3240–3254 Nucleic Acids Research, 2011, Vol. 39, No. 8 Published online 11 January 2011
doi:10.1093/nar/gkq1281

Insights into plant biomass conversion from the genome of the anaerobic thermophilic bacterium *Caldicellulosiruptor bescii* DSM 6725

Phuongan Dam[1,2,3], Irina Kataeva[2,3], Sung-Jae Yang[2,3], Fengfeng Zhou[1,2,3], Yanbin Yin[1,2,3], Wenchi Chou[1,3], Farris L. Poole II[2,3], Janet Westpheling[3,4], Robert Hettich[3], Richard Giannone[3], Derrick L. Lewis[3,5], Robert Kelly[3,5], Harry J. Gilbert[2,6], Bernard Henrissat[7], Ying Xu[1,2,3,*] and Michael W. W. Adams[2,3,*]

[1]Institute of Bioinformatics, [2]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602, [3]BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, TN 37831, [4]Department of Genetics, University of Georgia, Athens, GA 30602, [5]Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC 27695, [6]Complex Carbohydrate Research Center, University of Georgia, Georgia, Athens, GA 30602, USA and [7]Architecture et Fonction des Macromolécules Biologiques, CNRS and Universities of Aix-Marseille I & II, 13288 Marseille, France

Received August 8, 2010; Revised and Accepted November 25, 2010

ABSTRACT

***Caldicellulosiruptor bescii* DSM 6725 utilizes various polysaccharides and grows efficiently on untreated high-lignin grasses and hardwood at an optimum temperature of ~80°C. It is a promising anaerobic bacterium for studying high-temperature biomass conversion. Its genome contains 2666 protein-coding sequences organized into 1209 operons. Expression of 2196 genes (83%) was confirmed experimentally. At least 322 genes appear to have been obtained by lateral gene transfer (LGT). Putative functions were assigned to 364 conserved/hypothetical protein (C/HP) genes. The genome contains 171 and 88 genes related to carbohydrate transport and utilization, respectively. Growth on cellulose led to the up-regulation of 32 carbohydrate-active (CAZy), 61 sugar transport, 25 transcription factor and 234 C/HP genes. Some C/HPs were overproduced on cellulose or xylan, suggesting their involvement in polysaccharide conversion. A unique feature of the genome is enrichment with genes encoding multi-modular, multi-functional CAZy proteins organized into one large cluster, the products of which are proposed to act synergistically on different components of plant cell walls and to aid the ability of *C. bescii* to convert plant biomass. The high duplication of CAZy domains coupled with the ability to acquire foreign genes by LGT may have allowed the bacterium to rapidly adapt to changing plant biomass-rich environments.**

INTRODUCTION

Lignocellulosic plant biomass is the most abundant renewable alternative to petroleum as a source of fuel (1). It consists mainly of cellulose and hemicellulose in combination with up to 20% lignin. Biological conversion of this chemically and physically complex material, represents a major challenge (2,3). Expensive thermal and chemical pretreatments are needed to decrease its recalcitrance and expose the polysaccharides to carbohydrate-active enzymes (CAZy) and carbohydrate-binding modules (CBMs) that help destroy the plant cell walls (4,5). Despite intensive studies, many aspects of microbial and enzymatic biomass-to-biofuel conversion are still not understood. Thermophilic anaerobic bacteria hold great promise as they display higher bioconversion rates, minimize the risk of contamination, facilitate product recovery and synthesize highly thermostable enzymes (6,7). However, only a relatively small number of anaerobic thermophiles are able to convert crystalline cellulose into soluble fermentable sugars, and only a few of them are able to metabolize simultaneously the hexose and

*To whom correspondence should be addressed. Tel: +1 706 542 2060; Fax: +1 706 542 0229; Email: adams@bmb.uga.edu
Correspondence may also be addressed to Ying Xu. Tel: +1 706 542 9779; Fax: +1 706 542 9751; Email: xyn@bmb.uga.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

pentose sugars that are produced from cellulose and hemicellulose, respectively (1,8).

One of the best studied of the cellulolytic microbes is *Clostridium thermocellum,* which grows optimally at 60°C (9). It produces ethanol and is being used for the consolidated bioprocessing of plant biomass (6,8–10). Its cellulolytic system is a large multi-protein complex called the cellulosome, the enzymatic components of which act synergistically to degrade crystalline cellulose (9–11). The recent availability of genetic systems in *C. thermocellum* and a related thermophile (12) provides a much needed tool to investigate the mechanisms of cellulose degradation. Several members of the genus *Caldicellulosiruptor* are able to degrade cellulose at even higher temperatures (up to 90°C) and they also utilize pentose sugars (13–18). The genomes of *C. saccharolyticus* DSM 8903 (19) and *C. bescii* DSM 6725 have been sequenced (20) and some CAZy enzymes have been purified and characterized from both species (21,22). Representatives of this genus have potential utility in biomass-to-sugars conversion processes but more comprehensive studies are needed to understand the degradative mechanisms involved.

Caldicellulosiruptor bescii grows at temperatures up to 90°C and is the most thermophilic bacterium capable of growth on crystalline cellulose (16). It also utilizes xylan, pectin and starch and is also able to grow efficiently on untreated plant biomass with high lignin content (14,16). The bacterium is capable of using cellulose and xylan simultaneously. Its ability to grow on the hardwood poplar is of particular interest as this hardwood can be genetically manipulated to potentially decrease recalcitrance (23). For example, one transgenic poplar line overexpressing xyloglucanase is less recalcitrant to cellulolytic enzymes (24). In the present article we analyze the genome of *C. bescii* with a particular focus on genes encoding enzymes involved in plant biomass conversion. We also present transcriptomic and proteomic data and compare its genome with those of other anaerobic thermophiles, including its close relative *C. saccharolyticus*. This analysis will contribute to our understanding of plant biomass conversion at extreme temperatures and will provide a genetic basis for the plant biomass-degrading properties displayed by this remarkable organism.

MATERIALS AND METHODS

Growth of microorganism

Caldicellulosiruptor bescii strain DSM 6725 was obtained from the German Culture Collection (www.dsmz.de/index.htm). The organism was grown in the 516 medium as previously described (14) except that vitamin and trace mineral solutions were modified. Medium composition and growth conditions are given in the Supplementary Data.

Scanning electron microscopy (SEM) was performed as described elsewheree (25) (see also Supplementary Data).

DNA microarrays

RNA extraction and purification was carried out as described previously (26). RNA samples were converted to fluorescence-labeled cDNA and hybridized to a whole-genome *C. bescii* microarray according to the procedures previously described (27). Additional information is provided in the Supplementary Data.

Fractionation of *C. bescii* cell extract

Extracellular protein (ExtP) fractions were prepared from 1 l cultures grown for 24 h on different substrates. The residual insoluble substrates (if present) were removed by decantation. The cells and ExtP fractions were separated by centrifugation. The ExtP was filtered through a 0.2 μm membrane, concentrated using a 10 kDa membrane and dialyzed against 50 mM NH_4HCO_3, pH 8.0. To obtain intracellular protein (IntP) and membrane protein (MemP) fractions, samples were prepared from the sedimented cells (Supplementary Data for more details).

Proteomics

Samples for tandem mass-spectrometry were prepared as described earlier (28). Fragmentation spectra (MS/MS) obtained from each sample were searched against the DSM 6725 proteome using SEQUEST (29) and filtered using DTASelect (30). Filter levels were set at +1s 1.8, +2s 2.5 and +3s 3.5 to obtain a false-discovery rate of <5% at protein level. Additionally, a minimum of two unique peptides per locus were required in order to identify the protein (see Supplementary Data for more details).

Genome analysis

The genome sequence of *C. bescii* was determined by the Joint Genome Institute (JGI) (20) and the annotated version was downloaded from the NCBI genome database (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). Clusters of orthologous genes (COG)-based functional assignment (http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html), extracellular proteins, transmembrane helices (http://www.cbs.dtu.dk/services/TMHMM/) and insertion sequence (IS) elements (31) were predicted as described in the Supplementary Data.

Predictions of operons, CAZy-related proteins, transporters and transcription factors

Operons were predicted by our previously published method (32), which was ranked as the best available operon prediction program by an independent study (33). Carbohydrate-active enzymes were searched for using BLAST- and HMM-based tools and sequence libraries used for the updates of the CAZy database [http://www.cazy.org; (34)]. Further details are included in the Supplementary Data. Prediction of transporters and transcription factors (TFs) was carried out as described in the Supplementary Data.

Functional annotation of proteins

The KEGG assignment (http://www.genome.jp/kegg/) and the COG groups (http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html) were downloaded from the databases. Upon analysis, we found that 71–88% of operonic

gene pairs belong to the same KEGG pathways, whereas 48–75% of gene pairs that are not predicted to be in the same operons but having the intergenic distance between gene pairs <100 bp are in the same KEGG pathway (Supplementary Table S1). Our analysis suggested that operonic gene pairs and gene pairs with short intergenic distance are more likely to be functioned in the same pathway. Using this approach, we assigned the associated function for a hypothetical protein, if this protein is predicted to be in the same operon with genes assigned to KEGG pathway, or if the gene is near another gene with annotated function.

Microarray data analysis

To identify the genes expressed in *C. bescii* cells grown on glucose or cellulose (filter paper), we compared the gene expression profiles of *C. bescii* genes and their homologs in *Escherichia coli* K12. The data set of *E. coli* K12 grown on various carbon sources (GSE2037) was downloaded from NCBI, and we identified 278 genes whose expression was reduced when not growing on glucose, using SAM with the cut-off *P*-value of 0.05. Mapping this gene set to the *C. bescii* genome (cut-off *e*-value of 1e-20) results in 206 homologs termed glucose-related genes. The log-likelihood ratio at intensity i was calculated as $\ln(f_{1i}/f_{wi})$, where f_{1i} and f_{wi} are the respective frequencies of glucose-related genes and all genes having the probe intensity i. The cut-off intensity to consider that a gene is expressed was chosen so that $\ln(f_{1i}/f_{wi}) = 0$.

RESULTS AND DISCUSSION

General features and comparative genomics

The *C. bescii* genome contains a 2 919 718 bp circular chromosome with 35.2% GC content and is slightly smaller than the size of the average bacterial genome (3.3 Mb; www.ncbi.nlm.nih.gov/genomes/lproks.cgi). It contains two native circular plasmids termed AX710673 pBAL (8294 bp, 38.5% GC) and AX710687 pBAS2 (3653 bp, 42.9% GC), both of which were isolated and sequenced previously (35). The sequence of AX710687 pBAS2 reported here is identical but that of AX710673 pBAL has 8 deletions, 11 insertions and 7 mismatches (Supplementary Figure S2). AX710673 and AX710687 encode eight and four open reading frames (ORFs), respectively. AX710687 encodes exclusively uncharacterized proteins, while AX710673 encodes two putative regulators and two proteins involved in nucleic acid metabolism.

The chromosome is predicted to contain 2654 protein coding genes. Their arrangement on the two strands suggests that it has two equal replicores with a positive correlation between the direction of transcription and replication (Figure 1). The 16 S RNA sequences confirmed that what was formerly termed *Anaerocellum thermophilum* is a member of the phylum *Firmicutes*, class *Clostridia*, order *Clostridiales* and that it should be classified in the genus *Caldicellulosiruptor* (15). *C. hydrothermalis* and *C. kronotskiensis* are the closest known relatives of *C. bescii* and *C. saccharolyticus* DSM 8903 is the closest relative with a sequenced genome (15,19).

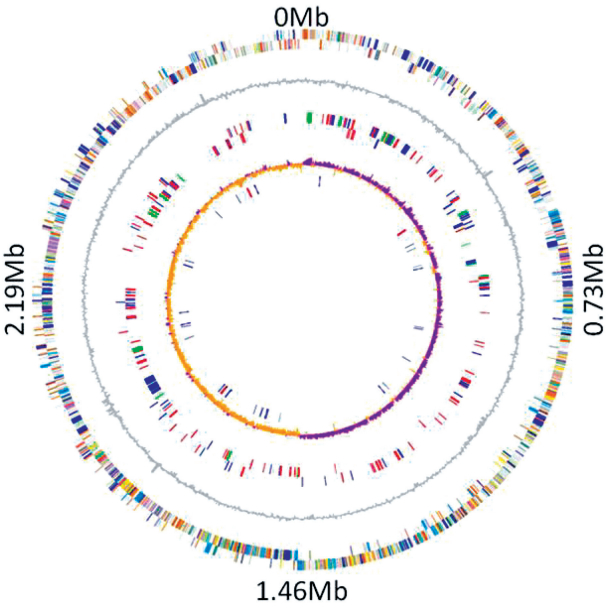

Figure 1. Diagram of *C. bescii* chromosome. From outside to inside, the circles show (i) COG categories (two circles), (ii) mean centered GC content of the genome, (iii) genes (two circles) with functions related to CAZy (green), sugar transporters (red) and cell-adhesion (blue), (iv) GC skew plot(orange-purple circle) and (v) RNA genes (ribosomal in red, tRNA in blue and others in aquamarine). The GenomeViz was used to construct the circular chromosome wheel (http://www.uniklinikum-giessen.de/genome/index.html).

We compared the general features of the *C. bescii* genome to those of five anaerobic thermophiles containing significant numbers of CAZy-related genes potentially involved in plant biomass degradation (Table 1): *C. saccharolyticus* DSM 8903, *C. thermocellum* ATCC 27405, *Thermotoga maritima* MSB8, *Thermoanaerobacter pseudethanolicus* ATCC 33223 and *T. tengcongensis* MB4. The genome size of *C. bescii* is similar to that of *C. saccharolyticus* (2.97 Mb) (19). Both utilize crystalline cellulose and xylan and are very closely related with over 2300 *C. bescii* genes having as their top Blast hit in the *C. saccharolyticus* sequence. The *C. bescii* genome is smaller than that of the cellulolytic but not xylanolytic *C. thermocellum* ATCC 27405 (T_{opt} 60°C, 3.8 Mb) (36), and larger than the genome of the xylanolytic but not cellulolytic *T. maritima* MSB8 (T_{opt} 80°C, 1.86 Mb) (37). By 16 S rRNA analysis, *C. bescii* is closely related to the *Thermoanaerobacter* genus. Its genome is most similar in size to that of *T. tengcongensis* MB4 (T_{opt} 75°C, 2.7 Mb) (38) and slightly smaller than that of *T. pseudethanolicus* ATCC 33223 (formerly *T. ethanolicus* strain 39E, T_{opt} 65°C, 2.4 Mb). Both *C. bescii* and *C. saccharolyticus* grow on polysaccharides such as starch but do not utilize cellulose or xylan.

Using the COG approach to predict gene function (39), we analyzed the genomes of 41 anaerobic thermophiles (Supplementary Tables S2 and S3). The genome of *C. bescii* is significantly enriched in genes encoding proteins involved with cell motility and secretion (COG

Table 1. General features and comparative genomics of *C. bescii DSM 6725*

General features	*Caldicellulosiruptor bescii DSM 6725*	*Caldicellulosiruptor saccharolyticus DSM 8903*	*Thermoanaerobacter pseudethanolicus ATCC 33223*	*Thermoanaerobacter tengcongensis* MB4	*Clostridium thermocellum ATCC 27405*	*Thermotoga maritima* MSB8
Length of chromosome (Mbp)	2.9	3.0	2.4	2.7	3.8	1.9
G+C content (%)	35.2	35.3	34.5	37.6	39.0	46.2
Coding density (%)	85.4	86.1	86.7	86.8	83.5	86.8
Total no. of predicted protein coding genes	2662	2679	2243	2588	3191	1858
Average length of protein-coding genes (bp)	942	957	915	905	1008	905
Total no. of predicted tRNA	47	46.0	56	55	56.0	46.0
Total no. of rRNA genes (no. of operons)	9 (3)	9 (3)	16 (4)	12 (5)	12 (4)	3 (3)
Secreted Proteome (SignalP prediction)	394	362	244	257	404	207
Membrane Proteins (TMHMM prediction)	344	348	282	401	481	239
Percentage secreted proteome (SignalP prediction)	14.80	13.51	10.88	9.93	12.66	11.14
Percentage membrane proteins (TMHMM prediction)	12.92	12.99	12.57	15.49	15.07	12.86
IS elements						
Full copies	34	91	42	69	100	3
Partial copies	132	130	77	104	56	22
Ismax-full-copy[a]	eISCsa4	eISCsa4	eISTps4/eISTps5	eISTps1	ISCth3	eISTma2
MaxCopy no.[b]	12	33	2	17	18	2
Growth on cellulose and xylan[c]	Cellulose, xylan	Cellulose, xylan	Does not grow	Does not grow	Cellulose	Xylan, CMC[d]

[a]ISmax is the IS element with the largest full copy number.
[b]MaxCopy no. is the largest full copy number.
[c]See text for references.
[d]CMC: carboxymethyl cellulose.

group N) and cell division (group D), in agreement with the SignalP result suggesting that *C. bescii* has a large number of secreted proteins. Interestingly, the cellulolytic *C. thermocellum* ATCC 27405 has a significantly higher than average number of genes responsible for DNA replication, recombination or repair, which may have facilitated the development of a genetic system for this organism (28). On the other hand, the *C. thermocellum* genome appears to have fewer genes involved in intracellular trafficking and in defense mechanisms. The genome of *C. bescii* has a lower number of uncharacterized genes (groups R and S) and a higher percentage of genes not assigned to COG categories.

Of the 2666 proteins encoded by the *C. bescii* genome, 394 (14.8%) and 344 (12.9%) are predicted to have signal peptides and transmembrane helices, respectively. Using a previously-developed program (32), it was found that the 2666 genes in *C. bescii* are predicted to be organized into 1209 transcriptional units, 577 of which are multi gene. The 259 genes with functions related to carbohydrate metabolism and sugar transport were predicted to form 180 transcriptional units, 111 of which are multi-gene and 69 are single-gene operons (Supplementary Table S4). Furthermore, *C. bescii* (and its close relative *C. saccharolyticus*) is predicted to contain 14 [12] sigma factors, 8 [7] anti-sigma modulators, 97 [60] putative transcription regulators and 18 [25] histidine kinases. Among the 18 putative histidine kinases, 11 are predicted to be membrane-associated. The high number of putative regulators suggests that *C. bescii* is highly responsive to changing environmental conditions and nutrient availability. This is in line with a recent report showing that the composition of *C. thermocellum* cellulosome varies with the growth substrate (40).

Genome dynamics

Insertion sequences (IS) have been found to be actively involved in the genomic recombination and horizontal gene transfer events in prokaryotic genomes (41). The coding region of an IS is flanked by fixed-length non-coding terminal regions, which are essential in mediating transposition and genomic recombination (31,42–44). Unfortunately, in many cases genome annotations include only the potential coding sequences carried by the elements and ignore their terminal regions. The statistics of IS elements in six genomes of anaerobic thermophiles are summarized in Table 1 (see also Supplementary Table S5). *Thermotoga maritima* harbors the smallest number of IS elements, with only 3 full copies and 22 partial copies. *Caldicellulosiruptor bescii* also harbors much fewer full IS copies [34] than the other four bacteria, especially compared to its closest relative *C. saccharolyticus* [91]. Full copies of IS elements are typically the results of recent proliferation. In light of these data, the genome of *C. bescii* is probably more stable than the other genomes, except for that of *T. maritima*.

The presence of IS elements suggests that all of the organisms listed in Table 1 likely have a history of

Table 2. Distribution of COG within some genomes of anaerobic thermophiles

	Computed frequency[a]						*P*-value[b]						Average	Function
COG	Cbes	Csac	Teth	TTE	Cthe	Tmar	Cbes	Csac	Teth	TTE	Cthe	Tmar		
A	0.00	0.00	0.00	0.00	0.00	0.00	0.240	0.240	0.240	0.240	0.240	0.240	0.056	RNA processing
B	0.00	0.11	0.00	0.10	0.05	0.07	0.167	0.390	0.167	0.388	0.747	0.304	0.149	Chromatin structure
C	6.08	5.98	6.67	6.91	6.02	8.16	0.069	0.063	0.892	0.871	0.065	0.726	9.581	Energy
D	2.75	1.96	2.60	2.25	1.80	1.54	0.018	0.768	0.967	0.108	0.679	0.499	1.537	Cell division
E	10.18	9.69	11.98	12.41	8.42	13.54	0.202	0.141	0.480	0.395	0.044	0.207	11.876	Amino acids
F	3.52	3.12	3.78	3.30	3.10	3.70	0.373	0.181	0.478	0.742	0.172	0.524	3.739	Nucleotides
G	12.29	12.18	10.32	9.22	7.40	12.28	0.099	0.895	0.750	0.629	0.406	0.099	8.169	Carbohydrates
H	5.63	4.87	5.66	3.40	4.58	3.98	0.394	0.245	0.600	0.068	0.198	0.881	6.126	Coenzymes
I	1.79	2.01	2.24	3.04	2.22	2.30	0.168	0.222	0.711	0.436	0.282	0.308	2.860	Lipids
J	8.90	7.99	9.03	8.17	7.77	9.42	0.311	0.220	0.675	0.763	0.201	0.631	10.513	Translation
K	8.45	8.31	7.73	8.43	8.75	5.93	0.158	0.813	0.663	0.161	0.107	0.152	7.207	Transcription
L	8.13	13.29	10.44	9.11	14.67	6.28	0.485	0.992	0.134	0.689	0.001	0.207	8.047	DNA
M	5.89	5.88	5.72	5.81	7.59	5.16	0.378	0.619	0.584	0.605	0.096	0.454	5.359	Cell membrane
N	8.51	3.76	3.07	3.72	4.35	4.12	0.016	0.598	0.489	0.592	0.314	0.348	3.136	Cell motility and secretion
O	3.84	3.18	3.78	4.19	4.07	3.77	0.181	0.036	0.159	0.671	0.726	0.157	4.521	Posttranslational modification
P	4.87	4.39	5.78	6.34	4.77	8.37	0.077	0.038	0.774	0.635	0.067	0.126	6.809	Inorganic ions
Q	3.59	1.32	1.24	1.78	1.48	1.47	0.075	0.202	0.179	0.646	0.750	0.245	2.153	Secondary metabolites
T	5.57	6.88	5.13	6.86	8.28	5.09	0.383	0.785	0.553	0.783	0.094	0.547	4.779	Signal transduction
U	0.00	2.12	3.01	2.20	2.31	2.51	0.033	0.695	0.067	0.273	0.767	0.171	1.658	Transport
V	0.00	2.81	1.83	2.72	2.27	2.30	0.035	0.878	0.449	0.859	0.276	0.736	1.707	Defense mechanism
Z	0.00	0.16	0.00	0.00	0.09	0.00	0.310	1.000	0.310	0.310	0.022	0.310	0.018	Cytoskeleton
R	9.30	11.00	10.79	11.63	10.00	13.02								General prediction only
S	5.20	6.61	7.94	6.88	5.77	17.00								Function unknown
Not in COG	32.11	18.48	13.64	14.61	22.23	9.85								Not assigned

[a]Frequency was computed as percentage of genes assigned to each COG group among all genes with COG assignment. When a gene was assigned to multiple COG groups, it would be counted multiple times.
[b]The *P*-value was calculated based on the assumption that the distribution of the frequence in each COG group follows a normal distribution.
Cbes, *Caldicellulosiruptor bescii* DSM 6725; Csac, *Caldicellulosiruptor saccharolyticus* DSM 8903; Teth, *T. pseudethanolicus* ATCC 33223; TTE, *T. tengcongensis* MB4; Cthe, *C. thermocellum* ATCC 27405; Tmar, *Thermotoga maritima* MSB8.

horizontal gene transfer events (Supplementary Table S5). Accordingly, the *C. bescii* genome has multiple sequences that are much more closely related to those in other genomes than they are to those in *C. saccharolyticus,* suggesting that these regions are the results of such events. As shown in Supplementary Figure S3 and Table S6, these include three of the thermophilic organisms listed in Table 1, *C. thermocellum* (23 genes), *T. tengcongensis* (21 genes) and *T. pseudethanolicus* (18 genes), as well as *Petrotoga mobilis* (11 genes), *Thermoanaerobacter sp.* X514 (11 genes) and *Dictyoglomus thermophilum* (14 genes). In addition, eleven *C. bescii* genes show the highest similarity to those in *C. phytofermentans* a mesophilic anaerobe that, like *C. bescii,* is both celluloytic and xylanolytic. These 'horizontally transferred' genes in *C. bescii* are predicted to encode ABC transporters [25], carbohydrate-active enzymes (CAZy) [17], mobile-element related [18], signal transducers and DNA binding (all containing a helix-turn-helix motif: 15), and genes encoding domains of unknown function like conserved domain UPF0236 [7], KWG leptospira repeat [6] and a radical SAM domain [6], many of which may be involved in various catabolic and anabolic pathways (45).

Distribution of CAZy genes within genomes of anaerobic thermophiles

The distribution of CAZy genes (http://www.cazy.org) related to plant biomass degradation within the genomes of the 41 anaerobic thermophiles is shown in Supplementary Tables S2 and S7. Glycoside transferases were not considered in this group as they are mainly involved in the biosynthesis of polysacchardies. Among these thermophiles, the 16 genomes of the archaea encode very few CAZy proteins. They do not contain polysaccharide lyases (PLs) and 13 of the 16 genomes do not encode CBMs, which are critical for degradation of insoluble polysaccharides. Six of the archaeal species grow on starch, although three of them are not predicted to contain genes that encode CBMs. Three of the genomes encode CBMs, glycosyde hydrolases (GHs) and carbohydrate esterases (CEs). Two of them, *Pyrococcus furiosus* and *Thermococcus kodakaraensis*, do not grow on cellulose or xylan, but do grow on starch, while the other, *Thermofilum pendens*, does not grow on any polysaccharide that has been examined although its genome encodes several GHs, CBMs and CEs. The presence of two GH13s, the recombinant forms of which are amylolytic enzymes, suggests that this organism can grow on starch or cyclodextrins (46).

In contrast to the anaerobic thermophilic archaea, all of the genomes of the 25 anaerobic thermophilic bacteria encode CBMs, GHs and CEs (Supplementary Tables S2 and S7). However, in many cases growth of these organisms on components of plant biomass has not been reported. Starch is the most common polysaccharide to be used by this group and 14 of them, including

C. bescii, have this ability and all contain α-amylase-type enzymes (GH13). PLs are identified in eight of these bacteria, and five of them have been shown to grow on pectin, including *C. bescii*, *C. saccharolyticus*, *C. thermocellum*, *T. lettingae* and *T. maritima*. The genome of *C. saccharolyticus* does not contain PLs but it does encode two GH28s that are putatively involved in hydrolysis of pectin backbone. Based on the numbers of CBMs and GHs that they contain, these anaerobic thermophilic bacteria (Supplementary Table S2) can be classified into three groups wherein (i) both CBMs and GHs are low (7 genomes); (ii) CBMs are low but GHs are high (15 genomes) and (iii) both CBMs and GHs are high (3 genomes). The latter category includes *C. bescii*, *C. saccharolyticus* and *C. thermocellum*. All representatives of the *Caldicellulosiruptor* genus grow on cellulose, xylan, pectin and starch (13,16). *C. thermocellum* does not grow on xylan as it cannot consume xylose, however, it depolymerizes xylan into xylose, xylobiose and xylooligosaccharides (9). Consequently, there is a clear correlation between the number of representatives of CAZy genes in a genome and the plant biomass-degrading abilities of a microorganism.

Comparison of CAZomes of *C. bescii* and *C. saccharolyticus*

The modular architecture of the 88 CAZy genes in *C. bescii* is shown in Supplementary Table S8. There is a comparable number of such genes in *C. saccharolyticus* [94]. Other common characteristics include (i) a similar module arrangement for CAZy-related proteins that do not contain CBMs, (ii) all proteins containing CBMs are predicted to be extracellular based on the presence of signal peptides (extracellular CAZy proteins in *C. bescii* are shown in Table 3); and (iii) the major CBM3s present in the enzymes from both organisms are derived from subfamilies 3a and 3b. CBM3a/3b bind tightly to crystalline cellulose and thus enhance the access of cellulases to their substrate relative to other cellulose-directed CBMs (47,48). In this respect, these two *Caldicellulosiruptor* species are similar to cellulolytic clostridia that produce cellulosomes, where CBM3 plays a pivotal role in substrate targeting of their respective cellulase complexes (10,11). The clostridial enzymes generally contain additional CBMs that direct the cellulose-tethered complex to specific regions of the cell wall, consistent with the activity of the enzyme containing these additional targeting modules (49,50). In contrast, *C. bescii* and *C. saccharolyticus* contain fewer of these additional, non-crystalline cellulose-binding CBM families (Supplementary Tables S2 and S8). The most significant of these are five CBM22s, and one CBM36 that likely targets xylan (Supplementary Table S8). Within this context it should be noted that CBM22s bind tightly to isolated xylan chains but not to hemicellulose within the plant cell wall (51). Thus, CBM22-containing enzymes likely target xylans that have been released from the plant cell wall. It appears, therefore, that CBM3s work in both of these bacteria as the primary mechanism for the attachment of enzymes to plant polysaccharides. Furthermore, the majority of CBM3-containing enzymes contain multiple CBM3s. These are likely to confer extremely tight binding to cellulose to offset the dissociation promoted by elevated temperatures. Indeed, it has been suggested that there is a general correlation between the growth temperature at which an organism and the frequency of finding enzymes with multiple CBM copies (52).

Notably, the CBM3s in the genomes of both *C. bescii* and *C. saccharolyticus* are concentrated only in one gene cluster and this encodes mainly CAZy proteins (Cbes_1853-_1867 and Csac_1076-_1085: see Figure 2 and Supplementary Figure S4). However, there is a significant difference in the arrangement of these gene clusters. In *C. bescii* this cluster is enriched in CBM3s, which are present as double or triple modules within one gene product, in comparison to the cluster of *C. saccharolyticus* (16 versus 10). Specifically, *C. bescii* contains three genes encoding PLs of different families that are absent from *C. saccharolyticus*. Moreover, of all thermophilic anaerobes, only *C. bescii* has PLs of three different families (Supplementary Table S2). The *C. bescii* cluster also contains three GH48s versus one in *C. saccharolyticus*. The GH48s are key enzymes in crystalline cellulose hydrolysis and are uniquely arranged in *C. bescii*. There is no other known example of three modules of this type in combination with a second catalytic module of different CAZy activity (Figure 2A and B). Interestingly, a deletion mutant of *C. thermocellum* lacking two GH48s was able to completely hydrolyze crystalline cellulose, albeit at a slower rate than the wild-type (28). This cluster in *C. bescii* also has three GH5 mannanases (versus one gene in *C. saccharolyticus*) and six genes encoding multifunctional CAZy proteins (versus three genes in *C. saccharolyticus*), each containing two catalytic modules of different hydrolytic activity separated by double or triple CBM3s.

Consequently, this CAZy-enriched gene cluster in *C. bescii* uniquely contains CBM3s that potentially mediate the binding of 13 catalytic modules to the insoluble substrate, while in *C. saccharolyticus* there are only eight catalytic modules attached to CBM3s. The *C. bescii* gene cluster also contains a GH74 module, which is a putative xyloglucanase (Table 3). This enzyme has an important role in biomass degradation as it hydrolyzes xyloglucan networks (53). In *C. bescii* the GH74 enzyme is part of a multi-modular protein with two CBM3s and GH48 (Cbes_1860), the combination of which is predicted to display synergism by binding the two catalytic modules to xyloglucan, hydrolyzing xyloglucan and releasing and hydrolyzing cellulose. In *C. saccharolyticus* the corresponding gene is truncated to GH74-CBM3 (Figure 2A and B). In Cbes_1867, GH48 is combined with GH9 via triplet of CBM3s. The combination of GH9 (endoglucanase) and GH48 (exoglucanase) assumes a synergy in hydrolysis of amorphous and crystalline parts of cellulose. Similarly, Cbes_1857 has the modular structure GH10-CBM3-CBM3-GH48 where GH10 is a xylanase, and the catalytic modules can act in a concert on mixed type xylan/cellulose substrates.

Table 3. Primary extracellular proteins of *C. bescii* involved in utilization of insoluble components of plant biomass

Gene	CAZy module architecture	CAZy module activity (www.cazy.org)		Transcriptomics		Proteomics	
		CBM	Catalycic (main activities)	Cell.	Signif.	ExtP	Membr.
Cbes_0089	GH11-CBM36	Xylan	Xylanase	Up	Yes		
Cbes_0182	GH43-CBM22[a]-GH43-CBM6[b]	Xylan[a,b], amorphous cellulose[b]	Xylanase, β-xylosidase, arabinanase	Up	Yes		
Cbes_0183	CBM22-CBM22-GH10	Xylan	Endo-1,4-, endo-1,3-β-xylanase	Up	Yes		
Cbes_0458	GH1		β-glucosidase, β-galactosidase, β-mannosidase, β-glucuronidase	Up	No	C	
Cbes_0594	GH5-CBM28-SLH-SLH-SLH	Amorphpus cellulose, cellooligosaccharides	Mannanase, cellulase, lichenase, xylanase	Up	Yes	C	
Cbes_0609	CBM41[a]-CBM48[b]-GH13-CBM20[c]	Starch[a,c], glycogen[b], cyclodextrines[c]	Starch	Up	Yes	C	C
Cbes_0610	CBM20	Starch, cyclodextins		Up	Yes	CX	
Cbes_0618	CBM22-CBM22-GH10	Xylan	Endo-1,4-, endo-1,3-β-xylanase	Up	Yes		X
Cbes_1439	GH23		Peptidoglycan lyase	Down	Yes		
Cbes_1462	CE4		Acetyl xylan esterase	Up	Yes		
Cbes_1829	CE4		Acetyl xylan esterase	Up	No		
Cbes_1853	PL11-CBM3	Cellulose	Rhamnogalacturonan lyase	Down	No	CX	
Cbes_1854	CBMX-PL3		Pectate lyase	Down	Y/N	CX	X
Cbes_1855	CBMX-PL9		Pectate lyase, exopolygalacturonate lyase	Down	Yes	CX	
Cbes_1857	GH10[a]-CBM3-CBM3-GH48[b]	Cellulose	endo-1,4-, endo-1,3-β-xylanase[a], cellobiohydrolase[b]	Up	Yes	CX	X
Cbes_1859	GH5[a]-CBM3-CBM3-GH44[b]	Cellulose	Mannanase[a]; Xyloglucanase, endoglucanase[b]	Up	Yes	CX	X
Cbes_1860	GH74[a]-CBM3-CBM3-GH48[b]	Cellulose	Xyloglucanase, endoglucanase[a]; cellobiohydrolase[b]	Up	Yes	CX	X
Cbes_1865	GH9[a]-CBM3-CBM3-CBM3-GH5[b]	Cellulose	Endoglucanase[a]; mannanase[b]	Up	No	CX	
Cbes_1866	GH5[a]-CBM3-CBM3-CBM3-GH5[b]	Cellulose	Mannanase[a], cellulase[b]	Down	Y/N	CX	
Cbes_1867	GH9[a]-CBM3-CBM3-CBM3-GH48[b]	Cellulose	Endoglucanase[a], cellobiohydrolase[b]	Up	Yes	CX	CX
Cbes_2593	GH13		Starch	Up	Yes		

Primary extracellular proteins are CAZy proteins where each contains a signal peptide and, in most cases, a CBM. The superscripts on the CBM and GH domains (a, b or c) indicate the corresponding CAZy module activity. Transcriptomics and proteomics show regulation of gene transcription on cellulose versus glucose, and protein identification using LC-MS/MS. N-terminal GH5 modules in Cbes_1859, Cbes_1865 and Cbes_1866 are identical to N-terminal GH5 module of Csac_1077, and C-terminal module in Cbes_1866 is identical to the C-terminal module in Csac_1077 which has been experimentally shown to display mannanase and cellulase/lichenase activities, respectively (40).
TMD, transmembrane domain; Membr., membrane protein fraction; C, cellulose; X, xylan; CBMX, an unknown module possibly pectin binding.

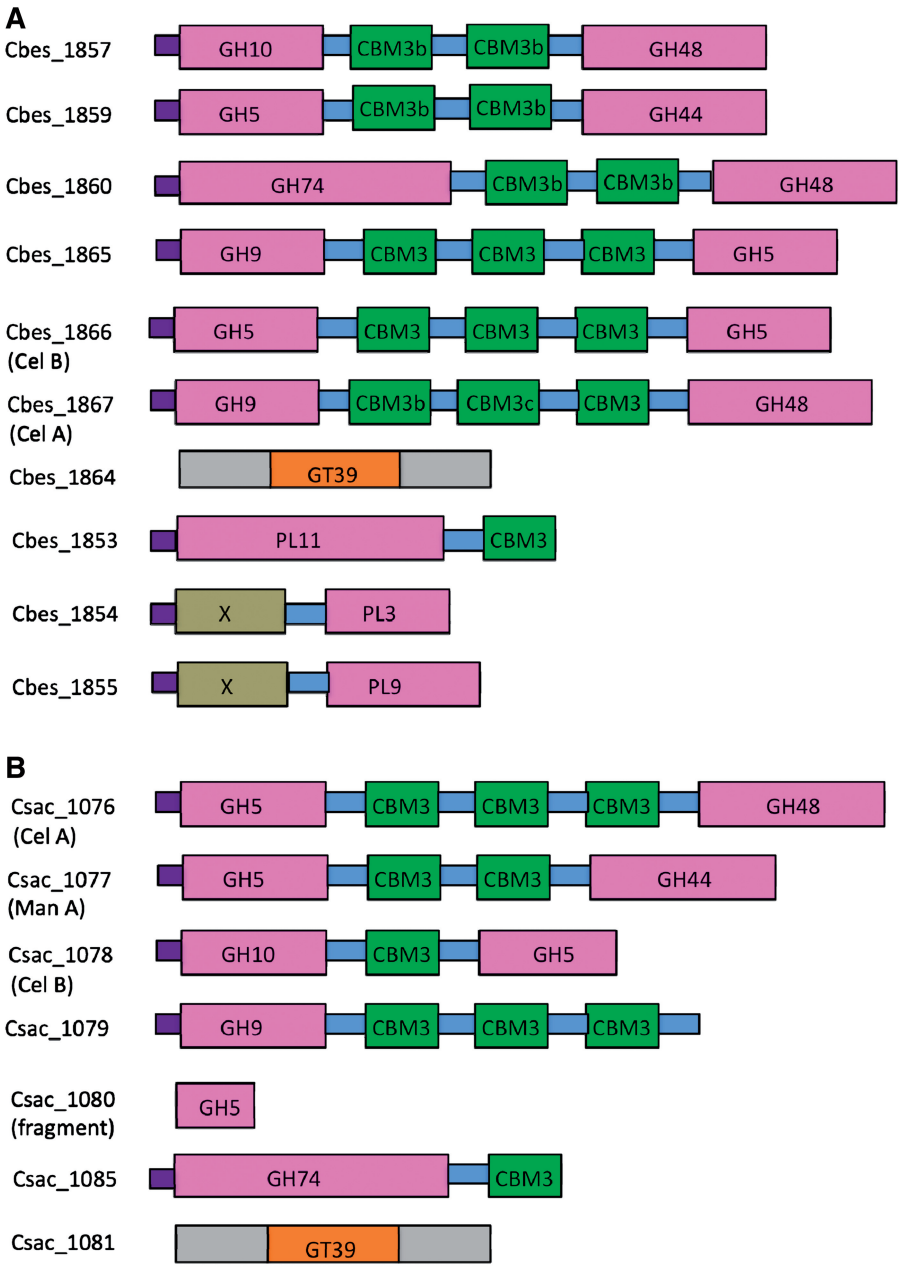

Figure 2. Comparison of the two relative gene clusters involved in biomass conversion in *C. bescii* DSM 6725 (**A**) and *C. saccharolyticus* DSM 8903 (**B**). Abbreviations: GH5, GH9, GH10, GH44, GH48 and GH74, glycoside hydrolases of families 5, 9, 10, 44, 48 and 74, respectively; CBM3, carbohydrate-binding module of family 3 where 'b' and 'c' are of subgroups within CBM3; GT39, glycosyl transferase of family 39; PL3, PL9 and PL11, polysaccharide lyases of families 3, 9 and 11; X, module of unknown function with homology to pfam CBM_4_9, Signal peptides, linkers and fragments of unknown function are shown in violet, blue and grey colors, respectively. CelA, CelB and ManA (encoding by Csac_1076, _1077 and _1078) are enzymes with experimentally demonstrated activities.

In general, the *C. bescii* gene cluster encodes a powerful set of CAZy enzymes active against major components of plant cell walls (cellulose, xylan, xylogluvan, pectin and mannan). In contrast, the analogous cluster in *C. saccharolyticus* is significantly truncated and lacks genes encoding some important biomass-related activities. All but of one the CBM3s in the *C. bescii* gene cluster has >99% identity, the three GH48s are 100% identical, and there are also three GH5s with high degree of sequence identity. Such gene duplication in the main CAZy-containing cluster in *C. bescii* suggests that both diversity of CAZys, and the 'dosage' of individual CAZy are important for this bacterium to adapt to new growth substrates, including various polysaccharides and related materials derived from plant biomass (54).

Our analysis of the CAZy-related genes in *C. bescii* revealed that the NCBI annotation of several of these sequences is incorrect and/or incomplete. For example, Cbes_1853 is annotated as cellulose 1,4-β-cellobiosidase, Cbes_1857, _1860 and _1867 are annotated as glycoside

hydrolases family 48 and Cbes_1865 is annotated as a glycoside hydrolase family 9. Based on comparisons with other sequences in the CAZy database, we propose that Cbes_1853 is a rhamnogalacturonan lyase; Cbes_1857, _1860, _1865 and _1867 are bifunctional enzymes containing GH10/GH48, GH74/GH48, GH9/GH5 and GH9/GH48, respectively. These changes are listed in Supplementary Table S8. Our new annotations suggest that these genes contain multiple domains, and such combination of multiple domains could be the key to biomass degradation.

Sugar transport

A total of 257 genes in the *C. bescii* genome are predicted to encode transporters including 171 involved in sugar transport (Supplementary Table S9). Cellular transport systems can be classified into seven main classes (http://www.chem.qmul.ac.uk/iubmb/mtp/). Although the total number of transporter genes is similar in the genomes of the two *Caldicellulosiruptor* species, *C. saccharolyticus* contains 18 more genes of family 3.A.1 that transport organic and inorganic molecules of various sizes, while *C. bescii* has 11 more genes of family 2.A.1 that transport molecules of small sizes including lactose (Supplementary Table S10).

ABC transporters in bacterial genomes are composed of an inner membrane component (IMC) and an ATPase component. In the *C. bescii* genome (Supplementary Table S9) in most cases the IMCs are paired and encoded by one operon suggesting that the ABC transporter system is tetrameric (two IMCs and two ATPases). The ATPases are typically not linked and are located remotely from the IMCs, suggesting that one ATPase serves multiple IMCs (55,56). Multiple solute-binding proteins (SBPs) were also identified in both genomes. They are generally located close to IMCs, but often are predicted in separate operons. Many SBPs belong to functional category COG1653 that includes putative proteins transporting various oligosaccharides and simple sugars. In many cases sugar transport systems are found in the same operon or in the vicinity of genes encoding CAZy related proteins. This observation suggests that these transporters are involved in the transport of sugars released by the corresponding enzymes encoded by these CAZy related operons or genes. In particular, the gene cluster Cbes_0050-_0063 contains ABC transporters and four glycosyl transferases of families GT2 and GT4 that transfer mannosyl, rhamnosyl, N-acetyl-glucosaminyl, β-galactosaminyl and galactosyl, glucosyl, mannosyl or xylosyl groups, respectively. It seems likely that ABC transporter elements located in the same operon are involved in transport of related sugars. The neighboring operons, Cbes_0174-_0181 and Cbes_0182-_0187 encode elements of ABC transporters and glycoside hydrolases GH43, GH39 and GH10, which encode xylanase, xylosidase and arabinofuranosidases, respectively. Transporter operon Cbes_1107-_1112 is located close to genes Cbes_1103 (GH51 with putative activities endoglucanase or arabinofuranosidase) and Cbes_1104 (GH4 displays activities of α-glucosidase, α-galactosidase and α-glucuronidase) (Supplementary Table S9). These observations imply that genes encoding sugar transport and sugar metabolism are typically closely associated.

Comparison of metabolic pathways

In comparing the pathways present in *C. bescii* and *C. sacharolyticus* assigned by the KEGG database, we found that both genomes are similar in term of the number of genes present in assigned pathways, as shown in Supplementary Table S11. However, there is one pathway present in *C. bescii* only. Its genome includes four genes essential for the biosynthesis of deoxythymidine-diphosphate rhamnose (dTDP-L-rhamnose) from glucose-1-phosphate, which is produced from cellobiose by cellobiose phosphorylase (Supplementary Table S11 and Figure S5). This is of particular interest as the activated sugar donor, glucose-1-phosphate, could be an energy source or could participate in the glycosylation of extracellular proteins and flagella biosynthesis (57,58), particularly since the genome of *C. bescii* is enriched in genes related to secretion and motility. In some bacteria, arabinogalactan is attached to peptidoglycan via a rhamnose-N-acetylglucosamine disaccharide linker unit (59) so it is not clear whether this pathway in *C. bescii* is essential for conversion of components of plant biomass. There is also a difference between the two *Caldicellulosiruptor* species in alanine metabolism. In particular, *C. bescii* and *C. sacharolyticus* contains eight and one copy, respectively, of homologs of alanine racemase (EC. 5.1.1.1), which reversibly converts L-alanine to D-alanine. However, they both contain only a single copy of D-alanine-D-alanine ligase (EC. 6.3.2.4), which converts D-alanine to D-alanyl-D-alanine, an enzyme involved in peptidoglycan metabolism in Gram-positive bacteria. The consequences of this are not clear at present.

Caldicellulosiruptor bescii CAZy and sugar transport genes with closest homologs in genomes other than *C. saccharolyticus*

Seventeen CAZy genes in the genome of *C. bescii* do not have their closest relatives in *C. saccharolyticus* (based on Blast analysis; see Supplementary Table S12). These genes were probably acquired from thermophilic [12] and mesophilic [5] microorganisms. Fourteen of these microorganisms degrade polysaccharides and three of them produce ethanol, but they also include three methanogenic archaea, which are not known to degrade polysaccharides. Ten of the 17 *C. bescii* genes are organized into three clusters that contain multiple CAZy-related proteins: Cbes_0052-_0061 and Cbes_0154-_0157 are all composed of GTs transferred from mesophiles, Cbes_1853-_1855 encodes PLs enzymes acquired from a thermophile and two mesophiles and Cbes_1853-_1855 was incorporated into a region containing multiple GH and CBM-containing genes. The latter gene cluster is discussed further below.

There are also 25 genes related to ABC transporters that do not have their closest relatives in *C. saccharolyticus*

(Supplementary Table S13). It is assumed that these were acquired by lateral gene transfer but in this case only from bacteria. The closest relatives of the 25 genes are found in 17 bacteria, many of which are capable of metabolizing polysaccharides with some generating ethanol as an end product. The ABC transporter genes appear to have been acquired predominantly [12] from mesophiles. Interestingly, 17 of the 25 genes are organized into 5 gene clusters, 3 of which are adjacent to 3 of the CAZy-gene clusters discussed above. In particular, cluster Cbes_2371-_2376 has four of its top Blast hits in *C. phytofermentans*, an organism that is capable of producing high concentrations of ethanol during cellulose fermentation. The same cluster encodes 3 ABC transporter genes, a GH43, a histidine kinase and a response regulator, suggesting that this six-gene cluster was horizontally transferred more or less intact from a *Clostridium* species, and may play a significant role in biomass degradation. Similarly, the Cbes_2076-_2094 cluster contains two ABC transporters (six genes), a GH2 and two integrase-related genes. A large number of genes in this cluster have their top Blast hits in two species, *B. subtilis* and *D. thermophilum,* indicating that this region could be a hot spot for DNA integration or genome rearrangement in *C. bescii*.

These data suggest that the exchange of genetic information has had a significant impact on the metabolic capabilities of *C. bescii,* and that this exchange has occurred between very different microorganisms, including (i) archaea and bacteria, (ii) aerobes and anaerobes, (iii) Gram-positive and Gram-negative bacteria and (iv) (hyper)thermophiles and mesophiles (and even psychrophiles). Moreover, these observations provide conclusive evidence for the divergent evolution of what appear to be two very closely-related species, *C. bescii* and *C. saccharolyticus*.

Genes encoding proteins potentially involved in cell–carbohydrate adhesion

In some cellulolytic microorganisms such as *C. thermocellum,* the strong interactions between the cells and the insoluble polysaccharide substrate are mediated by the cellulosome. The genome analyses of *C. bescii* shows that it does not produce a cellulosome complex as no dockerin- and cohesin-like domains of either types I or II were identified. In addition, genes encoding extracellular CAZy enzymes did not contain similar domains of unknown function that might encode new types of dockerins. However, microscopy studies show that *C. bescii* cells directly attach to xylan and switchgrass (Figure 3). The attachment is dynamic as many cells are also planktonic, enabling cell densities to be used as a measure of cell growth (14,15). Although the mechanism is not known, analysis of the genome of *C. bescii* reveals many genes that are predicted to encode modules that could be involved in such cell–substrate interactions (Table 4). They include surface-layer homology (SLH) domains which are known to mediate the binding of proteins to cell surfaces (60), fibronectin type 3-like (Fn3) domains containing binding sites for the cell surface (http://pfam.sanger.ac.uk), and lysine motif (LysM) domains found in a variety of enzymes involved in bacterial cell wall degradation that may have a general peptidoglycan-binding function (61). In addition, *C. bescii* contains Fn3-like domains that have sequence similarity to so-called 'X' domains, which have shown to bind carbohydrates (62).

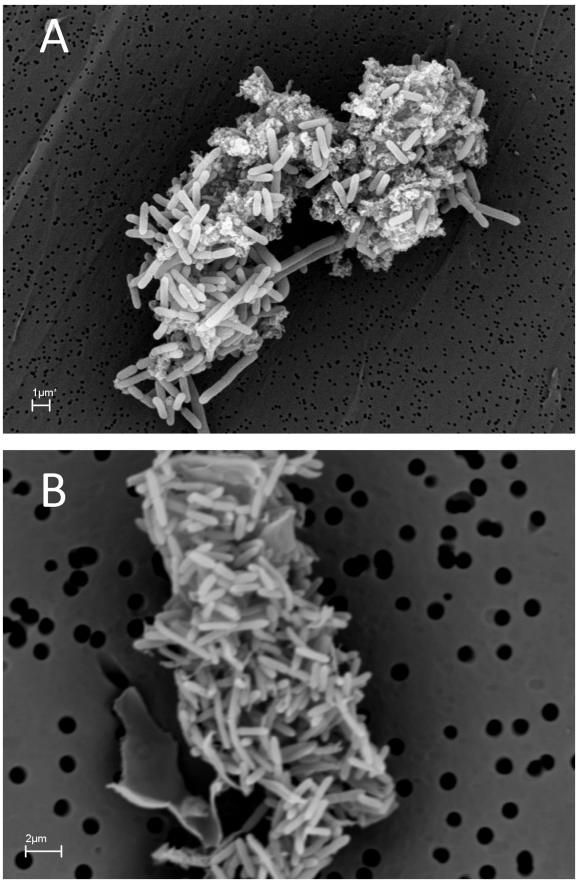

Figure 3. Scanning electron microscopy (SEM) images of *C. bescii* cells attached to xylan from oat spelts (**A**) and to switchgrass (**B**). The bars indicate (A) 1 μm and (B) 2 μm, respectively.

Specifically, Cbes_0594 has an SLH domain combined with GH5 and CBM28. Binding of *C. stercorarium* xylanase to the cell wall via its SLH domains has been demonstrated (63). Cbes_0174 and Cbes_0181 contain bacterial solute-binding domains (SBPs), which are typically attached to an outer membrane and are components of sugar transport systems (64). Cbes_0174 has an N-terminal and Cbes_0181 has C-terminal modules with BLAST hits to CBM6 and pfam CBM_4_9, respectively (designated here as CBM_X, Table 4). All three proteins are candidates for binding to both cell (by SLH, SBP) and polysaccharides (by CBM28, CBM_X). There are also many modules of unknown biological function listed in Table 4 (modules designated as 'X', LysM, Fn3, RHS, etc.) that contain signal peptides and could potentially be presented on the cell surface of *C. bescii* that may display novel catalytic and/or binding functions.

Table 4. *Caldicellulosiruptor bescii* genes encoding proteins with putative cell adhesion, protein–protein interaction or carbohydrate-binding function

Gene name	Protein (AAs)	SP	Annotation	Domain structure	Transcriptomics		Proteomics		
					FilterPaper	Significant	ExtP	Membrane	WC
Cbes_0012	3027	Y	Q466C0 Putative uncharacterized protein	**SLH-SLH-SLH-Fn3**-VWA-**RHS**	Up	Yes	CXn	CXn	
Cbes_0077	1710	Y	A4J714 S-layer domain protein	**SLH-SLH-SLH**-Transglut_core	Up	Yes	CXn	C	
Cbes_0594	755	Y	Q59154 Endoglucanase	GH5-**CBM28-SLH-SLH-SLH**	Up	Yes	C		
Cbes_0608	547	Y	A3DET8 Cellulose 1,4-beta-cellobiosidase	X-**SLH-SLH-SLH**	Up	Yes			
Cbes_0438	1157	Y	A4XG20 S-layer domain protein	**SLH-SLH**-X	Up	No	Xn		Yes
Cbes_1839	575	Y	A4XM24 S-layer domain protein	**SLH-SLH**-X	Up	Yes			
Cbes_1943	277	Y	A4XI88 S-layer domain protein	**SLH-SLH**	Yes	No			Yes
Cbes_2295	1074	Y	A4XM87 S-layer domain protein	**SLH-SLH**-X	Yes	Yes	Xn	C	
Cbes_2341	484	Y	A4XH32 S-layer domain protein	X-**SLH-SLH**	Y/N	No			
Cbes_1573	1055	Y	A4XH96 Putative uncharacterized protein	**SLH-SLH**- SpoVT_AbrB	Y/N	No			
Cbes_2303	1018	Y	A4XM93 S-layer domain protein	**SLH**-X	Y/N	No	CXn	CXn	
Cbes_2342	1010	Y	A4XH31 Putative uncharacterized protein	**SLH**-X	Y/N	No			Yes
Cbes_1944	1201	Y	A4XI87 Fibronectin, type III domain protein	X-**Fn3**-X	Y/N	No			Yes
Cbes_1945	1265	Y	A4XI87 Fibronectin, type III domain protein	X-**Fn3**-X	Y/N	No			
Cbes_0190	582	N	A4XM45 Peptidase M23B	X-**LysM**-G5-peptidase M3	UP	Yes			
Cbes_0508	203	Y	A4XIM2 Allergen V5/ Tpx-1 family protein	**LysM**-SH3_3	Y/N	No			Yes
Cbes-0560	507	Y	A4XHM2 Peptidoglycan-binding LysM	**LysM-LysM**	Yes	No			Yes
Cbes_1391	109	Y	A4XKU6 Peptidoglycan-binding LysM	**LysM**	Y/N	No			Yes
Cbes_2402	511	N	A4XGE4 Peptidoglycan-binding LysM	X-**LysM**	Y/N	No			Yes
Cbes_0174	951	Y	Extracellular solute-binding protein family 1	**CBM_X-SBP1**	Up	No	Xn	Xn	
Cbes_0181	595	Y	Extracellular solute-binding protein family 1	**SBP1-CBM_X**	Up	Yes	Xn	CXn	

SP, Signal Peptide; SLH, surface layer homology domain; SBP1, solute-binding protein of family 1; X, domain not present in PFAM; CBM_X, Pfam annotation of PF06204; RHS, multiple tandem 22-residue repeats each containing strongly conserved dipeptide YD; WC, cell-extract; C, cells grown on cellulose; Xn, cells grown on xylan.

Hypothetical genes and their location

According to the NCBI annotation, the *C. bescii* genome contains a total of 826 ORFs of unknown function that are annotated as encoding either hypothetical (723 HP) or conserved hypothetical proteins (103 CHP: Supplementary Table S14). We have now assigned a putative function to 46 of them via the KEGG [2] and COG [44] databases, and using the CAZy database another previously annotated CHP is annotated as a GT4 (Cbes_1572). In order to obtain some insight into the likely function of some of other C/HPs, we utilized the fact that genes transcribed in the same operon or gene cluster are often functionally related (65,66). A gene cluster is defined here as set of genes encoded on the same DNA strand with intergenic distances between adjacent genes of <300 bp (66). We found that 17 C/HPs are in the same operon with or located adjacent to CAZy genes, therefore, they are predicted to be functionally related to carbohydrate metabolism and potentially plant biomass conversion. As an example, Supplementary Figure S6 shows genes encoding a CHP (Cbes_0178) associated with genes encoding sugar transporters, suggesting that this CBP is likely involved in the same function. Consequently, using the KEGG and COG annotations, operon and gene cluster prediction analyses, putative functions can be assigned to a total of 295 HPs (41%, 428 remain unassigned) and 44 CHPs (43%, 59 remain unassigned: Supplementary Table S15).

Insights into gene function from proteomic and transcriptomic analyses

A total of 1429 (54%) of 2666 predicted protein-coding sequences (PCSs) were confirmed by proteomic analyses (Supplementary Table S16) and 1790 (67.1%) PCSs were confirmed by transcriptomic data (Supplementary Tables S17 and S18). Therefore, a total of 2196 (83%) of the annotated PCSs were confirmed, including 46.6% by both methods, 18.5% by proteomics and 34.9% by transcriptomics. Among 88 genes annotated as CAZy-related genes, 59 (67.0%) are confirmed experimentally, including 28.8% by both methods. Among 826 PCSs that were annotated as encoding C/HPs, 613 (74%) were expressed on different substrates according to the transcriptomic and proteomic results. These data also allowed us to correct putative transcription unit (TU or operon) boundaries for 18 gene pairs (or 5% of the gene pairs with proteomic data: Supplementary Table S19). This leads to the splitting and merging of 20 TUs into 30 TUs. The 257 genes predicted to encode sugar transporters and CAZys are organized into 180 TUs (Supplementary Table S4). Of the 171 transporters predicted to be sugar-related and 88 CAZy genes, expression at the RNA or protein level was shown for 136 (79%; Supplementary Table S9) and 84 (Supplementary Table S8), respectively, have been detected.

When *C. bescii* was grown on crystalline cellulose (filter paper) versus glucose, a total of 1203 genes had a significant change in expression level, as shown in Supplementary Table S17. These included 64 CAZys (32 down- and 32 up-regulated: Supplementary Table S8), 90 transporters (29 down- and 61 up-regulated: Supplementary Table S9) and 358 C/HPs (124 down- and 234 up-regulated: Supplementary Table S14). Among the 21 primary CAZy genes (encoding proteins with signal peptides and CBMs) (Supplementary Table S8), 16 were up-regulated on cellulose and 14 proteins were identified on both cellulose and xylan (Table 3). Of 21 genes putatively related to cell–substrate adhesion (Table 4), 12 genes were up-regulated on cellulose and 8 proteins were identified on cellulose and xylan.

More detailed analyses were conducted with operons/gene clusters (Supplementary Table S20, see also Tables S21 and S22). Among the gene clusters whose expression is up-regulated during growth on cellulose, there are six of potential interests. These include (i) Cbes_1856-Cbes_1864 encoding the majority of CAZy multi-modular multifunctional enzymes discussed above, as well as two HPs; (ii) Cbes-2371-Cbes_2375 encoding GH43 and two membrane components of ABC transporters, a gene cluster that is missing in *C. saccharolyticus*; (iii) Cbes_2413-Cbes_2421, Cbes_2494-Cbes_2500 and Cbes_0261-Cbes_0265, all of which encode HPs; and (iv) Cbes_2591-Cbes_2595 encoding an α-amylase, a DNA repair protein and three HPs. The up-regulation of these gene clusters on cellulose suggests that they are involved in plant cell wall conversion. Five clusters, including genes encoding proteins of different metabolic pathways, were down-regulated on cellulose indicating the plasticity of transcription regulation upon changing growth conditions. Upon switching from glucose to cellulose, four genes of the Cbes_1853-Cbes_1864 cluster are down-regulated while five other genes of the same cluster are up-regulated. This differential regulation validates our operon prediction that this cluster contains multiple transcription units.

Among 42 predicted TFs with significant changes in gene expression, 17 and 25 were down- and up-regulated, respectively, when cells were grown on crystalline cellulose versus glucose (Supplementary Table S23). It was previously suggested (67) that the level of expression of a TF is proportional to the number of operons that it regulates. This observation was used to predict the number of operons regulated by the TFs. Of 13 TFs with >4-fold changes, 7 and 6 were down- and up-regulated, respectively. In particular, TF Cbes_1856, a component of the major CAZy gene cluster, Cbes_1856-Cbes_1864, is up-regulated 3.4-fold suggesting that it is involved in plant biomass conversion. Cbes_2264 is up-regulated 17-fold. This TF is part of an operon encoding sugar transporters (Cbes-_2265-Cbes_2266), which are also up-regulated. These data suggest that these transporters utilize soluble oligomeric products of cellulose hydrolysis rather than glucose. In contrast, TF Cbes_2033 is down-regulated >8-fold, and it is located upstream of gene cluster Cbes_2029-Cbes_2031, which contains a predicted sugar transporter that is up-regulated 4-fold. This cluster is presumably not involved in cellulose metabolism. TF Cbes_1901 is down-regulated >9-fold, although the adjacent HP gene is up-regulated <2-fold, supporting the prediction that this TF regulates multiple operons (Supplementary Table S23).

It is also evident that the production of some CAZys, sugar transporters and C/HPs are sugar-specific (Supplementary Tables S8 and S16). Two of them were found only when cells were grown on xylan: Cbes_0618 (CBM22-CBM22-GH10) and Cbes_0152 (CE7). This is in accord with the CAZy annotation, as the CBM22 domain binds xylan, GH10 is an endo-xylanase and CE7 is an acetyl–xylan esterase. These proteins are assumed to play a pivotal role in hemicellulose degradation. Other proteins were detected only upon growth on cellulose and cellobiose, but not on xylan. They include Cbes_0097 (GH30) and Cbes_0458 (GH1), which are potential β-glucosidases related to cellulose degradation, Cbes_0468 (GH36, potential α-galactosidase) and Cbes_0609 (CBM41-CBM48-GH13-CBM20 with CBMs binding to α-linked polysaccharides and α-amylase). Production of the latter two proteins during growth on cellulose suggests that the stereospecificity of the sugar linkage is not important for the regulation of the respective genes. Cbes_0459 and Cbes_0460 are putative cellobiose/cellodextrin phosphorylases (GH94) detected in much higher amounts than β-glucosidases. This is consistent with the energetics of cellulose degradation as cellobiose/cellodextrin phosphorylases provide an advantage for anaerobic cellulolytic microorganisms. They convert cellobiose/cellodextrins into glucose and glucose-1-phosphate without utilizing valuable ATP, which can be conserved for energy-consuming reactions. In contrast, β-glucosidase hydrolyses cellobiose into two glucose molecules, which must be phosphorylated with

ATP before they can be utilized. Hence, in general, interpretation of the microarray and proteomic data is consistent with the CAZy database classification. Five and three sugar transporters were identified only after growth on xylan or on cellulose/cellobiose, respectively, consistent with the specificity of these proteins for certain oligosaccharides. Furthermore, 8 and 16 C/HPs were detected on xylan and cellulose/cellobiose, respectively. Two of the xylan-specific proteins (Cbes_2729 and _2368), and 2 cellulose/cellobiose specific proteins (Cbes_2630 and _1288) were detected at relatively high levels. These data suggest that these previously uncharacterized proteins play important roles in hemicellulose/cellulose metabolism, even though they have no recognizable CAZy domains.

CONCLUSIONS

Caldicellulosiruptor bescii is the most thermophilic anaerobic bacterium capable of utilizing cellulose as well as multiple polysaccharides and unprocessed plant biomass. From an analysis of its genome, coupled with transcriptomic and proteomic data, we suggest that not one particular feature but a combination of properties that act in synergy enables the bacterium to degrade various polysaccharides and plant biomass:

(i) Enrichment in multi-modular, multi-functional CAZy proteins each containing two catalytic modules specific to different components of plant cell walls combined with multiple CBMs.
(ii) Presence of three PLs of different families absent from the genome of C. *saccharolyticus*.
(iii) Concentration of all multi-modular, multifunctional CAZy genes including three PLs and all CBM3s in one large functional gene cluster.
(iv) Multiplication of CAZy modules within the large CAZy gene cluster increasing 'dosage' of particular CAZy modules, in particular, three cellobiohydrolase GH48s in combination with CBM3s.
(v) Absence of modules with motifs of dockerin or cohesin domains, which mediate the assembly of the cellulosome. This confers *C. bescii* with more flexibility to produce combinations of 'free' enzymes to degrade a variety of insoluble polysaccharides.
(vi) Binding of *C. bescii* to xylan and switchgrass is mediated by proteins containing conserved non-CAZy modules known to bind polysaccharides or cell wall components, and proteins with CBM and membrane-binding modules. This binding is dynamic, in contrast to the irreversible binding of the cellulosome to cellulose.
(vii) Hypothetical/conserved hypothetical genes located inside or in the vicinity of CAZy or sugar transport genes/operons are related to plant biomass conversion.
(viii) Hypothetical/conserved hypothetical genes that are highly regulated during growth on polysaccharides or related substrates are likely involved in plant biomass conversion.

Currently there is an increased interest in members of the *Caldicellulosiruptor* genus that display the ability to degrade multiple polysaccharides as well as plant biomass. Like the prototypical cellulose-degrader, *C. thermocellum,* these bacteria have a high potential for use in efficient two-step biomass-sugar–biofuel conversion processes. The data presented here are a valuable source of information that can be utilized for further characterization of the *Caldicellulosiruptor* species that will lead to a deeper understanding of the mechanisms of the non-cellulosomal plant biomass conversion process.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This work was supported by the Bioenergy Science Center (BESC), Oak Ridge National Laboratory, a US Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science (contract no. DE-PS02-06ER64304) (DOE 4000063512); the University of California, Lawrence Berkeley National Laboratory (contract no. DE-AC02-05CH11231); Lawrence Livermore National Laboratory (contract No. DE-AC52-07NA27344); Los Alamos National Laboratory (contract No. DE-AC02-06NA25396). Agence Nationale de la Recherche, e-TRICEL (grant No. AANR-07-BIOE-006, to B.H.); National Science Foundation, (DEB-0830024, DBI-0542119). Funding for open access charge: US Department of Energy (DE-AC05-00OR22725).

Conflict of interest statement. None declared.

REFERENCES

1. Demain,A.L. (2009) Biosolutions to the energy problem. *J. Ind. Microbiol. Biotechnol.*, **36**, 319–332.
2. Farrell,A.E., Plevin,R.J., Turner,B.T., Jones,A.D., O'Hare,M. and Kammen,M.D.M. (2006) Ethanol can contribute to energy and environmental goals. *Science*, **311**, 506–508.
3. Kumar,R., Singh,S. and Singh,O.V. (2008) Bioconversion of lignocellulosic biomass: biochemical and molecular perspectives. *J. Ind. Microbiol. Biotechnol.*, **35**, 377–391.
4. Wong,D.W. (2006) Feruloyl esterase: a key enzyme in biomass degradation. *Appl. Biochem. Biotechnol.*, **133**, 87–112.
5. Negro,M.J., Manzanares,P., Ballesteros,I., Oliva,J.M., Cabañas,A. and Ballesteros,M. (2003) Hydrothermal pretreatment conditions to enhance ethanol production from poplar biomass. *Appl. Biochem. Biotechnol.*, 105–108, 87–100.
6. Lynd,L.R. (2008) Energy biotechnology. *Curr. Opin. Biotechnol.*, **19**, 199–201.
7. Cooney,C.L. and Wise,D.L. (2004) Thermophilic anaerobic digestion of solid waste for fuel gas production. *Biotechnol. Bioeng.*, **17**, 1119–1135.
8. Blumer-Schuette,S.E., Kataeva,I., Westpheling,J., Adams,M.W. and Kelly,R.M. (2008) Extremely thermophilic microorganisms for biomass conversion: status and prospects. *Curr. Opin. Biotechnol.*, **19**, 210–217.
9. Wiegel,J., Mothershed,C.P. and Puls,J. (1985) Differences in xylan degradation by various noncellulolytic thermophilic anaerobes and *Clostridium thermocellum*. *Appl. Environ. Microbiol.*, **49**, 656–659.

10. Bayer,E.A., Lamed,R., White,B.A. and Flint,H.J. (2008) From cellulosomes to cellulosomics. *Chem. Rec.*, **8**, 364–377.
11. Bayer,E.A., Belaich,J.P., Shoham,Y. and Lamed,R. (2004) The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu. Rev. Microbiol.*, **58**, 521–554.
12. Lin,L., Song,H., Ji,Y., He,Z., Pu,Y., Zhou,J. and Xu,J. (2010) Ultrasound-mediated DNA transformation in thermophilic Gram-positive anaerobes. *PLoS ONE*, **5**, e12582.
13. Hamilton-Brehm,S.D., Mosher,J.J., Vishnivetskaya,T., Podar,M., Carroll,S., Allman,S., Phelps,T.J., Keller,M. and Elkins,J.G. (2010) *Caldicellulosiruptor obsidiansis* sp. nov., an anaerobic, extremely thermophilic, cellulolytic bacterium isolated from Obsidian pool, Yellowstone National Park. *Appl. Environ. Microbiol.*, **76**, 1014–1020.
14. Svetlichnyi,V.A., Svetlichnaya,T.P., Chernykh,N.A. and Zavarzin,G.A. (1990) *Anaerocellum thermophilum* gen. nov., sp. nov., an extremely thermophilic cellulolytic eubacterium isolated from hot-springs in the valley of Geysers. *Microbiol. (Translation of Mikrobiologia)*, **59**, 598–604.
15. Yang,S.J., Kataeva,I., Wiegel,J., Yin,Y., Dam,P., Xu,Y., Westpheling,J. and Adams,M.W. (2009) Classification of '*Anaerocellum thermophilum*' as *Caldicellulosiruptor bescii* strain DSM 6725T sp. nov. *Int. J. Syst. Evol. Microbiol.*, **60**, 2011–2015.
16. Yang,S.-J., Kataeva,I., Hamilton-Brehm,S.D., Engle,N.L., Tschaplinski,T.J., Doeppke,C., Davis,M., Westpheling,J. and Adams,M.W. (2009) Efficient degradation of lignocellulosic plant biomass, without pretreatment, by the thermophilic anaerobe "*Anaerocellum thermophilum*" DSM 6725. *Appl. Environ. Microbiol.*, **75**, 4762–4769.
17. Ivanova,G., Rakhely,G. and Kovacs,K.L. (2008) Hydrogen production from biopolymers by *Caldicellulosiruptor saccharolyticus* and stabilization of the system by immobilization. *Int. J. Hydrogen Energy*, **33**, 6953–6961.
18. Ivanova,G., Rakhely,G. and Kovacs,K. (2009) Thermophilic biohydrogen production from energy plants by *Caldicellulosiruptor saccharolyticus* and comparison with related studies. *Int. J. Hydrogen Energy*, **34**, 3659–3670.
19. van de Werken,H.J.G., Verhaart,M.R., VanFossen,A.L., Willquist,K., Lewis,D.L., Nichols,J.D., Goorissen,H.P., Mongodin,E.F., Nelson,K.E., vanNiel,E.W.J. *et al.* (2008) Hydrogenomics of the extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus*. *Appl. Environ. Microbiol.*, **74**, 6720–6729.
20. Kataeva,I.A., Yang,S.-J., Dam,P., Poole,F.L. 2nd, Yin,Y., Zhou,F., Chou,W.C., Xu,Y., Goodwin,L., Sims,D.R. *et al.* (2009) Genome sequence of the anaerobic, thermophilic, and cellulolytic bacterium "*Anaerocellum thermophilum*" DSM 6725. *J. Bacteriol.*, **191**, 3760–3761.
21. Luethi,E., Jasmat,N.B., Grayling,R.A., Love,D.R. and Bergquist,P.L. (1991) Cloning, sequence analysis, and expression in *Escherichia coli* of a gene coding for a beta-mannanase from the extremely thermophilic bacterium Caldocellum saccharolyticum. *Appl. Environ. Microbiol.*, **57**, 694–700.
22. Te'o,V.S.J., Saul,D.J. and Bergquist,P.L. (1995) celA, another gene coding for a multidomain cellulase from the extreme thermophile *Caldocellum saccharolyticum*. *Appl. Microbiol. Biotechnol.*, **43**, 291–296.
23. Sandquist,D., Filonova,L., von Schantz,L., Ohlin,M. and Daniel,G. (2010) Microdistribution of xyloglucan in different poplar cells. *BioResources*, **5**, 796–807.
24. Hayashi,T., Kaida,R., Kaku,T. and Baba,K. (2010) Loosening xyloglucan prevents tensile stress in tree stem bending but accelerates the enzymatic degradation of cellulose. *Russian. J. Plant Physiol.*, **57**, 316–320.
25. Kataeva,I.A., Seidel,R.D. III, Shah,A., West,L.T. and Li,X.L.L. (2002) The fibronectin type 3-like repeat from the *Clostridium thermocellum* cellobiohydrolase CbhA promotes hydrolysis of cellulose by modifying its surface. *Appl. Environ. Microbiol.*, **68**, 4292–4300.
26. Lee,H.S., Shockley,K.R., Schut,G.J., Conners,S.B., Montero,C.I., Johnson,M.R., Chou,C.J., Bridger,S.L., Wigner,N., Brehm,S.D. *et al.* (2006) Transcriptional and biochemical analysis of starch metabolism in the hyperthermophilic archaeon *Pyrococcus furiosus*. *J. Bacteriol.*, **188**, 2115–2125.
27. Conners,S.B., Montero,C.I., Comfort,D.A., Shockley,K.R., Johnson,M.R., Chhabra,S.R. and Kelly,R.M. (2005) An expression-driven approach to the prediction of carbohydrate transport and utilization regulons in the hyperthermophilic bacterium *Thermotoga maritima*. *J. Bacteriol.*, **187**, 7267–7282.
28. Olson,D.G., Tripathi,S.A., Giannone,R.J., Lo,J., Caiazza,N.C., Hogsett,D.A., Hettich,R.L., Guss,A.M., Dubrovsky,G. and Lynd,L.R. (2010) Deletion of the Cel48S cellulase from *Clostridium thermocellum*. *PNAS Early Edition*, **107**, 1–6.
29. Eng,J.K., McCormack,A.L. and Yates,J.R.I. (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.
30. Tabb,D.L., McDonald,W.H. and Yates,J.R. (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.*, **1**, 21–26.
31. Zhou,F., Olman,V. and Xu,Y. (2008) Insertion sequences show diverse recent activities in Cyanobacteria and Archaea. *BMC Genomics*, **9**, 36.
32. Dam,P., Olman,V., Harris,K., Su,Z. and Xu,Y. (2007) Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res.*, **35**, 288–298.
33. Brouwer,R.W., Kuipers,O.P. and vanHijum,S.A. (2008) The relative value of operon predictions. *Brief Bioinform.*, **9**, 367–375.
34. Cantarel,B.L., Coutinho,P.M., Rancurel,C., Bernard,T., Lombard,V. and Henrissat,B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.*, **37**, D233–D238.
35. Clausen,A., Mikkelsen,M.J., Schröder,I. and Ahring,B.K. (2004) Cloning, sequencing, and sequence analysis of two novel plasmids from the thermophilic anaerobic bacterium *Anaerocellum thermophilum*. *Plasmid*, **52**, 131–138.
36. Wagner,I.D. and Wiegel,J. (2008) Diversity of thermophilic anaerobes. *Ann. N. Y. Acad. Sci.*, **1125**, 1–43.
37. Nelson,K.E., Clayton,R.A., Gill,S.R., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Nelson,W.C. and Ketchum,K.A. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, **399**, 323–329.
38. Bao,Q., Tian,Y., Li,W., Xu,Z., Xuan,Z., Hu,S., Dong,W., Yang,J., Chen,Y., Xue,Y. *et al.* (2002) A complete sequence of the *T. tengcongensis* genome. *Genome Res.*, **12**, 689–700.
39. Anantharaman,V., Koonin,E.V. and Aravind,L. (2002) Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res.*, **30**, 1427–1464.
40. Blouzard,J.C., Coutinho,P.M., Fierobe,H.P., Henrissat,B., Lignon,S., Tardif,C., Pages,S. and de Philip,P. (2010) Modulation of cellulosome composition in Clostridium cellulolyticum: adaptation to the polysaccharide environment revealed by proteomic and carbohydrate-active enzyme analyses. *Proteomics*, **10**, 541–554.
41. Siguier,P., Perochon,J., Lestrade,L., Mahillon,J. and Chandler,M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.*, **34**, D32–D36.
42. Degnan,P.H., Leonardo,T.E., Cass,B.N., Hurwitz,B., Stern,D., Gibbs,R.A., Richards,S. and Moran,N.A. (2009) Dynamics of genome evolution in facultative symbionts of aphids. *Environ. Microbiol.*, **12**, 2060–2069.
43. Hill,K.K., Xie,G., Foley,B.T., Smith,T.J., Munk,A.C., Bruce,D., Smith,L.A., Brettin,T.S. and Detter,J.C. (2009) Recombination and insertion events involving the botulinum neurotoxin complex genes in *Clostridium botulinum* types A, B, E and F and *Clostridium butyricum* type E strains. *BMC Biol.*, **7**, 66.
44. Chen,Y., Zhou,F., Li,G. and Xu,Y. (2008) A recently active miniature inverted-repeat transposable element, Chunjie, inserted into an operon without disturbing the operon structure in *Geobacter uraniireducens* Rf4. *Genetics*, **179**, 2291–2297.
45. Sofia,H.J., Chen,G., Hetzler,B.G., Reyes-Spindola,J.F. and Miller,N.E. (2001) Radical SAM, a novel protein superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information visualization methods. *Nucleic Acids Res.*, **29**, 1097–1106.

46. Li,X., Li,D., Yin,Y. and Park,K.H. (2009) Characterization of a recombinant amylolytic enzyme of hyperthermophilic archaeon *Thermofilum pendens* with extremely thermostable maltogenic amylase activity. *Appl. Microbiol. Biotechnol.*, **85**, 1821–1830.
47. Carrard,G., Koivula,A., Soderlund,H. and Beguin,P. (2000) Cellulose-binding domains promote hydrolysis of different sites on crystalline cellulose. *Proc. Natl Acad. Sci. USA*, **97**, 10342–10347.
48. Blake,A.W., McCartney,L., Flint,J.E., Bolam,D.N., Boraston,A.B., Gilbert,H.J. and Knox,J.P. (2006) Understanding the biological rationale for the diversity of cellulose-directed carbohydrate-binding modules in prokaryotic enzymes. *J. Biol. Chem.*, **281**, 29321–29329.
49. Najmudin,S., Guerreiro,C.I., Carvalho,A.L., Prates,J.A., Correia,M.A., Alves,V.D., Ferreira,L.M., Romão,M.J., Gilbert,H.J., Bolam,D.N. *et al.* (2006) Xyloglucan is recognized by carbohydrate-binding modules that interact with beta-glucan chains. *J. Biol. Chem.*, **281**, 8815–8828.
50. Charnock,S.J., Bolam,D.N., Turkenburg,J.P., Gilbert,H.J., Ferreira,L.M., Davies,G.J. and Fontes,C.M. (2000) The X6 "thermostabilizing" domains of xylanases are carbohydrate-binding modules: structure and biochemistry of the *Clostridium thermocellum* X6b domain. *Biochemistry*, **39**, 5013–5021.
51. McCartney,L., Blake,A.W., Flint,J., Bolam,D.N., Boraston,A.B., Gilbert,H.J. and Knox,J.P. (2006) Differential recognition of plant cell walls by microbial xylan-specific carbohydrate-binding modules. *Proc. Natl Acad. Sci. USA*, **103**, 4765–4770.
52. Boraston,A.B., McLean,B.W., Chen,G., Li,A., Warren,R.A. and Kilburn,D.G. (2002) Co-operative binding of triplicate carbohydrate-binding modules from a thermophilic xylanase. *Mol. Microbiol.*, **43**, 187–194.
53. Whitney,S.E.C., Wilson,E., Webster,J., Bacic,A., Reid,J.S.G. and Gidley,M.J. (2006) Effects of structural variation in xyloglucan polymers on interactions with bacterial cellulose. *Am. J. Botany*, **93**, 1402–1414.
54. Mongodin,E.F., Shapir,N., Daugherty,S.C., DeBoy,R.T., Emerson,J.B., Shvartzbeyn,A., Radune,D., Vamathevan,J., Riggs,F., Grinberg,V. *et al.* (2006) Secrets of soil survival revealed by the genome sequence of *Arthrobacter aurescens* TC1. *PLoS Genet.*, **2**, e214.
55. Quentin,Y., Fichant,G. and Denizot,F. (1999) Inventory, assembly and analysis of *Bacillus subtilis* ABC transport systems. *J. Mol. Biol.*, **287**, 467–484.
56. Nataf,Y., Yaron,S., Stahl,F., Lamed,R., Bayer,E.A., Scheper,T.H., Sonenshein,A.L. and Shoham,Y. (2009) Cellodextrin and laminaribiose ABC transporters in *Clostridium thermocellum*. *Biochem. J.*, **422**, 73–82.
57. Audy,J., Labrie,S., Roy,D. and Lapointe,G. (2010) Sugar source modulates exopolysaccharide biosynthesis in *Bifidobacterium longum* subsp. longum CRC 002. *Microbiology*, **156**, 653–664.
58. Stabler,R., He,M., Dawson,L., Martin,M., Valiente,E., Corton,C., Lawley,T., Sebaihia,M., Quail,M., Rose,G. *et al.* (2009) Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. *Genome Biology*, **10**, R102, R102.104–R102.115.
59. McNeil,M., Daffe,M. and Brennan,P.J. (1990) Evidence for the nature of the link between the arabinogalactan and peptidoglycan components of mycobacterial cell walls. *J. Biol. Chem.*, **281**, 18200–18206.
60. Ito,Y., Tomita,T., Roy,N., Nakano,A., Sugawara-Tomita,N., Watanabe,S., Okai,N., Abe,N. and Kamio,Y. (2003) Cloning, expression, and cell surface localization of *Paenibacillus sp.* strain W-61 xylanase 5, a multidomain xylanase. *Appl. Environ. Microbiol.*, **69**, 6969–6978.
61. Joris,B., Englebert,S., Chu,C.P., Kariyama,R., Daneo-Moore,L., Shockman,G.D. and Ghuysen,J.M. (1992) Modular design of the *Enterococcus hirae* muramidase-2 and *Streptococcus faecalis* autolysin. *FEMS Microbiol. Lett.*, **70**, 257–264.
62. Devillard,E., Goodheart,D.B., Karnati,S.K., Bayer,E.A., Lamed,R., Miron,J., Nelson,K.E. and Morrison,M. (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.
63. Ali,M.K., Kimura,T., Sakka,K. and Ohmiya,K. (2001) The multidomain xylanase Xyn10B as a cellulose-binding protein in *Clostridium stercorarium*. *FEMS Microbiol. Lett.*, **198**, 79–83.
64. Bunai,K., Ariga,M., Inoue,T., Nozaki,M., Ogane,S., Kakeshita,H., Nemoto,T., Nakanishi,H. and Yamane,K. (2004) Profiling and comprehensive expression analysis of ABC transporter solute-binding proteins of *Bacillus subtilis* membrane based on a proteomic approach. *Electrophoresis*, **25**, 141–155.
65. Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
66. Sivashankari,S. and Shanmughavel,P. (2006) Functional annotation of hypothetical proteins—a review. *Bioinformation*, **1**, 335–338.
67. Janga,S.C., Salgado,H. and Martínez-Antonio,A. (2009) Transcriptional regulation shapes the organization of genes on bacterial chromosomes. *Nucleic Acids Res.*, **37**, 3680–3688.