
285 between the *Caldicellulosiruptor* genomes that would enable some species and not others to
286 hydrolyze crystalline cellulose, a baseline view of the genomes is required. The
287 *Caldicellulosiruptor* core and pan genomes (Fig. S1, also see Dataset 1), based on these eight
288 sequenced species, contain 1580 and 4009 genes, respectively (43, 86). The pan genome was
289 found to be open, such that the projected number of orthologous genes discovered with each
290 new species sequenced reaches an asymptote of 125 genes. This result is not surprising, given
291 that these species are isolated from dynamic environments, specifically environments with
292 variable nutrient types for organotrophic growth (28). Functional characterization of the core
293 *Caldicellulosiruptor* genome using COG analysis indicated that, while translation and amino acid
294 transport families are enriched in the core versus pan-genome, carbohydrate metabolism and
295 transport remain the major features of the genus *Caldicellulosiruptor* (Fig.S2 and Table S2). For
296 the core genome, approximately 120 genes are involved in carbohydrate transport and
297 metabolism according to COG classification (Table S2), which includes the main glycolysis
298 pathway, six ABC transporters and 30 CAZy-related proteins (Fig.S3). This suggests that the
299 core *Caldicellulosiruptor* genome is capable of extracellular xylan, glucan and starch hydrolysis,
300 xylan de-acetylation, as well as import of the resulting oligosaccharides and their catabolism
301 through central metabolism (Fig.3 and Fig.S4). Interestingly, all six of the core ABC transporters
302 are from the CUT1 group (Table S4) (91), which forms the basis for *Caldicellulosiruptor*
303 organotrophic import of oligosaccharides (75, 77) that are then further processed to their

JB Accepts published online ahead of print

304 respective monosaccharides. Of additional interest is the co-localization of glycoside hydrolases
305 and carbohydrate ABC transporters, especially among those included in the *Caldicellulosiruptor*
306 core genome (Fig.S3). A previous study (91) also observed this phenomenon in *C.*
307 *saccharolyticus*, and may be indicative of synergy between centralized carbohydrate
308 hydrolyzing and import systems. However, the core genome suggests that not all
309 *Caldicellulosiruptor* species are capable of crystalline cellulose hydrolysis, given that GHs
310 belonging to families known to exhibit these biocatalytic capabilities are not identifiable in
311 several genomes.

312 The convergence of the number of orthologs in the core genome and the open nature of
313 the *Caldicellulosiruptor* pan genome indicates that each species is endowed with a set of
314 specific genes beyond the core that relate to the types of carbohydrates present in their
315 environment. Comparisons between the frequencies of the *Caldicellulosiruptor*-unique KEGG
316 BLASTp hits in the core versus pan *Caldicellulosiruptor* genome showed an increase in unique
317 proteins in the pan versus core genome, with *C. bescii* possessing the largest number and
318 frequency of *Caldicellulosiruptor*-unique proteins (Table S3). Analysis of the top ranked
319 BLASTp hits from strongly cellulolytic versus weakly cellulolytic species did not exhibit any
320 trends based on cellulolytic capability. Top ranked KEGG BLASTp hits do highlight the major
321 phyla with homologs to proteins from the genus *Caldicellulosiruptor*, including Firmicutes,
322 Dictyoglomi, Thermotogae, Proteobacteria and Euryarchaeota. Since the genus
323 *Calidcellulosiruptor* is classified under the phylum Firmicutes, identifying the majority of best
324 BLASTp hits from Firmicutes was expected (48). Members of the phyla Dictyoglomi,
325 Thermotogae, Proteobacteria and Euryarchaeota are often isolated or identified from the same
326 locations as the genus *Caldicellulosiruptor* (33, 38). As such, *Caldicellulosiruptor* proteins that
327 are direct homologs to proteins from the above mentioned phyla are likely the result of historical
328 horizontal gene transfer (HGT) in their environment (53). Common biogeography influencing

JB Accepts published online ahead of print

16S rRNA and ANI-based phylogenetic analyses was not necessarily observed in context with the number of distinct phyla from KEGG best BLASTp hits, indicative of HGT that is not otherwise detected by phylogenetic analyses. For example, the highly related species *C. kristjanssonii* and *C. lactoaceticus* (ANI = 98 to 98.1%, Table S1), share similar frequencies of best BLASTp hits from the major related phyla (Table S3), however *C. kristjanssonii* has BLASTp best hits to a total of 31 phyla compared to *C. lactoaceticus* with 22 phyla. Due to the open nature of the *Caldicellulosiruptor* pan genome, HGT events are important for the evolution of *Caldicellulosiruptor* species capable of succeeding in their dynamic environments. Increasing the number of *Caldicellulosiruptor* genome sequences available would also further identify unique genes acquired through HGT, a fraction of which map back to specific aspects of carbohydrate hydrolysis, transport and metabolism.

Relationship between ABC carbohydrate transporter inventory and growth substrate range. Since non-core genes appear to be involved in a species' ability to hydrolyze crystalline cellulose, the inventory of carbohydrate transporters was first considered. Overall, the genus *Caldicellulosiruptor* has six core ATP-binding cassette (ABC) transporters out of 45 in the pan-genome (Table S4). Substrate preferences for five of these core transporters have previously been assigned based on transcriptomic analysis of *C. saccharolyticus* (91). Only *C. hydrothermalis*, *C. kronotskyensis* and *C. saccharolyticus* contain unique transporters not found in the other sequenced *Caldicellulosiruptor* species. As mentioned above, all of the core ABC transporters are of the CUT1 type, which are typically involved in oligosaccharide import (75, 77), although some of these CUT1 transporters from *C. saccharolyticus* will respond to monosaccharides (91). These transporters appear to import some, but not all, oligosaccharides that are generated by plant biomass hydrolysis. As there is a wide variety of CAZy-related genes found in the *Caldicellulosiruptor* genomes, there are also particular ABC transporters used by individual species to support growth on various types of plant biomass.

JB Accepts published online ahead of print

A connection between ABC transporter number and diversity and substrate range was evident in examining the genomes. *C. lactoaceticus* has the most restricted carbohydrate preferences (9, 58) and also has the fewest carbohydrate-related ABC transporters, in fact one-third of those found in *C. hydrothermalis*. This further supports the point that *C. lactoaceticus* has evolved as a specialist on higher-chain plant polysaccharides and cannot use glucose to support growth due to a lack of a transporter for glucose. The next closest related species to *C. lactoaceticus*, *C. kristjanssonii*, has only three more transporters than *C. lactoaceticus*, and is capable of growth on glucose (9, 12), strongly implicating one of those three additional transporters in glucose uptake. Two of these transporters have previously been implicated in glucose import for *C. saccharolyticus*, and this finding further confirms that result (91).

C. hydrothermalis contains the most transporters of any member of the genus, with 39 ABC transporters predicted to be involved in carbohydrate transport. On a whole, the G+C content of *C. hydrothermalis* is higher than the rest of the genus (Table 1), implying that it has obtained genes through HGT. Indeed, seven ABC transporters from *C. hydrothermalis* are unique to the genus and could potentially be the result of HGT. Interestingly, *C. hydrothermalis* grows weakly on Avicel (Fig.2A), and does not visibly deconstruct filter paper (Fig.2B), indicating that transporter inventory does not correlate to the ability to hydrolyze crystalline cellulose. Instead, it appears that *C. hydrothermalis* may have evolved by either importing diverse types of carbohydrates into the cell or using multiple transporters to maximize uptake of specific carbohydrates.

Overall, no common transporter set could be identified that was only present in cellulolytic, but not weakly cellulolytic *Caldicellulosiruptor* species (Table S4). This finding seems reasonable, given that all isolated species have been described as having the ability to grow on cellobiose (12, 28, 32, 62, 70, 81, 98). Since these bacteria are assumed to live in plant biomass degrading communities, even if a species is lacking strong cellulolytic machinery, it would be beneficial to maintain the ability to import cellulose hydrolysis products. In addition, no

JB Accepts published online ahead of print

correlation between the number of transporters and cellulolytic ability was evident (Table 2). However, the diversity of carbohydrate transporters in weakly cellulolytic species merits further consideration, when considering the design of a biocatalyst for CBP. By incorporating a large number of diverse carbohydrate transporters, flux through many different catabolic pathways could be maintained, supported by the fact that the genus does not exhibit carbon catabolite repression (CCR) (89, 91).

Similarities and subtle differences in core metabolism influence carbohydrate preferences. Since carbohydrate transporter diversity did not appear to correlate with specific determinants for cellulolytic ability, an examination of the metabolic capacity should be considered. However, based on the information reported here and for the previously sequenced *Caldicellulosiruptor* genomes (16, 89), the core metabolic pathways across the genus appear to be highly conserved. All species are capable of glycolysis through the Embden-Meyerhof-Parnas (EMP) pathway, fermentation of xylose through a non-oxidative pentose phosphate pathway (PPP), uronic acid metabolism and oxidation of acetate-CoA/ reduce pyruvate through an incomplete citric acid cycle (TCA) (Fig.S4). The highly conserved EMP pathway would be responsible for oxidation of glucose liberated from cellulose or starch, and highlights the importance of both α-D- and β-D-glucose as an energy source.

Outside of metabolism of cellulose and pectin, there are some differences between *Caldicellulosiruptor* species with respect to various monosaccharide metabolic pathways involved in hemicellulose metabolism. One subtle difference concentrates on the xylose isomerase (XI) of *C. saccharolyticus*, which is a class I XI, in contrast to the other species which use a class II XI (29, 41). The significance of this is currently unknown; however, all *Caldicellulosiruptor* species grow well on xylose (9) and the analgous complex polysaccharide xylan (Fig.2A), indicating that both types of XI are able to catalyze efficient xylose metabolism for the genus *Caldicellulosiruptor*.

JB Accepts published online ahead of print

Three other alternative pathways that feed into the PPP involve other aldopentoses: D-ribose, L-arabinose and D-arabinose. To metabolize L-arabinose, a component of pectin and arabinoxylan, a putative L-fucose isomerase (MCL group 1847, Dataset 1) appears to be used by most species (Fig.S4). In contrast, the Icelandic *Caldicellulosiruptor* species lack the genes to metabolize any of these aldopentoses, which also explains their inability to grow on D/L-arabinose and ribose (12, 58). This apparent lack of D/L-arabinose-specific isomerases and kinases would then theoretically reduce their capacity to metabolize a portion of the hydrolysis products from arabinoxylan.

Another example of upstream carbohydrate conversion pathways influencing carbohydrate growth profiles is metabolism of deoxy-sugars, such as L-fucose and L-rhamnose. The plant cell wall component pectin can contain L-rhamnose, and xyloglucans can also be fucosylated (99), making the catabolism of deoxy-sugars important for the complete metabolism of all biomass-related carbohydrates. While some species possess complete pathways to metabolize deoxy-sugars, not all species have been described as being capable of growth on them; for example, *C. bescii* was described as being unable to grow on fucose (81). In addition, other species with incomplete deoxy-sugar pathways have been described as capable of growth on rhamnose, *C. owensensis* being one such example (32). This highlights the overall need for a better understanding of the alternate upstream carbohydrate conversion pathways.

Glycoside Hydrolase (GH) inventory reflects the capacity for crystalline cellulose hydrolysis. Ultimately, the answer to what makes an organism weakly or strongly cellulolytic rests to a large extent on their enzymatic inventory. As discussed above, the inventory of carbohydrate transporters and metabolic pathways only give clues about the metabolic capacity of the organism after deconstruction of plant biomass. Therefore, a comparative analysis of their glycoside hydrolase (GH) inventory should highlight distinct determinants for cellulose deconstruction. The pan-genome of the genus *Caldicellulosiruptor* encodes 134 carbohydrate-

JB Accepts published online ahead of print

active enzymes (CAZy) (13), here classified as glycoside hydrolases (GHs), carbohydrate esterases (CEs), polysaccharide lyases (PLs), and carbohydrate binding modules (CBMs); 48 of these contain signal peptides and are predicted to be extracellular (Table 2). Carbohydrate-active enzyme inventory of the pan genome constitutes the collective capacity of the genus to metabolize complex and simple carbohydrates, including various types of plant biomass. In a preliminary screen of carbohydrate-active enzyme inventory from the genus based on draft genome sequence data, GH family 48 and CBM family 3 were implicated as being essential elements for crystalline cellulose hydrolysis by *Caldicellulosiruptor* species (9). With eight finished genome sequences, a more complete assessment can be done.

As might be expected of microorganisms capable of plant biomass degradation, each *Caldicellulosiruptor* species contains a significant number of GH domains and CBM modules in their genomes, ranging from 38 and 26, respectively, for *C. kristjanssonii*, and up to 84 and 63 respectively, for *C. kronotskyensis* (Table 2). These numbers are high in comparison to other thermophilic anaerobes, but are smaller when compared to fungi, such as *Trichoderma reesei* (~200) (16, 55). The genome of *C. kronotskyensis* contains 84 GH domains that represent 38 different GH families, which is also the highest diversity of GH domains found in an anaerobic thermophile (13, 61). This is about 50% more GH domains than many other *Caldicellulosiruptor* species (Table 2). However, the diversity of GH families does not necessarily map back to cellulolytic capability, as *C. hydrothermalis* and *C. saccharolyticus* possess 60 and 61 families, respectively, and have vastly different plant biomass deconstruction capabilities (Fig.2B).

Approximately, one-fourth of the 121 CAZy-related ORFs are conserved across all eight sequenced *Caldicellulosiruptor* genomes and constitute the core. These 30 ORFs include 26 enzymes containing GH domains, three containing carbohydrate esterase (CE) domains, and one with only a single CBM domain (Fig.3A). Four ORFs from this core are predicted to be extracellular, including Csac_0678 and its orthologs: a bifunctional GH5 domain enzyme (64); a

JB Accepts published online ahead of print

455 putative xylanase, a putative pullulanase, and a carbohydrate esterase (Fig.3A). In theory, these
456 genes represent the minimal set of CAZy-related genes required for biomass deconstruction by
457 a *Caldicellulosiruptor* species. While it may be tempting to use this list as an indication of the
458 minimal set of extracellular enzymes required by the genus to support a plant biomass-
459 degrading lifestyle, functional homology of non-core enzymes must also be considered. Indeed,
460 *C. kristjanssonii* has 11 GH domain-containing enzymes above the core *Caldicellulosiruptor* set,
461 the lowest number of total GH domain containing enzymes in the genus. Note that the minimal
462 set of carbohydrate-active enzymes in the genus does not equip the microbe for crystalline
463 cellulose hydrolysis, although the GH5-containing enzyme does allow for random cleavage of
464 amorphous cellulose (64). *C. lactoaceticus*, a closely related species to *C. kristjanssonii*, is
465 cellulolytic and possesses only 6 more CAZy-related ORFs above that of *C. kristjanssonii* (Table
466 2). Comparison with core-cellulolytic enzymes will highlight which of these six additional CAZy-
467 related ORFs are those important for cellulose hydrolysis. It appears that both species isolated
468 from Iceland are “minimalists” with respect to gene inventory for carbohydrate hydrolysis.

469 Seven additional genes conserved among the cellulolytic species comprise the core
470 cellulolytic carbohydrate-active enzyme list (Fig.3B). This set includes full or partial homologs of
471 enzymes with GH9 and GH48 domains linked by CBM3 modules, GH74 and GH48 domains
472 linked by CBM3 modules, and a GH9 and GH5 linked by CBM3 modules (Fig.3B). Indeed, as a
473 previous preliminary analysis suggested, those species that are strongly cellulolytic also
474 possess enzymes with GH9, GH48 catalytic domains and CBM3 modules (9) (see Tables S5
475 and S6). In fact, these enzymes are co-localized in loci that contain anywhere from four to
476 seven modular multi-domain enzymes, all of which possess CBM3 modules (Fig.4). One weakly
477 cellulolytic species, *C. kristjanssonii*, also has some CBM3-linked enzymes; however, none also
478 have a GH48 domain, which appears to be the absolute determinant for crystalline cellulose
479 hydrolysis in the genus (see Table S5). In the comparison between *C. kristjanssonii* and *C.*
480 *lactoaceticus*, where six additional ORFs are present in the cellulolytic *C. lactoaceticus*, three

JB Accepts published online ahead of print

481 are core to cellulolytic species and only two multi-domain multifunctional ORFs are only
482 encoded by cellulolytic *Caldicellulosiruptor* species, the GH74:GH48 enzyme and CelA, a
483 GH9:GH48 enzyme (Fig.3B). Family 48 GHs are often characterized as cellobiohydrolases (2),
484 supporting the theory that this particular family is responsible for the strong cellulolytic
485 phenotype. Indeed, mutations in GH48-containing enzymes have disrupted the cellulolytic ability
486 of *Ruminococcus albus* 8 (17), and reduced the cellulolytic efficiency of *Clostridium*
487 *thermocellum* (61) and *Clostridium cellulolyticum* (67). There are cases, however, where the
488 sole presence of a GH48 domain is not enough to promote a strong cellulolytic phenotype, as is
489 the case for the cellulosomal *Clostridium acetobutylicum* (60, 74), even though the GH48
490 enzyme was expressed and secreted as part of the cellulosome (46). Evidently, even in the
491 case of the strongly cellulolytic *Caldicellulosiruptor* species, additional determinants beyond the
492 presence of GH domains and CBM modules most likely exist that promote crystalline cellulose
493 hydrolysis.

494 **Identification of secreted proteins provides insights into substrate attachment and**
495 **hydrolysis.** To further probe what determinants exist beyond the cellulolytic GH family-
496 containing enzymes in the genus *Caldicellulosiruptor*, Avicel-induced proteins were identified via
497 bottom-up proteomics. Avicel was used as a model plant biomass substrate due to the large
498 proportion of cellulose in plant cell walls, and previous studies on *T. reesei* cellulase systems
499 demonstrating strong affinity of cellulases for Avicel (39, 63). A strong, potential irreversible
500 interaction between *Caldicellulosiruptor* proteins and Avicel would be ideal for proteomic
501 screening to identify substrate-bound proteins, since their affinity for Avicel would have to
502 withstand washing steps to remove cells. Previous proteomic screens from members of the
503 genus focused on the cell-free extracellular and whole cell fractions of cellulolytic
504 *Caldicellulosiruptor* species (16, 44, 45). We previously reported on differential two dimensional
505 SDS-PAGE profiles of cell-free supernatant from cells grown on Avicel, in an attempt to capture

JB Accepts published online ahead of print

506 protein-level differences of weakly to strongly cellulolytic *Caldicellulosiruptor* species (9). In
507 order to fully capture differential protein expression between weakly and strongly cellulolytic
508 *Caldicellulosiruptor* species an expanded proteomic screen was employed. Here, we describe
509 the first comprehensive genus-wide screen of Avicel-induced proteins identified not only from
510 cell-free supernatant (SN), and whole cell lysate (WC) but also from the Avicel-bound (SB)
511 fraction from four selected strongly cellulolytic and three weakly cellulolytic *Caldicellulosiruptor*
512 species.

513 Overall, between 36 to 48% of total protein coding sequences predicted from
514 *Caldicellulosiruptor* genomes was detected as peptides from the SB, SN and WC fractions using
515 mass spectrometry (see Dataset 2). This is lower than 54% detection for *C. bescii* (16) or 65%
516 detection for *C. obsidiansis* (44) however, previous experiments included two or more growth
517 substrates analyzed and/or measurements at various growth stages, whereas this study only
518 included one growth substrate, Avicel. Peptides identified in the SB-fraction range from 16 to
519 24% of total protein coding sequences detected, however the numbers could be inflated by the
520 presence of intercellular proteins released by cells adhered irreversibly to Avicel. Weakly
521 cellulolytic *Caldicellulosiruptor* species had the lowest frequency (16.7 and 20.1% for *C.*
522 *owensensis* and *C. hydrothermalis*, respectively) of proteins detected in the SB fraction. This
523 result is not unexpected. A weakly cellulolytic species would not be expected to produce many
524 proteins that interact with cellulose, including the above mentioned multi-domain modular
525 enzymes with CBM family 3 motifs. However, another weakly cellulolytic species, *C.*
526 *kristjanssonii*, had the largest frequency of protein coding sequences detected in the SB
527 fraction, again potentially from intercellular leakage. In fact, when the average substrate-bound
528 enrichment score (SBE) for *C. kristjanssonii* is considered, it is lower than the average SBE over
529 the entire genus, indicative of intercellular protein contamination of the SB fraction.

JB Accepts published online ahead of print

Identification of glycoside hydrolases bound to cellulose. Peptides classified as CAZy-related ORFs were detected at higher frequencies than the complete proteome, ranging from 54 to 83% detection (see Dataset 2). As expected, one of the highest detected fractions of extracellular peptides corresponded to proteins encoded by the gene cluster containing enzymes with CBM3 modules (MCL cluster 4, Dataset 2). These GHs were also enriched in the SB fraction versus the SN fraction (weighted percentages of 88%, 3%, and 9% total NSAF for SB, SN, and WC, respectively), agreeing with the cellulose-binding function of CBM family 3 modules (88). One particular group, orthologs of CelA (GH9-CBM3-CBM3-CBM3-GH48, Fig.3B), was the most abundant CBM3-containing enzyme detected in the SB-fraction. Previous studies identifying extracellular proteins in *C. bescii* and *C. obsidiansis* grown on cellulose also found that CelA is the most abundant CBM3-containing enzyme produced by cellulolytic *Caldicellulosiruptor* species (45). Enrichment of cellobiohydrolases bound to Avicel has been noted before; in competitive binding assays using *T. reesi* cellulases, including cellobiohydrolases and endo-glucanases, the cellobiohydrolases bound with a higher affinity to Avicel (39). This observation further highlights the association of modular multi-domain enzymes containing both GH48 and CBM3 domains to crystalline cellulose and emphasizes their important role in its hydrolysis.

One benefit of identifying proteins in the SB fraction is the discovery of previously overlooked enzymes, such as the enrichment of a modular multi-domain mannanase (GH26) enzyme on cellulose (23). This cellulolytic enzyme contains two CBM families, CBM27 and CBM35, which are found in the genomes of *C. hydrothermalis*, *C. kristjanssonii*, *C. lactoaceticus* and *C. obsidiansis* (MCL Group 2116, Dataset 2). Enrichment of this enzyme in the SB fraction was significantly higher in two weakly cellulolytic species, *C. hydrothermalis* and *C. kristjanssonii* (NSAF= 1.57×10^{-2} and 4.82×10^{-3} respectively), and significantly lower (almost non-existent) in the cellulolytic *C. lactoaceticus* (NSAF= 2.35×10^{-4}). Furthermore, there was no detection of this protein either in the SN or WC fractions of strongly cellulolytic *C. obsidiansis*

JB Accepts published online ahead of print

grown on cellobiose, cellulose or switchgrass, as shown in another study (44). At minimum, this indicates that there are different regulatory mechanisms for weakly versus strongly cellulolytic species; those species lacking CBM3 protein loci are likely compensating with other enzymes. As mentioned above, previous reports using an orthologous enzyme from *Caldicellulosiruptor* sp. Rt8B.4 (23), characterized this enzyme as a mannanase, and there has been no further description of enzyme activity beyond that on gluco- and galactomannans (79). However, when carbohydrate binding specificity of the CBM motifs was investigated, it was noted that the N-terminus of the protein, comprised of the CBM motifs, demonstrated affinity for not only mannan, but glucans such as soluble cellulose and β-glucan (80). It is not unusual for non-cellulolytic enzymes to be targeted to cellulose in order to decouple cellulose from surrounding polysaccharides, as is the case for some of the multi-modular enzymes containing CBM3 motifs (MCL group 4, Dataset 1 and 2).

Non-catalytic proteins bound to cellulose. Other proteins that have been theorized to be involved in microbe-substrate interactions were also enriched in the substrate-bound fraction (Fig.5). The major protein that forms the S-layer (MCL group 219, see Dataset 2) was found in the extracellular fractions of all species in significant amounts. In fact, this protein alone constituted over 9% of the total spectra collected across all organisms, with overall fractional partitioning of 35%, 53%, and 12% (SB, SN, and WC, respectively). However, as previously observed with 2-dimensional SDS-PAGE analysis (9), the supernatants of *C. saccharolyticus* and *C. owensensis* are enriched with the S-layer protein (see Dataset 2) relative to the other *Caldicellulosiruptor spp*. The recently characterized, S-layer-located cellulolytic enzyme, Csac_0678 (64), was also enriched in the SB-fraction (MCL group 1342, Dataset 2), as expected. Interestingly, only the ortholog from *C. owensensis* was strongly enriched in the SN-fraction, potentially as a result of the truncated CBM28 motif, highlighting the importance of this particular CBM family in adherence to non-crystalline portions of Avicel (11). A role for four other

JB Accepts published online ahead of print

581 S-layer-associated proteins in substrate attachment can be assigned as well from their
582 observed binding to Avicel (Fig.5). Though the majority does not have identifiable carbohydrate
583 binding modules, they all strongly partition towards the SB fraction (86% of their total SpC
584 collected overall).

585 Another group of proteins potentially involved in substrate attachment are those
586 assembled into flagella (Fig.5). Surprisingly, proteins comprising the flagella were detected
587 primarily in the SN fraction for strongly cellulolytic species (22%, 67%, and 11% for SB, SN, and
588 WC, respectively based on total NSAF), while for the weakly cellulolytic species the proteins
589 were enriched in the SB fraction (54%, 37%, and 9% for SB, SN, and WC, respectively).
590 Enrichment of the flagella in the SN fraction of strongly cellulolytic species indicates that
591 cellulose will induce expression of flagellar genes, although in this case, the flagella were not
592 detected as playing a role in cellular adhesion. In contrast, enrichment of flagella components in
593 the SB fraction indicates a more important role for flagella in cellulose adhesion for weakly
594 cellulolytic species. A two-stage mechanism for cell-surface attachment has been proposed for
595 the Proteobacterium, *Caulobater crescentus*, with the reversible primary surface attachment
596 mechanism involving the flagella, followed by attachment by type IV pili prior to biofilm formation
597 in the irreversible secondary phase (42). Clearly, there are differing mechanisms for cellulose-
598 attachment even within the genus *Caldicellulosiruptor*, and the enrichment of flagella-related
599 proteins in the SB fraction from weakly cellulolytic species may be indicative of an extended
600 reversible attachment phase.

601 Formation of a cellulolytic biofilm by the strongly cellulolytic species *C. obsidiansis* on
602 cellulose surfaces has been shown previously (92). Since this irreversible secondary stage of
603 cell-surface attachment occurs with a strongly cellulolytic species, we looked at the abundance
604 of type IV pilus-related proteins to determine if these structures play a role. Indeed, fewer pre-
605 pilin subunits were detected for two of the weakly cellulolytic species as compared to strongly
606 cellulolytic species. In addition, the pre-pillin subunits were enriched in the SN fraction for all

JB Accepts published online ahead of print

607 species (5%, 93%, and 2% for SB, SN, and WC, respectively). However, almost 7-fold less of
608 these proteins were detected for the weakly cellulolytic species (MCL groups 443, 1652 and
609 1819, Fig.5 and Dataset 2).

610 Proteins from the type IV pilus genomic region that were enriched in the SB fraction
611 (82% of total NSAF for MCL 1820; 97% of total NSAF for MCL 1653) belonged almost
612 exclusively to the strongly cellulolytic species (Table 3). Annotated as hypothetical proteins, they
613 have no significant homology to proteins outside of the genus. Orthologs from highly cellulolytic
614 speices (*C. bescii*, *C. kronotskyensis*, *C. obsidiansis* and *C. saccharolyticus*) had identity scores
615 ranging 81% to 95% (MCL 1820) and 85% to 99% (MCL 1653), whereas orthologs from species
616 isolated in Iceland were highly homologous with each other (99% identity) yet much more
617 divergent from the highly cellulolytic group with identity scores ranging from 36% to 37% (MCL
618 1820) and 40% (MCL 1653). Indeed, when predicted parameters such as molecular weight and
619 isoelectric point are compared within MCL group 1820 and 1653, orthologs from *C.*
620 *lactoaceticus* and *C. kristjanssonii* are the smallest proteins, and in the case of MCL group 1820
621 the most positively charged with a predicted pI over 8 (Table 3).

622 Orthologous MCL group 1820 is expressed by all species examined, and was enriched
623 in the SB fraction, in some cases over 90% of total NSAF. Since an ortholog in MCL group 1820
624 is also expressed and found enriched in the SB fraction from the weakly cellulolytic *C.*
625 *kristjanssonii*, these proteins do not impart a strong cellulolytic phenotype. However, the ORF
626 directly downstream, represented by orthologous MCL group 1653, was only detected by
627 proteomic screening in the highly cellulolytic species examined and was also enriched in the SB
628 fraction (Table 3). The demonstrated differential expression of this MCL group during growth on
629 cellulose implicates MCL group 1653 in a *Caldicellulosiruptor* species' ability to hydrolyze
630 crystalline cellulose. Based on genomic proximity of the ORFs to the type IV pilus locus and the
631 enrichment of these proteins in the SB fraction, we propose that these proteins are novel

JB Accepts published online ahead of print

adhesins that mediate attachment of type IV pili to cellulose (MCL groups 1820 and 1653, Fig.5 and Dataset 2). Gram positive species sequenced so far generally have one genomic locus that contains the cluster of type IV pilus assembly genes, including hypothetical proteins located adjacent to the locus (66). When the genomic neighborhood of type IV pili genes were examined, it appears that the adhesins and the type IV pili locus also reside directly upstream of the cellulase gene cluster in strongly cellulolytic species lending evidence towards a synergistic expression pattern (see Table S7). No orthologs of these adhesins are found in the genome of *C. owensensis*, a weakly cellulolytic species, which instead possesses other adhesin-like proteins located downstream of the type IV pili locus (Table 3 and Table S7). However, both adhesins from *C. owensensis* were enriched in the SN fraction and the sole adhesin from *C. hydrothermalis* was not detected in any of the protein fractions (Table 3). A potential role for those adhesin-like proteins cannot be ruled out, and indeed low levels of mRNA for Calhy_0908 were detected when *C. hydrothermalis* was grown on cellobiose or switchgrass (data not shown) In the case of *C. owensensis*, while type IV pilus-proximate proteins were not enriched in the SB fraction, these proteins are expressed in response to the detection of cellooligosaccharides and may mediate attachment to other polysaccharides found in biomass such as xylan, pectin or mannans. Determination of the polysaccharide specificity of these putative adhesins, as well as further characterization of the interplay between neighboring adhesins are the subjects of ongoing experiments.

Was the ancestral *Caldicellulosiruptor* cellulolytic? The genomic neighborhoods of type IV pili and CBM3-containing enzymes present an interesting case of presumed genomic rearrangement of cellulases in a weakly cellulolytic species, *C. kristjanssonii* and the closely related strongly cellulolytic species, *C. lactoaceticus*. Since the CBM3-containing enzymes of *C. kristjanssonii* and *C. lactoaceticus* are found in blocks throughout their respective genomes instead of a single locus, genomic rearrangement can explain the separation of the type IV

JB Accepts published online ahead of print

locus and CBM3-containing enzymes. Genomic rearrangement in this locus could explain the lack of GH48-containing enzymes in *C. kristjanssonii* and, hence, weak growth on crystalline cellulose (Fig.2A). Since the genomic identity is very close (ANI of ~98%, see Table S1) between these two species with vastly different phenotypes on cellulose, it begs the question of which phenotype came first in the *Caldicellulosiruptor* lineage: strongly or weakly cellulolytic?

Two clusters of CAZy-related enzymes exist in the pan-genome; one cluster includes primarily glucan-degrading enzymes (GDL) with CBM 3 (CBM3) domains (Fig.4) and the other that contains xylan-degrading enzymes (XDL) and xylooligosaccharide transporters (91). Since *Caldicellulosiruptor* species from more than one continental location contain one or both clusters, it is likely that the ancestral *Caldicellulosiruptor* species contained both clusters. This also suggests that the ancestral *Caldicellulosiruptor* species was strongly cellulolytic and capable of crystalline cellulose deconstruction, and that weakly cellulolytic species have lost that ability through gene deletion events.

Most members of the genus, except *C. hydrothermalis* and *C. owensensis*, have at least one homolog contained within the GDL, which would mean that *C. hydrothermalis* and *C. owensensis* either branched off from the *Caldicellulosiruptor* lineage prior to acquisition of those genes by the ancestral species, or that they lost the entire region after speciation. As mentioned before, the type IV pilus operon is also located directly upstream of the GDL in strongly cellulolytic species. The separation of these co-located regions, in addition to further genomic rearrangements in the GDL of Icelandic species, makes it likely that *C. hydrothermalis* and *C. owensensis* lost the GDL after speciation. In addition to the loss of the GDL, these two species also lost one or both cellulose-associating adhesins from the type IV pilus loci, indicating that gene loss occurred further upstream than just the GDL. Furthermore, if the weakly cellulolytic *Caldicellulosiruptor* species were the result of a separate lineage in the genus, one would expect the weakly cellulolytic species to be more genetically similar to one another, which 16S

JB Accepts published online ahead of print

682 phylogeny and ANI both disprove (Fig.1 and Table S1). It is also interesting that many genes
683 located in the GDL cluster of the strongly cellulolytic *Caldicellulosiruptor* species appear to be
684 the result of various recombination events after gene duplication of glycoside hydrolase
685 domains with CBM3 domains (24, 36, 61) (Fig.4). Microsynteny in the GDL between *C.*
686 *saccharolyticus* and *C. kronotskyensis*, two geographically distinct species (Fig.4), indicates that
687 there has been additional rearrangement in the GDL of *C. bescii* after speciation.

688 **Conclusions.** Eight whole genome sequences from the genus *Caldicellulosiruptor*,
689 ranging from weakly to strongly cellulolytic species (Fig.2A), were assessed for determinants of
690 cellulolytic capability. While biogeography was determined to play a role in the level of
691 relatedness between species based on 16S phylogeny and ANI (Fig.1), it was not a reliable
692 metric to predict phenotype. Using detailed comparative analysis of the genomes, carbohydrate
693 transport and catabolic pathways were indicative of carbohydrate metabolic capabilities.
694 However, genomic analysis is not enough to predict cellulolytic capability. This is not to say that
695 there is no benefit of such an analysis, since metabolic engineering of a CBP organism will
696 require detailed characterization of the import and metabolism of carbohydrates.

697 Further analysis of the CAZy-related gene inventory did reaffirm previously predicted
698 determinants for cellulolytic ability, namely enzymes possessing GH family 48 domains with
699 CBM family 3 modules. Indeed, when the GDL for the cellulolytic species *C. lactoaceticus* is
700 compared to the highly related *C. kristjanssonii*, the presence of a GH48-containing enzyme, a
701 GH5-containing enzyme and an additional GH9 enzyme in *C. lactoaceticus* are the main
702 differences. Since *C. kristjanssonii* already possesses a GH9 linked to CBM3 modules and
703 other GH5-containing enzymes in its genome, it is unlikely that these were the determinants for
704 a cellulolytic phenotype. Most likely, it is the presence of a GH48-containing enzyme that makes
705 the difference, since GH family 48 are most often characterized as cellobiohydrolases (13).
706 Additionally, when species that grow better than *C. lactoaceticus* on cellulose are considered

JB Accepts published online ahead of print

(Fig.2A), the enzyme CelA, which links a GH9 and GH48 with three CBM3 modules (Fig.4B), appears to be the determinant for strong cellulolytic growth. Lastly, proteomic-based identification of substrate-bound extracellular proteins revealed additional determinants for a strong cellulolytic phenotype, including two type IV pilus associated adhesins. As more *Caldicellulosiruptor* species genomes become available, the insights discussed here can be further evaluated.

ACKNOWLEDGEMENTS

This work was supported by the Bioenergy Science Center (BESC), Oak Ridge National Laboratory, a U.S. Department of Energy Bioenergy Research Center funded by the Office of Biological and Environmental Research in the DOE Office of Science [contract no. DE-PS02-06ER64304][DOE 4000063512]. We gratefully acknowledge the efforts of Lynne Goodwin (JGI-LANL) and Karen Walston Davenport (LANL) toward the *Caldicellulosiruptor* sequencing project. We also thank Dhaval Mistry and Dustin Nelson (NCSU) for their technical assistance in gathering physiological data.

JB Accepts published online ahead of print

REFERENCES

1. **Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389-3402.
2. **Barr, B. K., Y. L. Hsieh, B. Ganem, and D. B. Wilson.** 1996. Identification of two functionally different classes of exocellulases. Biochemistry **35:**586-592.
3. **Bayer, E. A., R. Lamed, B. A. White, and H. J. Flint.** 2008. From cellulosomes to cellulosomics. Chem. Rec. **8:**364-377.
4. **Bendtsen, J. D., H. Nielsen, G. von Heijne, and S. Brunak.** 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. **340:**783-795.
5. **Bennett, S.** 2004. Solexa Ltd. Pharmacogenomics **5:**433-438.
6. **Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers.** 2011. GenBank. Nucleic Acids Res. **39:**D32-37.
7. **Berka, R. M., I. V. Grigoriev, R. Otillar, A. Salamov, J. Grimwood, I. Reid, N. Ishmael, T. John, C. Darmond, M.-C. Moisan, B. Henrissat, P. M. Coutinho, V. Lombard, D. O. Natvig, E. Lindquist, J. Schmutz, S. Lucas, P. Harris, J. Powlowski, A. Bellemare, D. Taylor, G. Butler, R. P. de Vries, I. E. Allijn, J. van den Brink, S. Ushinsky, R. Storms, A. J. Powell, I. T. Paulsen, L. D. H. Elbourne, S. E. Baker, J. Magnuson, S. LaBoissiere, A. J. Clutterbuck, D. Martinez, M. Wogulis, A. L. de Leon, M. W. Rey, and A. Tsang.** 2011. Comparative genomic analysis of the thermophilic biomass-degrading fungi *Myceliophthora thermophila* and *Thielavia terrestris*. Nat. Biotechnol. **29:**922-927.
8. **Blumer-Schuette, S. E., I. Kataeva, J. Westpheling, M. W. Adams, and R. M. Kelly.** 2008. Extremely thermophilic microorganisms for biomass conversion: status and prospects. Curr. Opin. Biotechnol. **19:**210-217.

JB Accepts published online ahead of print

9. **Blumer-Schuette, S. E., D. L. Lewis, and R. M. Kelly.** 2010. Phylogenetic, microbiological, and glycoside hydrolase diversities within the extremely thermophilic, plant biomass-degrading genus *Caldicellulosiruptor*. Appl. Environ. Microbiol. **76:**8084-8092.

10. **Blumer-Schuette, S. E., I. Ozdemir, D. Mistry, S. Lucas, A. Lapidus, J.-F. Cheng, L. A. Goodwin, S. Pitluck, M. L. Land, L. J. Hauser, T. Woyke, N. Mikhailova, A. Pati, N. C. Kyrpides, N. Ivanova, J. C. Detter, K. Walston-Davenport, S. Han, M. W. W. Adams, and R. M. Kelly.** 2011. Complete genome sequences for the anaerobic, extremely thermophilic plant biomass-degrading bacteria *Caldicellulosiruptor hydrothermalis*, *Caldicellulosiruptor kristjanssonii*, *Caldicellulosiruptor kronotskyensis*, *Caldicellulosiruptor owensensis*, and *Caldicellulosiruptor lactoaceticus*. J. Bacteriol. **193:**1483-1484.

11. **Boraston, A. B., M. Ghaffari, R. A. J. Warren, and D. G. Kilburn.** 2002. Identification and glucan-binding properties of a new carbohydrate-binding module family. Biochem. J. **361:**35-40.

12. **Bredholt, S., J. Sonne-Hansen, P. Nielsen, I. M. Mathrani, and B. K. Ahring.** 1999. *Caldicellulosiruptor kristjanssonii* sp. nov., a cellulolytic, extremely thermophilic, anaerobic bacterium. Int. J. Syst. Bacteriol. **49** 991-996.

13. **Cantarel, B. L., P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat.** 2009. The Carbohydrate-Active enZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. **37:**D233-D238.

14. **Claudel-Renard, C., C. Chevalet, T. Faraut, and D. Kahn.** 2003. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res. **31:**6633-6639.

15. **Consortium, T. U.** 2010. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. **39:**D214-D219.

JB Accepts published online ahead of print

16. **Dam, P., I. Kataeva, S. J. Yang, F. Zhou, Y. Yin, W. Chou, F. L. Poole, 2nd, J. Westpheling, R. Hettich, R. Giannone, D. L. Lewis, R. Kelly, H. J. Gilbert, B. Henrissat, Y. Xu, and M. W. Adams.** 2011. Insights into plant biomass conversion from the genome of the anaerobic thermophilic bacterium *Caldicellulosiruptor bescii* DSM 6725. Nucleic Acids Res. **39:**3240-3254.

17. **Devillard, E., D. B. Goodheart, S. K. R. Karnati, E. A. Bayer, R. Lamed, J. Miron, K. E. Nelson, and M. Morrison.** 2004. *Ruminococcus albus* 8 mutants defective in cellulose degradation are deficient in two processive endocellulases, Cel48A and Cel9B, both of which possess a novel modular architecture. J. Bacteriol. **186:**136-145.

18. **Elkins, J. G., A. Lochner, S. D. Hamilton-Brehm, K. W. Davenport, M. Podar, S. D. Brown, M. L. Land, L. J. Hauser, D. M. Klingeman, B. Raman, L. A. Goodwin, R. Tapia, L. J. Meincke, J. C. Detter, D. C. Bruce, C. S. Han, A. V. Palumbo, R. W. Cottingham, M. Keller, and D. E. Graham.** 2010. Complete genome sequence of the cellulolytic thermophile *Caldicellulosiruptor obsidiansis* OB47^T. J. Bacteriol. **192:**6099-6100.

19. **Eng, J. K., A. L. McCormack, and J. R. Yates.** 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. **5:**976-989.

20. **Gasteiger, E., A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel, and A. Bairoch.** 2003. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. **31:**3784-3788.

21. **Geslin, C., M. Le Romancer, G. Erauso, M. Gaillard, G. Perrot, and D. Prieur.** 2003. PAV1, the first virus-like particle isolated from a hyperthermophilic euryarchaeote, *Pyrococcus abyssi*. J. Bacteriol. **185:**3888-3894.

22. **Giannone, R. J., H. Huber, T. Karpinets, T. Heimerl, U. Küper, R. Rachel, M. Keller, R. L. Hettich, and M. Podar.** 2011. Proteomic characterization of cellular and molecular

JB Accepts published online ahead of print

processes that enable the *Nanoarchaeum equitans-Ignicoccus hospitalis* relationship. PLoS One **6:**22942.

23. **Gibbs, M. D., A. U. Elinder, R. A. Reeves, and P. L. Bergquist.** 1996. Sequencing, cloning and expression of a beta-1,4-mannanase gene, *manA*, from the extremely thermophilic anaerobic bacterium, *Caldicellulosiruptor* Rt8B.4. FEMS Microbiol. Lett. **141:**37-43.

24. **Gibbs, M. D., R. A. Reeves, G. K. Farrington, P. Anderson, D. P. Williams, and P. L. Bergquist.** 2000. Multidomain and multifunctional glycosyl hydrolases from the extreme thermophile *Caldicellulosiruptor* isolate Tok7B.1. Curr. Microbiol. **40:**333-340.

25. **Gibbs, M. D., D. J. Saul, E. Luthi, and P. L. Bergquist.** 1992. The beta-mannanase from "*Caldocellum saccharolyticum*" is part of a multidomain enzyme. Appl. Environ. Microbiol. **58:**3864-3867.

26. **Griffiths-Jones, S., A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy.** 2003. Rfam: an RNA family database. Nucleic Acids Res. **31:**439-441.

27. **Haft, D. H., J. D. Selengut, and O. White.** 2003. The TIGRFAMs database of protein families. Nucleic Acids Res. **31:**371-373.

28. **Hamilton-Brehm, S. D., J. J. Mosher, T. Vishnivetskaya, M. Podar, S. Carroll, S. Allman, T. J. Phelps, M. Keller, and J. G. Elkins.** 2010. *Caldicellulosiruptor obsidiansis* sp. nov., an anaerobic, extremely thermophilic, cellulolytic bacterium isolated from Obsidian Pool, Yellowstone National Park. Appl. Environ. Microbiol. **76:**1014-1020.

29. **Hartley, B. S., N. Hanlon, R. J. Jackson, and M. Rangarajan.** 2000. Glucose isomerase: insights into protein engineering for increased thermostability. Biochim. Biophys. Acta **1543:**294-335.

30. **Himmel, M. E., S. Y. Ding, D. K. Johnson, W. S. Adney, M. R. Nimlos, J. W. Brady, and T. D. Foust.** 2007. Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science **315:**804-807.

JB Accepts published online ahead of print

31. **Hobbie, J. E., R. J. Daley, and S. Jasper.** 1977. Use of nuclepore filters for counting bacteria by fluorescence microscopy. Appl. Environ. Microbiol. **33:**1225-1228.

32. **Huang, C. Y., B. K. Patel, R. A. Mah, and L. Baresi.** 1998. *Caldicellulosiruptor owensensis* sp. nov., an anaerobic, extremely thermophilic, xylanolytic bacterium. Int. J. Syst. Bacteriol. **48:**91-97.

33. **Hugenholtz, P., C. Pitulle, K. L. Hershberger, and N. R. Pace.** 1998. Novel division level bacterial diversity in a Yellowstone hot spring. J. Bacteriol. **180:**366-376.

34. **Hyatt, D., G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11:**119.

35. **Kanehisa, M., S. Goto, Y. Sato, M. Furumichi, and M. Tanabe.** 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. **40:**D109-D114.

36. **Kataeva, I., X.-L. Li, H. Chen, S.-K. Choi, and L. G. Ljungdahl.** 1999. Cloning and sequence analysis of a new cellulase gene encoding CelK, a major cellulosome component of *Clostridium thermocellum*: evidence for gene duplication and recombination. J. Bacteriol. **181:**5288-5295.

37. **Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer.** 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. **305:**567-580.

38. **Kublanov, I. V., A. A. Perevalova, G. B. Slobodkina, A. V. Lebedinsky, S. K. Bidzhieva, T. V. Kolganova, E. N. Kaliberda, L. D. Rumsh, T. Haertle, and E. A. Bonch-Osmolovskaya.** 2009. Biodiversity of thermophilic prokaryotes with hydrolytic activities in hot springs of Uzon Caldera, Kamchatka (Russia). Appl. Environ. Microbiol. **75:**286-291.

JB Accepts published online ahead of print

39. **Kyriacou, A., R. J. Neufeld, and C. R. MacKenzie.** 1989. Reversibility and competition in the adsorption of *Trichoderma reesei* cellulase components. Biotechnol. Bioeng. **33:**631-637.

40. **Lagesen, K., P. Hallin, E. A. Rødland, H.-H. Staerfeldt, T. Rognes, and D. W. Ussery.** 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. **35:**3100-3108.

41. **Lewis, D.** 2010. Functional genomics analysis of extremely thermophilic fermentative microorganisms from the archaeal genus *Pyrococcus* and bacterial genus *Caldicellulosiruptor*. North Carolina State University, Raleigh, NC.

42. **Li, G., P. J. B. Brown, J. X. Tang, J. Xu, E. M. Quardokus, C. Fuqua, and Y. V. Brun.** 2012. Surface contact stimulates the just-in-time deployment of bacterial adhesins. Mol. Microbiol. **83:**41-51.

43. **Li, L., C. J. Stoeckert, and D. S. Roos.** 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. **13:**2178-2189.

44. **Lochner, A., R. J. Giannone, M. Keller, G. Antranikian, D. E. Graham, and R. L. Hettich.** 2011. Label-free quantitative proteomics for the extremely thermophilic bacterium *Caldicellulosiruptor obsidiansis* reveal distinct abundance patterns upon growth on cellobiose, crystalline cellulose, and switchgrass. J. Proteome Res. **10:**5302-5314.

45. **Lochner, A., R. J. Giannone, M. Rodriguez, Jr., M. B. Shah, J. R. Mielenz, M. Keller, G. Antranikian, D. E. Graham, and R. L. Hettich.** 2011. Use of label-free quantitative proteomics to distinguish the secreted cellulolytic systems of *Caldicellulosiruptor bescii* and *Caldicellulosiruptor obsidiansis*. Appl. Environ. Microbiol. **77:**4042-4054.

46. **Lopez-Contreras, A. M., K. Gabor, A. A. Martens, B. A. M. Renckens, P. A. M. Claassen, J. Van Der Oost, and W. M. De Vos.** 2004. Substrate-induced production

JB Accepts published online ahead of print

and secretion of cellulases by *Clostridium acetobutylicum*. Appl. Environ. Microbiol. **70:**5238-5243.

47. **Lowe, T. M., and S. R. Eddy.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:**955-964.

48. **Ludwig, W., K.-H. Schleifer, and W. Whitman.** 2009. Revised Road Map to the Phylum *Firmicutes*. *In* P. de Vos, G. Garrity, D. Jones, N. R. Krieg, W. Ludwig, F. A. Rainey, K.-H. Schleifer, and W. B. Whitman (ed.), Bergey's Manual of Systematic Bacteriology, vol. 3. Springer, New York, USA.

49. **Lykidis, A., K. Mavromatis, N. Ivanova, I. Anderson, M. Land, G. DiBartolo, M. Martinez, A. Lapidus, S. Lucas, A. Copeland, P. Richardson, D. B. Wilson, and N. Kyrpides.** 2007. Genome sequence and analysis of the soil cellulolytic Actinomycete *Thermobifida fusca* YX. J. Bacteriol. **189:**2477-2486.

50. **Lynd, L. R., P. J. Weimer, W. H. van Zyl, and I. S. Pretorius.** 2002. Microbial cellulose utilization: fundamentals and biotechnology. Microbiol. Mol. Biol. Rev. **66:**506-577.

51. **Lynd, L. R., C. E. Wyman, and T. U. Gerngross.** 1999. Biocommodity engineering. Biotechnol. Prog. **15:**777-793.

52. **Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376-380.

JB Accepts published online ahead of print

53. **Markowitz, V. M., I. M. A. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, I. Anderson, A. Lykidis, K. Mavromatis, N. N. Ivanova, and N. C. Kyrpides.** 2010. The integrated microbial genomes system: an expanding comparative analysis resource. Nucleic Acids Res. **38:**D382-D390.

54. **Markowitz, V. M., I. M. A. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, B. Jacob, J. Huang, P. Williams, M. Huntemann, I. Anderson, K. Mavromatis, N. N. Ivanova, and N. C. Kyrpides.** 2012. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res. **40:**D115-D122.

55. **Martinez, D., R. M. Berka, B. Henrissat, M. Saloheimo, M. Arvas, S. E. Baker, J. Chapman, O. Chertkov, P. M. Coutinho, D. Cullen, E. G. Danchin, I. V. Grigoriev, P. Harris, M. Jackson, C. P. Kubicek, C. S. Han, I. Ho, L. F. Larrondo, A. L. de Leon, J. K. Magnuson, S. Merino, M. Misra, B. Nelson, N. Putnam, B. Robbertse, A. A. Salamov, M. Schmoll, A. Terry, N. Thayer, A. Westerholm-Parvinen, C. L. Schoch, J. Yao, R. Barabote, M. A. Nelson, C. Detter, D. Bruce, C. R. Kuske, G. Xie, P. Richardson, D. S. Rokhsar, S. M. Lucas, E. M. Rubin, N. Dunn-Coleman, M. Ward, and T. S. Brettin.** 2008. Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). Nat. Biotechnol. **26:**553-60.

56. **McDonald, W. H., R. Ohi, D. T. Miyamoto, T. J. Mitchison, and J. R. Yates Iii.** 2002. Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. Int. J. Mass Spectrom. **219:**245-251.

57. **Miroshnichenko, M. L., I. V. Kublanov, N. A. Kostrikina, T. P. Tourova, T. V. Kolganova, N. K. Birkeland, and E. A. Bonch-Osmolovskaya.** 2008. *Caldicellulosiruptor kronotskyensis* sp. nov. and *Caldicellulosiruptor hydrothermalis* sp.

JB Accepts published online ahead of print

926 nov., two extremely thermophilic, cellulolytic, anaerobic bacteria from Kamchatka
927 thermal springs. Int. J. Syst. Evol. Microbiol. **58:**1492-1496.
928 58. **Mladenovska, Z., I. M. Mathrani, and B. K. Ahring.** 1995. Isolation and
929 characterization of *Caldicellulosiruptor lactoaceticus* sp. nov., an extremely thermophilic,
930 cellulolytic, anaerobic bacterium. Arch. Microbiol. **163:**223-230.
931 59. **Morris, D. D., M. D. Gibbs, M. Ford, J. Thomas, and P. L. Bergquist.** 1999. Family 10
932 and 11 xylanase genes from *Caldicellulosiruptor* sp. strain Rt69B.1. Extremophiles
933 **3:**103-111.
934 60. **Nolling, J., G. Breton, M. V. Omelchenko, K. S. Makarova, Q. Zeng, R. Gibson, H. M.**
935 **Lee, J. Dubois, D. Qiu, J. Hitti, Y. I. Wolf, R. L. Tatusov, F. Sabathe, L. Doucette-**
936 **Stamm, P. Soucaille, M. J. Daly, G. N. Bennett, E. V. Koonin, and D. R. Smith.** 2001.
937 Genome sequence and comparative analysis of the solvent-producing bacterium
938 *Clostridium acetobutylicum*. J. Bacteriol. **183:**4823-4838.
939 61. **Olson, D. G., S. A. Tripathi, R. J. Giannone, J. Lo, N. C. Caiazza, D. A. Hogsett, R. L.**
940 **Hettich, A. M. Guss, G. Dubrovsky, and L. R. Lynd.** 2010. Deletion of the Cel48S
941 cellulase from *Clostridium thermocellum*. Proc. Natl. Acad. Sci. USA **107:**17727-17732.
942 62. **Onyenwoke, R. U., Y. J. Lee, S. Dabrowski, B. K. Ahring, and J. Wiegel.** 2006.
943 Reclassification of *Thermoanaerobium acetigenum* as *Caldicellulosiruptor acetigenus*
944 comb. nov. and emendation of the genus description. Int. J. Syst. Evol. Microbiol.
945 **56:**1391-1395.
946 63. **Otter, D. E., P. A. Munro, G. K. Scott, and R. Geddes.** 1989. Desorption of
947 *Trichoderma reesei* cellulase from cellulose by a range of desorbents. Biotechnol.
948 Bioeng. **34:**291-298.
949 64. **Ozdemir, I., S. E. Blumer-Schuette, and R. M. Kelly.** 2012. S-layer homology (SLH)
950 domain proteins Csac_0678 and Csac_2722 implicated in plant polysaccharide

JB Accepts published online ahead of print

deconstruction by the extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus*. Appl. Environ. Microbiol. **78:**768-777.

65. **Pati, A., N. N. Ivanova, N. Mikhailova, G. Ovchinnikova, S. D. Hooper, A. Lykidis, and N. C. Kyrpides.** 2010. GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. Nat. Methods **7:**455-457.

66. **Pelicic, V.** 2008. Type IV pili: e pluribus unum? Mol. Microbiol. **68:**827-837.

67. **Perret, S., H. Maamar, J.-P. Belaich, and C. Tardif.** 2004. Use of antisense RNA to modify the composition of cellulosomes produced by *Clostridium cellulolyticum*. Mol. Microbiol. **51:**599-607.

68. **Petersen, T. N., S. Brunak, G. v. Heijne, and H. Nielsen.** 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods **8:**785-786.

69. **Punta, M., P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn.** 2011. The Pfam protein families database. Nucleic Acids Res. **40:**D290-D301.

70. **Rainey, F. A., A. M. Donnison, P. H. Janssen, D. Saul, A. Rodrigo, P. L. Bergquist, R. M. Daniel, E. Stackebrandt, and H. W. Morgan.** 1994. Description of *Caldicellulosiruptor saccharolyticus* gen. nov., sp. nov: an obligately anaerobic, extremely thermophilic, cellulolytic bacterium. FEMS Microbiol. Lett. **120:**263-266.

71. **Raman, B., C. Pan, G. B. Hurst, M. Rodriguez, Jr., C. K. McKeown, P. K. Lankford, N. F. Samatova, and J. R. Mielenz.** 2009. Impact of pretreated switchgrass and biomass carbohydrates on *Clostridium thermocellum* ATCC 27405 cellulosome composition: a quantitative proteomic analysis. PLoS One **4:**e5271.

72. **Richter, M., and R. Rosselló-Móra.** 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. USA **106:**19126-19131.

JB Accepts published online ahead of print

73. **Rincon, M. T., T. Cepeljnik, J. C. Martin, Y. Barak, R. Lamed, E. A. Bayer, and H. J. Flint.** 2007. A novel cell surface-anchored cellulose-binding protein encoded by the sca gene cluster of *Ruminococcus flavefaciens*. J. Bacteriol. **189:**4774-4783.

74. **Sabathe, F., A. Belaich, and P. Soucaille.** 2002. Characterization of the cellulolytic complex (cellulosome) of *Clostridium acetobutylicum*. FEMS Microbiol. Lett. **217:**15-22.

75. **Saier, M. H.** 2000. A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol. Mol. Biol. Rev. **64:**354-411.

76. **Saul, D. J., L. C. Williams, R. A. Grayling, L. W. Chamley, D. R. Love, and P. L. Bergquist.** 1990. *celB*, a gene coding for a bifunctional cellulase from the extreme thermophile "*Caldocellum saccharolyticum*". Appl. Environ. Microbiol. **56:**3117-3124.

77. **Schneider, E.** 2001. ABC transporters catalyzing carbohydrate uptake. Res. Microbiol. **152:**303-310.

78. **Suen, G., P. J. Weimer, D. M. Stevenson, F. O. Aylward, J. Boyum, J. Deneke, C. Drinkwater, N. N. Ivanova, N. Mikhailova, O. Chertkov, L. A. Goodwin, C. R. Currie, D. Mead, and P. J. Brumm.** 2011. The complete genome sequence of *Fibrobacter succinogenes* S85 reveals a cellulolytic and metabolic specialist. PLoS ONE **6:**e18814.

79. **Sunna, A.** 2010. Modular organisation and functional analysis of dissected modular beta-mannanase CsMan26 from *Caldicellulosiruptor* Rt8B.4. Appl. Microbiol. Biotechnol. **86:**189-200.

80. **Sunna, A., M. D. Gibbs, and P. L. Bergquist.** 2001. Identification of novel beta-mannan- and beta-glucan-binding modules: evidence for a superfamily of carbohydrate-binding modules. Biochem. J. **356:**791-798.

81. **Svetlichnyi, V. A., T. P. Svetlichnaya, N. A. Chernykh, and G. A. Zavarzin.** 1990. *Anaerocellum thermophilum* gen. nov sp. nov: an extremely thermophilic cellulolytic eubacterium isolated from hot springs in the Valley of Geysers. Microbiology **59:**598-604.

JB Accepts published online ahead of print

1002 82. **Tabb, D. L., W. H. McDonald, and J. R. Yates.** 2002. DTASelect and Contrast: Tools
1003 for assembling and comparing protein iIdentifications from shotgun proteomics. J.
1004 Proteome Res. **1:**21-26.
1005 83. **Tamura, K., J. Dudley, M. Nei, and S. Kumar.** 2007. MEGA4: Molecular Evolutionary
1006 Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. **24:**1596-1599.
1007 84. **Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V.**
1008 **Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S.**
1009 **Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale.** 2003.
1010 The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:**41.
1011 85. **Te'O, V. S., D. J. Saul, and P. L. Bergquist.** 1995. *celA*, another gene coding for a
1012 multidomain cellulase from the extreme thermophile *Caldocellum saccharolyticum*. Appl.
1013 Microbiol. Biotechnol. **43:**291-296.
1014 86. **Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V.**
1015 **Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. DeBoy, T. M. Davidsen, M.**
1016 **Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram,**
1017 **W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A.**
1018 **Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H.**
1019 **Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R.**
1020 **Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L.**
1021 **Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser.** 2005. Genome analysis of
1022 multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial
1023 "pan-genome". Proc. Natl. Acad. Sci. USA **102:**13950-13955.
1024 87. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the
1025 sensitivity of progressive multiple sequence alignment through sequence weighting,
1026 position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673-
1027 4680.

JB Accepts published online ahead of print

1028 88. **Tormo, J., R. Lamed, A. J. Chirino, E. Morag, E. A. Bayer, Y. Shoham, and T. A.**
1029 **Steitz.** 1996. Crystal structure of a bacterial family-III cellulose-binding domain: a
1030 general mechanism for attachment to cellulose. EMBO J. **15:**5739-5751.
1031 89. **van de Werken, H. J., M. R. Verhaart, A. L. VanFossen, K. Willquist, D. L. Lewis, J.**
1032 **D. Nichols, H. P. Goorissen, E. F. Mongodin, K. E. Nelson, E. W. van Niel, A. J.**
1033 **Stams, D. E. Ward, W. M. de Vos, J. van der Oost, R. M. Kelly, and S. W. Kengen.**
1034 2008. Hydrogenomics of the extremely thermophilic bacterium *Caldicellulosiruptor*
1035 *saccharolyticus*. Appl. Environ. Microbiol. **74:**6720-6729.
1036 90. **VanFossen, A. L., I. Ozdemir, S. L. Zelin, and R. M. Kelly.** 2011. Glycoside hydrolase
1037 inventory drives plant polysaccharide deconstruction by the extremely thermophilic
1038 bacterium *Caldicellulosiruptor saccharolyticus*. Biotechnol. Bioeng. **108:**1559-1569.
1039 91. **VanFossen, A. L., M. R. Verhaart, S. M. Kengen, and R. M. Kelly.** 2009. Carbohydrate
1040 utilization patterns for the extremely thermophilic bacterium *Caldicellulosiruptor*
1041 *saccharolyticus* reveal broad growth substrate preferences. Appl. Environ. Microbiol.
1042 **75:**7718-7724.
1043 92. **Wang, Z.-W., S.-H. Lee, J. G. Elkins, and J. L. Morrell-Falvey.** 2011. fSpatial and
1044 temporal dynamics of cellulose degradation and biofilm formation by *Caldicellulosiruptor*
1045 *obsidiansis* and *Clostridium thermocellum*. AMB Express **1:**30.
1046 93. **Washburn, M. P., D. Wolters, and J. R. Yates.** 2001. Large-scale analysis of the yeast
1047 proteome by multidimensional protein identification technology. Nat. Biotechnol. **19:**242-
1048 247.
1049 94. **Watson, B. J., H. Zhang, A. G. Longmire, Y. H. Moon, and S. W. Hutcheson.** 2009.
1050 Processive endoglucanases mediate degradation of cellulose by *Saccharophagus*
1051 *degradans*. J. Bacteriol. **191:**5697-5705.
1052 95. **Weiner, R. M., L. E. Taylor, B. Henrissat, L. Hauser, M. Land, P. M. Coutinho, C.**
1053 **Rancurel, E. H. Saunders, A. G. Longmire, H. Zhang, E. A. Bayer, H. J. Gilbert, F.**

JB Accepts published online ahead of print

Larimer, I. B. Zhulin, N. A. Ekborg, R. Lamed, P. M. Richardson, I. Borovok, and S. Hutcheson. 2008. Complete genome sequence of the complex carbohydrate-degrading marine bacterium, *Saccharophagus degradans* strain 2-40T. PLoS Genet. **4:**e1000087.

96. **Xie, G., D. C. Bruce, J. F. Challacombe, O. Chertkov, J. C. Detter, P. Gilna, C. S. Han, S. Lucas, M. Misra, G. L. Myers, P. Richardson, R. Tapia, N. Thayer, L. S. Thompson, T. S. Brettin, B. Henrissat, D. B. Wilson, and M. J. McBride.** 2007. Genome sequence of the cellulolytic gliding bacterium *Cytophaga hutchinsonii*. Appl. Environ. Microbiol. **73:**3536-3546.

97. **Yang, S. J., I. Kataeva, S. D. Hamilton-Brehm, N. L. Engle, T. J. Tschaplinski, C. Doeppke, M. Davis, J. Westpheling, and M. W. Adams.** 2009. Efficient degradation of lignocellulosic plant biomass, without pretreatment, by the thermophilic anaerobe "*Anaerocellum thermophilum*" DSM 6725. Appl. Environ. Microbiol. **75:**4762-4769.

98. **Yang, S. J., I. Kataeva, J. Wiegel, Y. Yin, P. Dam, Y. Xu, J. Westpheling, and M. W. Adams.** 2010. Reclassification of '*Anaerocellum thermophilum*' as *Caldicellulosiruptor bescii* strain DSM 6725T sp. nov. Int. J. Syst. Evol. Microbiol. **60:**2011-2015.

99. **York, W. S., H. van Halbeek, A. G. Darvill, and P. Albersheim.** 1990. Structural analysis of xyloglucan oligosaccharides by 1H-n.m.r. spectroscopy and fast-atom-bombardment mass spectrometry. Carbohydrate Res. **200:**9-31.

100. **Zdobnov, E. M., and R. Apweiler.** 2001. InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17:**847-848.

101. **Zybailov, B., A. L. Mosley, M. E. Sardiu, M. K. Coleman, L. Florens, and M. P. Washburn.** 2006. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. J. Proteome Res. **5:**2339-2347.

JB Accepts published online ahead of print

Table 1. General *Caldicellulosiruptor* genome characteristics

Species	Culture Accession	Genome Accession	Genome Size (Mb)	Protein Coding	G+C Content (%)	Ref(s).
C. bescii	DSM-6725	CP001393	2.93	2776	35.2	(16)
C. hydrothermalis	DSM-18901	CP002219	2.77	2625	36.5	(10)
C. kristjanssonii	DSM-12137	CP002326	2.80	2648	36.0	(10)
C. kronotskyensis	DSM-18902	CP002330	2.84	2583	35.0	(10)
C. lactoaceticus	DSM-9545	CP003001	2.62	2492	36.1	(10)
C. obsidiansis	ATCC BAA-2073	CP002164	2.53	2331	35.2	(28)
C. owensensis	DSM-13100	CP002216	2.43	2264	35.4	(10)
C. saccharolyticus	DSM-8903	CP000679	2.97	2760	35.2	(89)

Table 2. Carbohydrate-related domains and transporter inventory

Species	Number of ORFs with: [a]					Total [b]	SigP [c]	C.T. [d]
	GH	CBM	PL	CE	GT			
C. bescii	52	22	4	7	29	68	23	20
C. hydrothermalis	62	17	1	6	28	74	15	39
C. kristjanssonii	37	15	3	5	31	48	14	15
C. kronotskyensis	77	28	4	9	35	93	32	28
C. lactoaceticus	44	18	4	4	27	54	17	12
C. obsidiansis	47	18	2	5	29	59	16	20
C. owensensis	51	16	4	8	31	67	19	18
C. saccharolyticus	59	17	1	6	30	70	17	25

[a] GH, glycoside hydrolase; CBM, carbohydrate binding module; PL, polysaccharide lyase; CE, carbohydrate esterase; GT, glycosyl transferase. Numbers of carbohydrate-active protein domains were retrieved from the CAZy database, http://www.cazy.org (13).

[b] Indicates the total number of ORFs that contain either glycoside hydrolases, carbohydrate-binding modules, polysaccharide lyases or carbohydrate esterases.

[c] SigP, signal peptide

[d] C.T., ATP binding cassette (ABC) carbohydrate transporter

JB Accepts published online ahead of print

Table 3. *Caldicellulosiruptor* adhesins located downstream of type IV pilus gene clusters 1086

MCL[a]	Gene Locus	Protein Properties					Protein Abundance[e]		
		AA	Mol Wt (kDa)[b]	pI[b]	SigP[c]	TMD[d]	SB	SN	WC
1820	Athe_1871	642	70.1	5.37	N	Y	2.26E-03	1.07E-04	3.16E-06
	Calkr_0826	634	68.9	8.3	N	Y	5.37E-03	8.54E-04	7.07E-05
	Calkro_0844	642	69.6	5.18	N	Y	4.41E-03	8.77E-06	2.83E-06
	Calla_1507	634	69.0	8.02	N	Y	9.80E-03	2.32E-03	5.45E-04
	COB47_1678	642	69.8	5.13	N	Y	n/a[f]	n/a	n/a
	Csac_1073	642	69.9	5.13	N	Y	4.29E-03	2.49E-03	5.99E-05
1653	Athe_1870	649	70.3	6.37	N	Y	2.04E-03	1.05E-05	3.13E-06
	Calhy_0908	638	71.0	5.8	Y	Y	nd[g]	Nd	nd
	Calkr_0827	622	68.9	5.7	Y	N	nd	Nd	nd
	Calkro_0845	649	70.5	7.02	N	Y	6.95E-04	8.67E-06	2.80E-06
	Calla_1506	628	69.5	6.01	N	Y	nd	Nd	nd
	COB47_1675	649	70.3	5.72	N	Y	n/a	n/a	n/a
	Csac_1074	649	70.4	5.58	N	Y	1.84E-04	1.75E-05	4.80E-05
	Calow_1589	667	71.7	9.23	Y	N	4.70E-03	1.60E-02	2.07E-04
	Calow_1590	900	100.2	5.12	N	Y	2.93E-04	7.64E-04	7.79E-05

1087

1099 [a] OrthoMCL group numbers for orthologous *Caldicellulosiruptor* proteins (see Data Set 1). No orthologous groups were assigned to the two
1100 proteins detected from *C. owensensis*

1101 [b] Predictions for Mol Wt and isoelectric point (pI) used the ExPASy Compute pI/Mw tool (20, [http://web.expasy.org/compute_pi/])

1102 [c] SigP, signal peptide; predicted using SignalP (68, [http://www.cbs.dtu.dk/services/SignalP/])

1103 [d] TMD, transmembrane domain; predicted using the TMHMM server (37, [http://www.cbs.dtu.dk/services/TMHMM/])

1104 [e] Protein abundance is reported as NSAF for each fraction screened: SB, substrate-bound; SN, supernatant; WC, whole cell lysate

1105 [f] n/a, protein abundance not available

1106 [g] nd, not detected in protein fractions using proteomics

1107

JB Accepts published online ahead of print

1108 **FIGURE LEGENDS**

1109

1110 **Figure 1. Biogeography of sequenced *Caldicellulosiruptor* species.**

1111 **(A)** Global distribution of cellulolytic and weakly cellulolytic species. Squares denote cellulolytic
1112 species and circles denote weakly cellulolytic species. Colors shading the shapes indicate
1113 common isolation locations.

1114 **(B)** Phylogenetic tree using 16S ribosomal RNA sequences from sequenced species plus
1115 related outliers. MEGA4 was used to calculate distances and built the phylogenetic tree (83).

1116 **(C)** Phylogenomic heat plot using ANI as a measure of relatedness. Red indicates closer related
1117 species while gray to blue indicates more distantly related species, the percentage of homology
1118 for each pairing of species can be found in Table S1. Abbreviated species names follows the
1119 assigned locus tags and are as follows: Cbes, *C. bescii*; Calhy, *C. hydrothermalis*; Calkr, *C.*
1120 *kristjanssonii*; Calkro, *C. kronotskyensis*; Calla, *C. lactoaceticus*; COB47, *C. obsidiansis*; Calow,
1121 *C. owensensis*; Csac, *C. saccharolyticus*; Cthe, *Clostridium thermocellum*; Dtur, *Dictyoglomus*
1122 *turgidum*; Teth39, *Thermoanaerobacter pseudethanolicus*.

1123

1124 **Figure 2. Capacity for crystalline cellulose deconstruction and growth of**
1125 ***Caldicellulosiruptor* species on complex substrates.**

1126 **(A)** Cell density (cells/ml) for each species after 24 hours of growth on: Avicel, microcrystalline
1127 cellulose; Xylan, birchwood xylan; SWG, acid-treated switchgrass. Standard deviations are
1128 equal to one third or less of cell density. Abbreviations follow the assigned locus numbering
1129 system and are as follows: C, control; 1, Cbes, *C. bescii*; 2, Calhy, *C. hydrothermalis*; 3, Calkr,
1130 *C. kristjanssonii*; 4, Calkro, *C. kronotskyensis*; 5, Calla, *C. lactoaceticus*; 6, COB47, *C.*
1131 *obsidiansis*; 7, Calow, *C. owensensis*; 8, Csac, *C. saccharolyticus*.

1132

JB Accepts published online ahead of print

(B) Microbial deconstruction of Whatman #1 filter paper during growth. Fibers released from the substrate at the bottom of the Hungate culture tube are indicative of enzymatic activity against crystalline cellulose.

Figure 3. Core carbohydrate-active enzymes and carbohydrate binding motif-containing proteins from all eight *Caldicellulosiruptor* species.

(A) Core glycoside hydrolases (GH), polysaccharide lyases (PL), carbohydrate esterases (CE) and carbohydrate binding motifs (CBM). Numbers refer to protein families established and curated by CAZy (13, [http://www.cazy.org])

(B) Core glycoside hydrolases for strongly cellulolytic species, dashed parentheses indicate gene truncations in †, *C. lactoaceticus*, and ‡, *C. saccharolyticus*; solid parenthesis indicates an additional CBM family 3 domain in *C. lactoaceticus*.

Figure 4. **Gene clusters of CBM3-containing glycoside hydrolases**. Locus tags are as follows: Cbes, Athe_1867-Athe_1853; Calkr, Calkr_0017, Calkr_1847~Calkr_1849, Calkr_2455, Calkr_2522; Calkro, Calkro_0850~Calkro_0864; Calla,. Calla_0015~Calla_0018, Calla_1251~Calla_1249, Calla_2311~Calla_2308, Calla_2385; COB47, COB47_1673~COB47_1662; Csac, Csac_1076~Csac_1085. CBM3 modules are denoted by white diamonds, dashed means that orthologs posses the CBM3 module; green ovals, GH5; red ovals, GH9; lilac ovals, GH10; blue ovals, GH44; grey ovals, GH48; purple ovals, GH74; blue rectangles, polysaccharide lyase; beige arrow, GT39; brown rectangle, AraC transcriptional regulator.

Figure 5. Extracellular, cell membrane-bound proteins involved in microbe-cellulose interactions of strongly cellulolytic *Caldicellulosiruptor*.

Highlighted proteins were detected in the supernatant or substrate-bound proteome. Proteins found enriched in the substrate proteome are shaded red, those enriched in the supernatant are shaded in green, proteins shaded blue indicate enrichment in the cell lysate. Noted proteins shaded in grey were detected in all three protein fractions and were not determined to be enriched in one fraction over another. Numbers in parentheses above proteins are nominal labels given to orthologous families of proteins as determined by the OrthoMCL program (43). Exact locus tag numbers for each orthologous protein family are found in Dataset 1 and NSAF for each MCL group are found in Dataset 2.

JB Accepts published online ahead of print

JB Accepts published online ahead of print

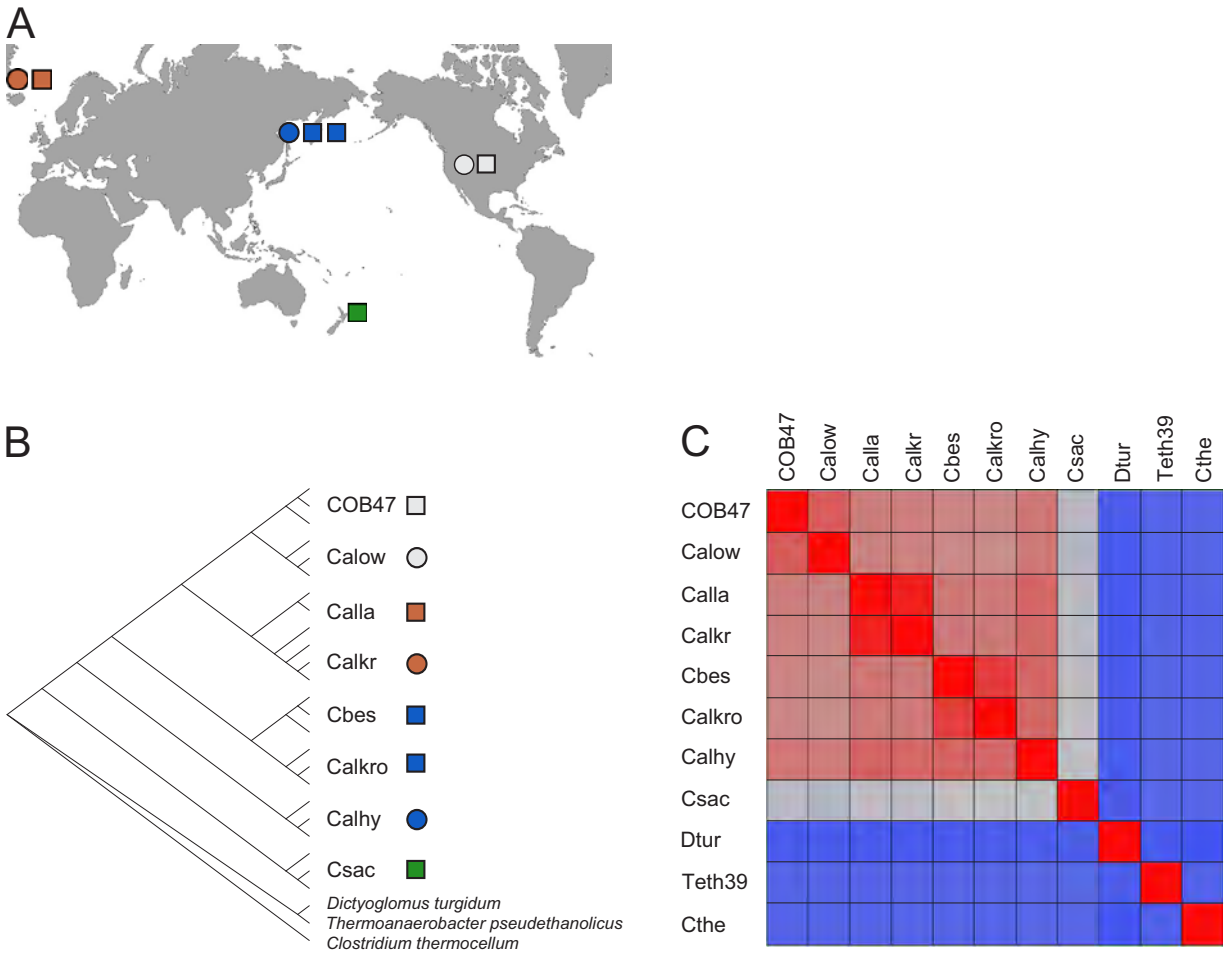

Figure 1

JB Accepts published online ahead of print

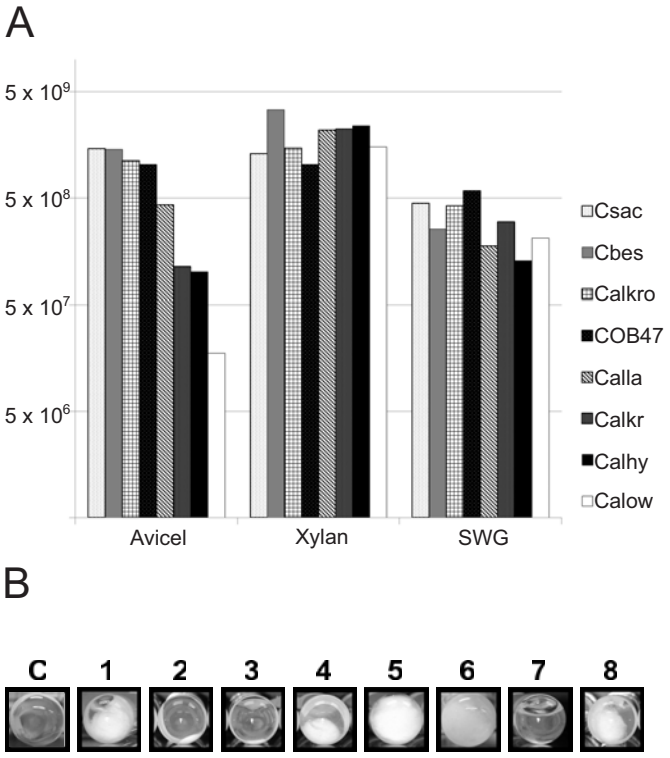

Figure 2

JB Accepts published online ahead of print

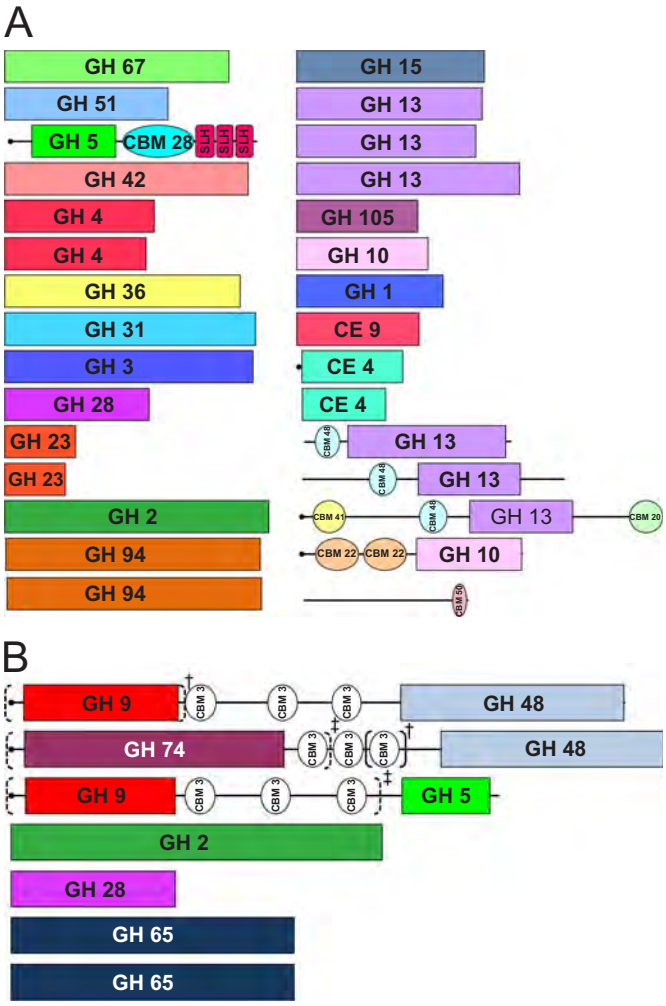

Figure 3

JB Accepts published online ahead of print

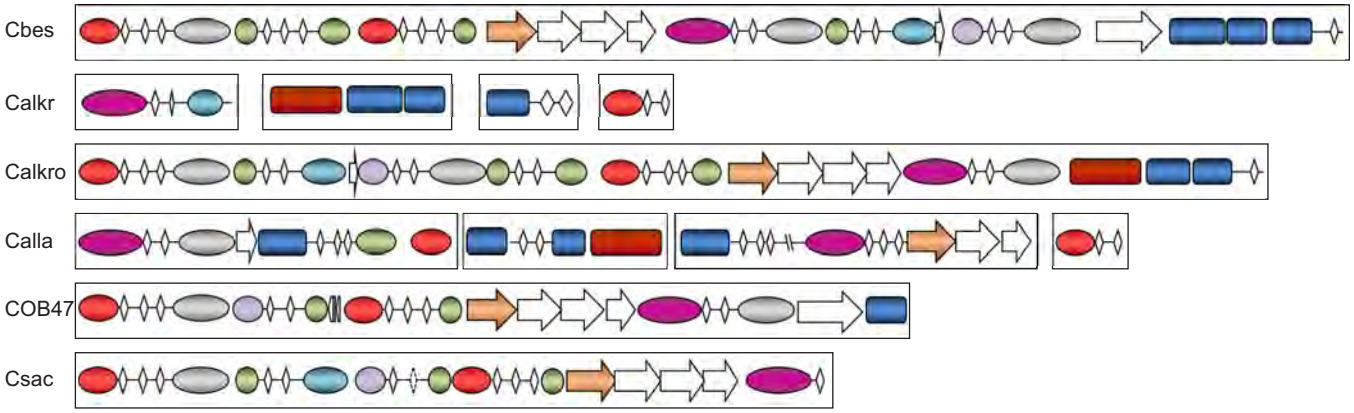

Figure 4

JB Accepts published online ahead of print

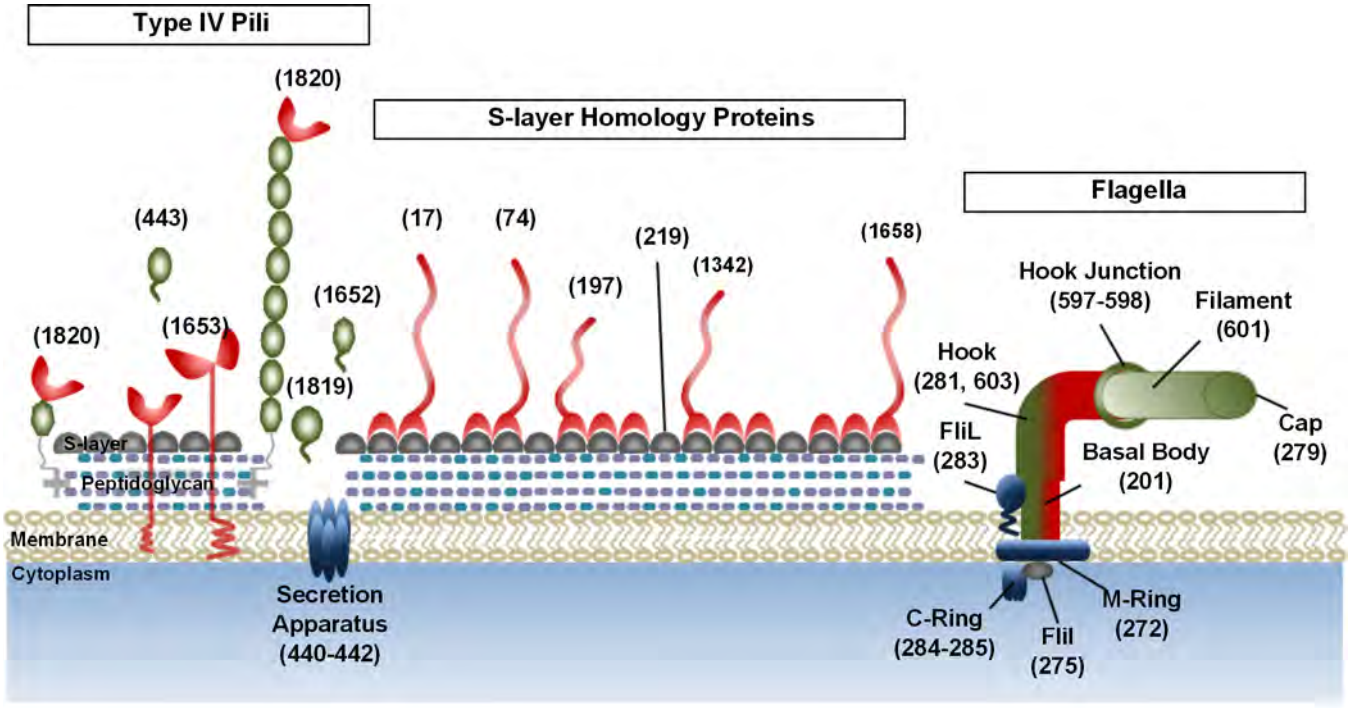

Figure 5