Adv Biochem Eng Biotechnol
DOI: 10.1007/10_2015_5007

Comparative Genomics of Core Metabolism Genes of Cellulolytic and Non-cellulolytic *Clostridium* Species

Sadhana Lal and David B. Levin

Abstract Microbial production of fuels such as ethanol, butanol, hydrogen (H_2), and methane (CH_4) from waste biomass has the potential to provide sustainable energy systems that can displace fossil fuel consumption. Screening for microbial diversity and genome sequencing of a wide-range of microorganisms can identify organisms with natural abilities to synthesize these alternative fuels and/or other biotechnological applications. *Clostridium* species are the most widely studied strict anaerobes capable of fermentative synthesis of ethanol, butanol, or hydrogen directly from waste biomass. *Clostridium termitidis* CT1112 is a mesophilic, cellulolytic species capable of direct cellulose fermentation to ethanol and organic acids, with concomitant synthesis of H_2 and CO_2. On the basis of 16S ribosomal RNA (*rRNA*) and chaperonin 60 (*cpn*60) gene sequence data, phylogenetic analyses revealed a close relationship between *C. termitidis* and *C. cellobioparum*. Comparative bioinformatic analyses of the *C. termitidis* genome with 18 cellulolytic and 10 non-cellulolytic *Clostridium* species confirmed this relationship, and further revealed that the majority of core metabolic pathway genes in *C. termitidis* and *C. cellobioparum* share more than 90% amino acid sequence identity. The gene loci and corresponding amino acid sequences of the encoded enzymes for each pathway were correlated by percentage identity, higher score (better alignment), and lowest e-value (most significant "hit"). In addition, the function of each enzyme was proposed by conserved domain analysis. In this chapter we discuss the comparative analysis of metabolic pathways involved in synthesis of various useful products by cellulolytic and non-cellulolytic biofuel and solvent producing *Clostridium* species. This study has generated valuable information concerning the core metabolism genes and pathways of *C. termitidis* CT1112, which is helpful in developing

S. Lal and D.B. Levin (✉)
Department of Biosystems Engineering, University of Manitoba, Winnipeg, MB, Canada R3T 5V6
e-mail: david.levin@umanitoba.ca

metabolic engineering strategies to enhance its natural capacity for better industrial applications.

Keywords Biomass, CAZymes, Cellulolytic bacteria, Cellulosomes, Genetic manipulation

Contents

Abbreviations

16S rRNA	16S ribosomal RNA
ACT	Artemis comparison tool
CAZymes	Carbohydrate-active enzymes
CDS	Coding sequence
COG/COGs	Clusters of orthologous groups
Cpn60	Chaperonin 60
cpnDB	Chaperonins database
GenePRIMP	Gene PRediction IMprovement Pipeline
IMG-ER	Integrated microbial genomes-expert review
JGI	Joint Genome Institute
KEGG	Kyoto encyclopedia of genes and genomes
NCBI	National Center for Biotechnology Information
PEP	Phosphoenolpyruvate
PPDK	Pyruvate phosphate dikinase
PPK	Pyruvate directly by pyruvate kinase
RDP	Ribosomal database project
UT	Universal target

1 Introduction

Low-cost, high-throughput sequencing is now being used extensively for investigating bacterial genomes. In silico screening of microbial diversity has enabled analyses of genetic variability and can help identify organisms with natural ability to synthesize fuels and bio products of interest. High throughput sequencing has also revealed how bacteria acquire and transmit genes via horizontal gene transfer over short periods of time. Such organisms can be exploited through genome shuffling for transgenic expression and efficient generation of clean fuel and other diverse biotechnological applications.

Clostridium is the second largest genus within the Phylum *Firmicutes*. *Clostridium* species are present in diverse environments and are known to utilize a wide-variety of substrates and to synthesize a great diversity of metabolites, such as ethanol, butanol, hydrogen, carbon dioxide, acetate, formate, lactate, propionate, butyrate, isobutyrate, and isovalerate [1, 2] (Table 1). The genus *Clostridium* consists of Gram-positive obligate anaerobes and contains common free-living bacteria and important pathogens. The great metabolic capabilities of *Clostridium* species have attracted researchers to investigate their diverse metabolisms for biofuel production and cellulose degradation. As of May 2015, the Integrated Microbial Genomes (IMG) system listed a total of 534 genome sequences for species in the genus *Clostridium* [40]. Of these, 56 genomes were listed as "finished genomes," 33 were listed as "draft genomes," and 410 were listed as "permanent drafts." IMG also lists 13 "finished genomes" of *Clostridium* plasmids and 22 "finished genomes" of bacteriophages which infect *Clostridium* species. Of the bacterial genome sequences listed, 247 sequences were from *C. difficile*, 39 were from *Clostridium* species, 34 were from *C. botulinum*, 25 were from *C. perfringens*, and 10 were from *C. clostridioforme*.

In this chapter we have discussed the draft genome sequence of *C. termitidis* and analyzed the genomes of 18 cellulolytic and 10 non-cellulolytic *Clostridium* species which synthesize fermentation end-products of interest as potential fuels [41] (Table 1). *C. termitidis* strain CT1112 (DSM 5398) was isolated from the gut of the wood-feeding termite *Nasutitermes lujae* from the Mayombe tropical rainforest, Congo, Central Africa [21]. *C. termitidis* was examined for H_2 and other end-products such as acetate, CO_2, formate, and ethanol formation on cellobiose, α-cellulose, xylan, xylose, and glucose by Ramachandran et al. [22] in our laboratory. Experimental data showed the produced amount of H_2, ethanol, acetate, formate, and CO_2 was comparable to other cellulolytic *Clostridium* species. Thus, *C. termitidis* could be used as a potential candidate for biofuel (H_2 and ethanol) production from biomass through consolidated bioprocessing [22, 42]. Comparative analysis of experimental data revealed that *C. termitidis* is distinct from *C. cellulolyticum*, which takes more time to grow on cellulose and produces less H_2 than ethanol during cellulose fermentation [22].

Identifying the core metabolic pathways using an experimental approach in any organisms is a major challenge and a time-consuming process. In this study we used

Table 1 Genome metadata for sequenced cellulolytic and non-cellulolytic *Clostridium* species

NCBI Ref Seq	Genome name	Fermentation product	Isolation	Gene size	GC %	Gene count	CDS count	16s rRNA count	COG count	References
Cellulolytic *Clostridium* species										
NC_014393.1[a]	*C. cellulovorans* 743B, ATCC 35296	H_2, CO_2, acetate, butyrate, formate, lactate, ethanol	Woody biomass digester	5.26	30	4,500	4,389	9	2,514	[3–5]
NC_011898.1[a]	*C. cellulolyticum* H10	H_2, CO_2, ethanol, acetate, lactate, formate	Decayed grass in compost pile (packaged for 3–4 months)	4.07	37	3,575	3,488	8	2,036	[6, 7]
NC_014376.1[a]	*C. saccharolyticum* WM1, DSM 2544	H_2, CO_2, ethanol	methanogenic cellulose-enrichment culture of sewage sludge	4.66	45	4,388	4,299	6	2,785	[8]
NC_010001.1[a]	*C. phytofermentans* ISDg	H_2, CO_2, ethanol, acetate, formate, lactate	Forest soil near the Quabbin Reservoir in Massachusetts	4.8	35	3,991	3,902	8	2,497	[9]
NC_009012.1[a]	*C. thermocellum* ATCC 27405[c]	Ethanol, organic acids	Yellowstone National Park	3.84	39	3,335	3,236	4	1,892	[10, 11]
NC_017304.1[a]	*C. thermocellum* LQ8, DSM 1313[c]	Ethanol, organic acids	Evolved mixed population	3.56	39	3,102	3,031	4	1,814	[12]
NC_016627.1[a]	*C. clariflavum* EBR 45, DSM 19732[c]	H_2, CO_2, lactate, acetate, ethanol, formate	Methanogenic sludge of a cellulose-degrading bioreactor	4.89	36	4,229	4,131	6	2,294	[13, 14]

NC_020134.1[a]	*Clostridium stercorarium* DSM 8532[c]	H_2, CO_2, ethanol, acetate, lactate, L-alanine	Compost	2.97	42	2,763	2,706	3	1,750	[15–17]
CP003259.1[a]	*Clostridium* species BNL1100	biofuel production	Corn stover enrichment culture	4.61	37	4,117	4,025	8	2,372	[18]
NZ_ACXX00000000.2[b]	*C. papyrosolvens* DSM 2782	H_2, CO_2, ethanol, acetate, lactate	Estuarine sediment, River Don, Aberdeenshire, Scotland	4.87	37	4,479	4,423	1	2,393	[19, 20]
NZ_AORV00000000.1[b]	*C. termitidis* CT1112, DSM 5398	H_2, CO_2, acetate, formate, lactate, ethanol	Gut of termite, *Nasutitermes lujae*, Congo, Africa	6.42	41	5,389	5,327	2	3,392	[21–23]
AUVG00000000.1[b]	*C. cellulosi* CS-4-4[c]	H_2, CO_2, ethanol, acetic acid		5.68	42	5,216	5,106	4	3,374	[24]
NZ_JHYD00000000.1[b]	*C. cellobioparum* ATCC 15832	H_2, CO_2, ethanol, acetic acid	Rumen of cattle, 1998, L. Ganqiu	6.13	41	5,220	5,134	4	3,224	[25]
2509887034[b] (from JGI/IMG)	*C. alkalicellulosi* Z-7026, DSM 17461	Cellulose degrader		5.31	32	4,473	4,390	2	2,214	[26]
2541046955[b] (from JGI/IMG)	*C. josui* JCM 17888[c]	H_2, CO_2, ethanol, acetate, butyrate	Thai compost	4.47	36	3,991	3,895	7	2,274	[27]
Genome and cpn60 not available	*C. chartatabidum*	H_2, ethanol, acetate, butyrate	Chloroform treated rumen contents	–	–	–	–	–	–	[28]
Genome not available	*C. aldrichii*	H_2, CO_2, acetate, propionate, butyrate, isobutyrate, isovalerate, lactate, succinate	Wood fermenting anaerobic digester	–	–	–	–	–	–	[29]

(continued)

Table 1 (continued)

NCBI Ref Seq	Genome name	Fermentation product	Isolation	Gene size	GC %	Gene count	CDS count	16s rRNA count	COG count	References
Genome and cpn60 not available	*C. celerecrescens* DSM 5628	H_2, CO_2, ethanol, acetate, formate, butyrate, isobutyrate, isovalerate, caproate, lactate, succinate	Methanogenic cellulose-enriched culture	–	–	–	–	–	–	[30]
Non-cellulolytic *Clostridium* species										
NC_003030.1[a]	*C. acetobutylicum* ATCC 824	Gelatin hydrolysis positive	Plant-derived foodstuff (corn meal)	4.13	31	4,022	3,848	11	2,589	[31]
NC_020291.1[a]	*C. saccharoperbutylacetonicum* N1-4(HMT)	Acetone, butanol, ethanol producer	Derived from existing strain (derived from ATCC 13564)	6.67	30	5,981	5,843	11	3,652	[32]
NC_009706.1[a]	*Clostridium kluyveri* DSM 555	H_2, butyrate, caproate	Mud of a canal in Delft, The Netherlands	4.02	32	4,073	3,913	7	2,261	[33]
NC_022571.1[a]	*C. saccharobutylicum* DSM 13864	Solvent producers such as acetone, butanol, ethanol	Soya bean field	5.11	29	4,552	4,430	12	2,667	[31]
NC_021182.1[a]	*C. pasteurianum* BC1	Nitrogen fixation	Coal-cleaning residues	5.04	31	4,966	4,851	9	2,914	[34]

NC_014328.1[a]	*C. ljungdahlii* PETC, DSM 13528	Acetogen, ethanol production	Chicken yard waste	4.63	31	4,283	4,184	9	2,662	[35]
NC_022592.1[a]	*C. autoethanogenum* DSM 10061	Ethanol and acetate	Enriched from rabbit feces and isolated using carbon monoxide as the sole source of energy and carbon	4.35	31	4,131	4,042	9	2,577	[36]
NC_021047.1[a]	*C.* cf. *saccharolyticum* K10	Metablolize a wide range of sugars	Mixed cellulolytic culture and relies on the cellulolytic microorganism to provide sugars for growth	3.77	50	3,124	3,073	1	1,675	–
NZ_ADEK00000000.1[b]	*C. carboxidivorans* P7, DSM 15243	Acetogenic, solvent producer	Agricultural settling lagoon at Oklahoma State University, Aquatic, USA	4.41	30	4,234	4,174	2	2,503	[37, 38]
NZ_APJA00000000.1[b]	*C. intestinale* URNW	Hydrogen, ethanol production	Contaminated cellobiose stock solution	4.67	30	4,660	4,549	10	2,941	[39]

[a]Sequencing status is "finished" (F)
[b]Sequencing status is "permanent draft" (PD)
[c]Thermophilic *Clostridium* species ; remaining *Clostridium* species are mesophilic

an in silico approach to explore the core metabolic pathways in a draft genome sequence of *C. termitidis* using *C. cellulolyticum* and *C. thermocellum* as reference organisms, as they are well studied for their metabolic pathways in comparison to other cellulolytic *Clostridium* species. We have selected six cellulolytic *Clostridium* species known for their ability to produce H_2, CO_2, acetate, ethanol, formate, lactate, or butyrate for comparative genomic analysis, which revealed the presence or absence of enzymes involved in glycolysis, pyruvate formation and catabolism, acetate, ethanol, and hydrogen synthesis. The information generated by the computational approach concerning the core metabolism genes and pathways of *C. termitidis* CT1112 could be helpful in developing metabolic engineering strategies to enhance the natural capacity of *C. termitidis* for better industrial applications. This study demonstrates that comparative genomics analysis is a very useful tool to generate large amount of highly informative data in less time, allowing quick functional prediction for many hypothetical or putative proteins in poorly studied organisms [43].

2 Genome Annotation and Phylogenetic Analysis of *Clostridium* Species

2.1 *Annotation and Manual Curation of Draft Genome of* C. termitidis *Strain CT1112*

The genome of *C. termitidis* CT1112 was sequenced by the Genome Québec/McGill University platform using a Roche/454s GS-FLX Titanium sequencer by a whole-genome shotgun strategy. A 454 standard flow-gram format (.sff) read file was assembled using Newbler v2.3. The assembled draft genome of *C. termitidis* strain CT1112 was deposited to the Joint Genome Institute's (JGI) Integrated Microbial Genomes-Expert Review (IMG-ER) platform for annotation using their annotation pipeline (http://img.jgi.doe.gov) [40] which generates protein coding genes (CDS) and assigns names to gene products. The annotated genome was subsequently submitted to the JGI's Gene Prediction Improvement pipeline (GenePRIMP) [44] which allows automated correction of genes, including insertion of missed genes, extension of "short" genes, and identification of putative pseudogenes [45]. GenePRIMP generated 360 anomalies of different types, including short genes, long genes, unique genes, dubious genes, split genes interrupted by frameshifts and stopcodons ("broken genes"), split genes interrupted by transposases "(interrupted genes"), missed genes, missed CrispR elements, and overlaps. Anomalies generated by the GenePRIMP pipeline were manually curated using Artemis Comparison Tool (ACT) [46]. Manual correction of 1,847 genes was accomplished using several lines of evidence, including BLAST searches of

nucleotides (BLAST-N), protein (BLAST-P), conserved domain analyses using the Conserved Domain Database (CDD), protein domains and motifs, Interpro (Pfam, prosite, SMART etc.), and characterized proteins in some cases (e.g., UNIPROT entries).

The manually curated draft genome has been deposited at DDBJ/EMBL/GenBank under the accession no. NZ_AORV00000000.1 [23]. The final draft genome of *C. termitidis* strain CT1112 has 78 contigs with a total size of 6,415,858 bp and G + C content of 41.18%. It was predicted to encode 5,389 genes, with 5,327 CDSs (98.85% of predicted genes). The coding region covers 87.94% of the genome sequence. Out of 5,327 CDSs, 4,403 were assigned with functions, although no functional prediction could be assigned to 924 CDSs. Among the 4,403 CDSs, 3,392 (62.94%) genes could be classified into COG families. A total of seven rRNA genes, including three 5S rRNAs, two 16S rRNAs (243 bp and 1,342 bp, of which 1,342 bp was used in phylogenetic tree), and two 23S rRNAs, were present on the *C. termitidis* chromosome. In addition, 55 tRNA genes that represent all 20 amino acids were identified.

2.2 *Genome Features of* C. termitidis *Strain CT1112 and Other* Clostridium *Species*

Comparative analyses of a range of genome features were conducted with 18 cellulolytic and 10 non-cellulolytic sequenced genomes of *Clostridium* species (Table 1). The genomic information included genome size, % G + C, total numbers of genes, numbers of coding sequences (CDS), numbers of Clusters of Orthologous Groups (COG), and 16S rRNAs derived from the IMG-ER platform (http://img.jgi.doe.gov; Table 1).

Comparative analyses of general features of *Clostridium* species revealed that the non-cellulolytic bacterium, *C. saccharoperbutylacetonicum* N1-4(HMT) had the largest genome size (6.67 Mb) with highest number of protein-encoding genes (5,843) and COG count (3,652), whereas the thermophilic, cellulolytic bacterium, *C. stercorarium*, had the smallest genome size (2.97 MB) with the lowest number of protein coding genes (2,706). *C. termitidis* had the largest genome size (6.42 Mb) of the cellulolytic *Clostridium* species [23, 42]. The finished genome of *C.* cf. *saccharolyticum* K10 had the highest G + C content (50%), with the lowest number of protein coding genes (3,073) and COG count (1,675). Interestingly, in this analysis we observed that the 16S rRNA count was higher in non-cellulolytic *Clostridium* species than in cellulolytic *Clostridium* species (Table 1). In addition, the average percentage G + C in non-cellulolytic *Clostridium* species (30%) was lower than the percentage G + C in cellulolytic *Clostridium* species (36%), except for *C. cellulovorans* 743B, which has 30% G + C [5].

2.3 *Phylogenetic Analyses of Cellulolytic and Non-cellulolytic* Clostridium *Species*

16S *rRNA* sequences have been used for phylogenetic studies and for sequence-based taxonomy for many years [47–50]. The dynamic nature of the genomes and the impact of lateral gene transfer on genomic evolution [51] have forced researchers to use more than one conserved target to understand the taxonomy and phylogenetics of bacterial diversity. The chaperonin-60 universal target (*cpn60* UT, also known as *GroEL* or *HSP60*) nucleotide sequence (549–567 bp), is highly conserved in bacteria. It can differentiate even more closely related isolates of the same bacterial species and thus find more reliable targets for phylogenetic studies, microbial identification, microbial ecology, and evolution [50, 52, 53] than 16S rRNA [54–56]. *Cpn60* UT sequence alignments have been shown to correlate to whole genome sequence alignments and resolve ambiguities associated with 16S rDNA gene phylogeny in bacteria [54].

To determine the evolutionary relationship between *C. termitidis* and sequenced 18 cellulolytic and 10 non-cellulolytic *Clostridium* species (Table 1), a phylogenetic tree was constructed based on 16S *rRNA* and *cpn60* universal target (UT) gene sequences. The 16S *rRNA* sequences chosen for this study were retrieved from the ribosomal database (RDP) (http://rdp.cme.msu.edu/) [57] and NCBI (http://www.ncbi.nlm.nih.gov/) [58]. The *cpn60* "universal target" (UT) sequences were retrieved from a Chaperonins database (cpnDB) (http://www.cpndb.ca/cpnDB/home.php) [50] and IMG database (http://img.jgi.doe.gov) [59]. The 16S rRNA and *cpn60* sequences of *Acidothermus cellulolyticus* was used as an out group for phylogenetic analysis.

16S rRNA and *cpn60* sequences of 28 cellulolytic and non-cellulolytic *Clostridium* species were aligned using the BioEdit v.7.0.9.0 program [60]. Phylogenetic trees were generated using the PHYLIP 3.69 package [61]. Evolutionary distances between all species were calculated with the DNADIST and the resultant distance matrix was then used to draw Neighbor Joining trees with the program NEIGHBOR [62]. The program SEQBOOT was used for statistical testing of the trees by resampling the dataset 1,000 times [62]. The trees were viewed through TreeView Version 1.6.6 [63]. Phylogenetic analyses with 16S *rRNAs* revealed that *C. termitidis* strain CT1112 is closely related to *C. cellobioparum* DSM 1351 (99%) and with other cellulolytic *Clostridium* species (Fig. 1). Moreover, the 16S *rRNA* phylogenetic tree showed that cellulolytic and non-cellulolytic species are very well separated, except for *C. cellulovorans* and *C. chartatabidum*. Interestingly, cellulolytic *C. cellulovorans* and *C. chartatabidum* clustered with non-cellulolytic species, such as *C. intestinale* URNW, *C. saccharoperbutylacetonicum* N1-4(HMT), and *C. saccharobutylicum* DSM 13864, in Clade 2. Thus, the 16S *rRNA* phylogeny suggests that *C. cellulovorans* and *C. chartatabidum* are different from other cellulolytic *Clostridium* species [2]. This observation is consistent with a previous study in

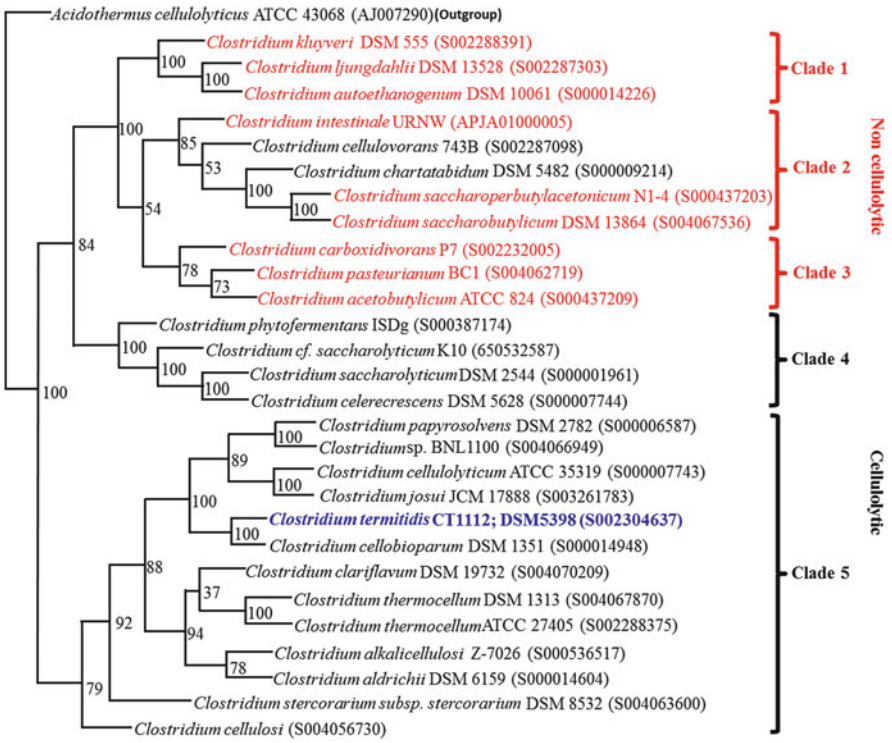

Fig. 1 Neighbor-joining tree based on 16S *rRNA* gene sequences, showing the phylogenetic position of *C. termitidis* with respect to species of the genus *Clostridium*. The non-cellulolytic *Clostridium* species are highlighted in *red* and the cellulolytic *Clostridium* species are indicated in *black*. The evolutionary history of the taxa analyzed was represented using 1,000 replicates obtained from the bootstrap consensus tree. Accession numbers are given in *parentheses*

which the genome sequence of *C. cellulovorans* 743B was compared with other *Clostridium* species [5] and revealed the G + C content of *C. cellulovorans* (30.0%) was very similar to the G + C content of the non-cellulolytic species, *C. acetobutylicum* ATCC 824 (30.9%).

Phylogenetic trees based on *cpn60* of 26 *Clostridium* species, clearly separated the cellulolytic and non-cellulolytic *Clostridium* species into two groups, except for the cellulolytic bacterium *C. cellulolvorans* 743B, which clustered with the non-cellulolytic species *C. acetobutylicum* ATCC 824 in Clade 3. Non-cellulolytic *Clostridium* species in the *cpn60* tree clustered into four Clades (1, 2, 3, and 4), whereas cellulolytic *Clostridium* species grouped into five Clades (5, 6, 7, 8, and 9; Fig. 2). The placement of *C. cellulovorans* with non-cellulolytic *Clostridium* species detected in the 16S *rRNA* tree is clearly supported by the *cpn60* tree (Fig. 2).

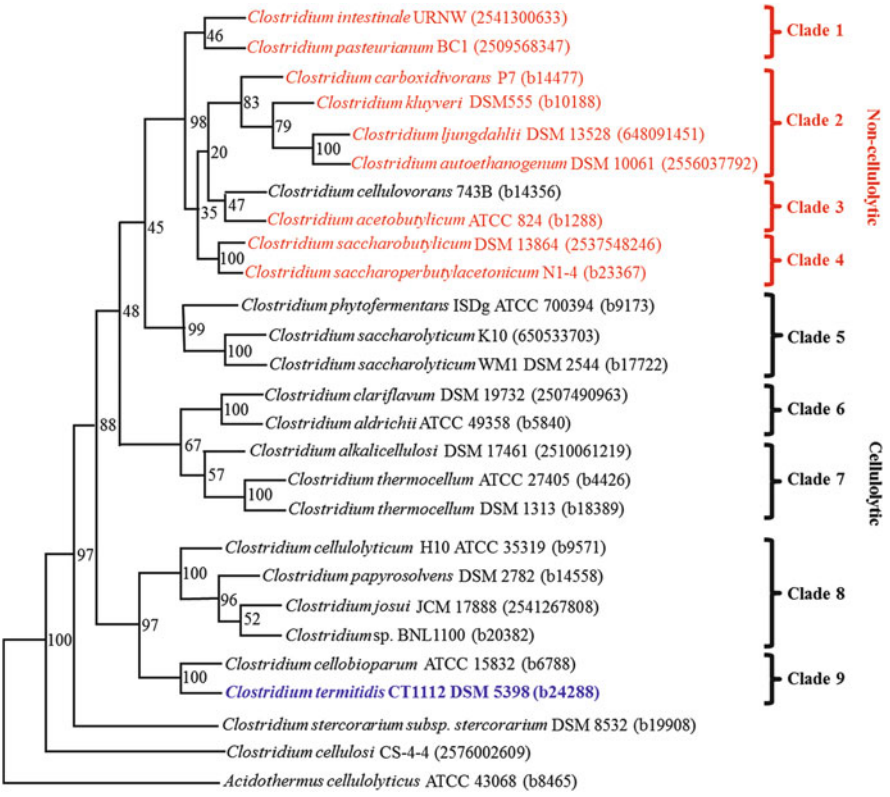

Fig. 2 Neighbor-joining tree based on UT region of the *cpn60* gene sequences, showing the phylogenetic position of *C. termitidis* with respect to species of the genus *Clostridium*. The non-cellulolytic *Clostridium* species are highlighted in *red* and the cellulolytic *Clostridium* species are indicated in *black*. The evolutionary history of the taxa analyzed was represented using 1,000 replicates obtained from the bootstrap consensus tree. Accession numbers are given in *parentheses*

3 Comparative Analysis of Core Metabolisms in *Clostridium* Species

The major fermentation and cellulose degradation end-products commonly observed in cellulolytic *Clostridium* species are H_2, CO_2, acetate, formate, and ethanol, whereas most of the non-cellulolytic *Clostridium* species are solvent producers that synthesize H_2, CO_2, butyrate, butanol, acetone, and ethanol (Table 1).

We have selected two cellulolytic *Clostridium* species – mesophilic *C. cellulolyticum* H10 and thermophilic *C. thermocellum* ATCC 27405 – as references to search for homologous genes of interest from glycolysis, pyruvate formation and catabolism, acetate, ethanol, and hydrogen synthesis in six *Clostridium* species. Some enzymes in certain pathways were found to be missing in *C. cellulolyticum* and, for these specific cases, we used *C. thermocellum* ATCC 27405 as the reference organism.

The corresponding gene loci and enzymes for each pathway in *C. termitidis*, *C. cellobioparum*, *C. phytofermentans*, *C. cellulovorans*, and *C. papyrosolvens* were identified using different tools from the IMG-ER platform. These tools included alignments against BLAST-P and RPS-BLAST using the conserved domain database [64] from NCBI and Pfams [65] and IMG Terms [40]. Amino acid sequences for each gene product were retrieved from the JGI genome portal (http://genome.jgi-psf.org/ [45]) and the NCBI database (http://www.ncbi.nlm.nih.gov/genomes). Confirmation of CDS features of *C. termitidis* was accomplished by comparing with six cellulolytic *Clostridium* species – *C. cellulolyticum* H10, *C. thermocellum* ATCC 27405, *C. phytofermentans* ISDg, *C. cellulovorans* 743B, *C. cellobioparum* ATCC 15832, and *C. papyrosolvens* DSM 2782.

The functions of predicted genes were manually assessed using different databases such as Clusters of Orthologous Groups (COG) [66], KEGG Orthology (KO) assignments [67], and TIGRFAMs [68]. Clusters of Orthologous Groups (COGs) analyses were conducted using "Abundance Profiles" tools from IMG, which provide a comprehensive examination of the functional components of genomes between strains. Genome clustering was done using cluster 3.0 analysis software [69] within the IMG-ER platform using the COG profile for each selected clostridial genome. In this analysis, genome sequences of cellulolytic mesophilic (*C. termitidis*, *C. cellulotylicum* H10, *C. cellobioparum* ATCC 15832, *C. cellulovorans* 743B, *C. phytofermentans* ISDg, *C. papyrosolvens* DSM 2782) and thermophilic (*C. thermocellum* ATCC 27405) *Clostridium* species were used for comparative analyses of core metabolic pathways. The corresponding protein sequences for enzymes were retrieved from the KEGG, JGI/IMG, and NCBI databases. Manual construction of metabolic pathways from the annotated *C. termitidis* CT1112 genome was accomplished using pathways from the KEGG database as a reference [70].

Genome sequence analysis of *C. termitidis* revealed the presence of all open reading frame (ORFs) encoding proteins for all metabolic pathways, except pyruvate phosphate dikinase (PPDK) when compared with *C. cellulolyticum* genome. However, a gene encoding PPDK was found in the *C. termitidis* genome when *C. thermocellum* was used as the reference organism. The absence of PPDK in *C. termitidis* with *C. cellulolyticum* was observed because of the sequence variation in functional domain. Amino acid sequence of PPDK from *C. termitidis* showed 85% identity when compared with the *C. thermocellum* PPDK gene. Some genes encoding phosphoenolpyruvate synthase, the γ- and δ-subunits of pyruvate ferredoxin oxidoreductase and aldehyde dehydrogenase enzymes, were absent in the *C. cellulolyticum* genome. In such cases, *C. thermocellum* was used as a reference to search for these enzymes in other *Clostridium* species (Table 2).

3.1 Pyruvate Metabolism and End-Product Synthesis

For in silico comparative analysis, it is very important to select more than one reference organism. The gene loci and sequences corresponding to each enzyme

Table 2 Comparative analysis of core metabolisms in cellulolytic *Clostridium* species

Metabolic enzymes for different metabolisms	EC no.	Locus tag							Amino acid identity (%)					
		Ccel_	CpapDRAFT_	T344DRAFT_	Cter_	Cthe_	Clocel_	Cphy_	Ccel/ CpapDRAFT	Ccel/ T344DRAFT	Ccel/ Cter	Ccel/ Cthe	Ccel/ Clocel	Ccel/ Cphy
Glycosis														
Cellobiose phosphorylase	2.4.1.20	2109	3781	03023	4494	0275	0032	0430	94	88	88	76	71	66
Glucose kinase/ROK domain containing protein	2.7.1.2	0700	3365	01639	3170	2938	3765	0329	86	75	74	47	32	29
ROK domain containing protein	2.7.1.2	3430	3726	3251	4330	0390	0594	3420	91	75	75	31	30	24
Glucose-6-phosphate isomerase	5.3.1.9	1445	2002	4897	3865	0217	1364	0419	96	90	90	79	68	66
6-Phosphofructokinase	2.7.1.11	2223	3751	00156	4719	0347	1603	3345	96	95	95	83	45	48
6-Phosphofructokinase	2.7.1.11	2612	4116	05203	0067	1261	0388	0336	96	90	90	67	54	56
Fructose-1,6-bisphosphate aldolase, class II	4.1.2.13	2222	3750	00157	4718	0349	0553	3683	97	93	93	85	36	38
Triose phosphate isomerase	5.3.1.1/ 2.7.2.3	2260	2776	00088	4786	0138	0720	2875	96	91	91	83	62	59
Glyceraldehyde-3-phosphate dehydrogenase	1.2.1.12	2275	4374	00072	4809	0137	0719	2876	96	96	95	84	49	45
Phosphoglycerate mutase	5.4.2.1	2259	2775	00089	4785	0140	0722	2868	95	88	87	76	61	63
Phosphoglycerate mutase	5.4.2.1	0619	0778	03891	2329	0946	3764	1900	87	74	74	67	33	30
Enolase	4.2.1.11	2254	2767	00096	4779	0143	0730	3001	96	93	93	85	70	67
Pyruvate formation														
Pyruvate kinase	2.7.1.40	2569	0002	03464	0649	–	0389	0741	90	83	83	–	53	56
Pyruvate phosphate dikinase[a]	2.7.9.1	2388	–	00787	–	–	–	–	–	29	–	–	–	–
		–	–	01049	0809	1308	1454	0651	–	86	85	–	74	68
Phosphoenolpyruvate synthase/pyruvate phosphate dikinase[a]	2.7.9.2	–	–	02136	5054	1253	–	–	–	46	44	–	–	–
Phosphoenolpyruvate carboxykinase [GTP]	4.1.1.32	0212	4012	01445	1146	2874	–	–	96	91	91	83	–	–
Oxaloacetate carboxykinase (α-subunit)	4.1.1.3	1736	0378	02498	0730	0701	2828	2433	92	86	87	84	35	67

Malate dehydrogenase	1.1.1.37	0137	4023	01233	0412	0345	2700	1117	94	91	90	65	45	40
Malic enzyme (NAD)	1.1.1.40	0138	4022	01232	0411	0344	0393	0409	95	95	89	77	57	60
Pyruvate catabolism														
L-Lactate dehydrogenase	1.1.1.27	2485	0740	00623	2504	1053	1533	1117	96	86	86	72	49	39
Pyruvate ferredoxin oxidoreductase, gamma subunit[a]	1.2.7.1	–	3734	02440	1025	2390	2840	–	32	88	88	–	26	–
Pyruvate ferredoxin oxidoreductase, delta subunit[a]	1.2.7.1	–	3735	02439	1024	2391	–	3558	46	74	74	–	–	37
Pyruvate ferredoxin oxidoreductase, alpha subunit[a]	1.2.7.1	1164	3736	02438	1023	2392	1684	0603	37	79	79	30	31	30
Pyruvate ferredoxin oxidoreductase, beta subunit[a]	1.2.7.1	1164	3737	02437	1022	2393	2840	3558	43	84	84	29	29	29
Pyruvate flavodoxin/ferredoxin oxidoreductase domain protein	1.2.1.51	1164	3182	03616	3589	3120	2840	0603	94	84	84	64	63	67
Pyruvate flavodoxin/ferredoxin oxidoreductase domain protein	1.2.1.51	0016	3065	03616	3589	3120	1684	3558	94	55	55	55	57	61
Pyruvate formate lyase	2.3.1.54	2582	0044	03479	0038	0505	1811	2823	95	90	90	80	61	67
Acetate and ethanol production														
Aldehyde dehydrogenase[a]	1.2.1.3	–	–	02407	0993	2238	0629	3041	–	43	43	–	26	46
Alcohol dehydrogenase	1.1.1.1	1083	2833	03763	1650	0423	3817	1029	91	82	82	39	48	48
Alcohol dehydrogenase	1.1.1.1	3337	1270	04004	2586	0101	1140	2463	92	49	49	54	42	43
Alcohol dehydrogenase	1.1.1.1	0894	3919	03683	3280	0394	4197	1029	95	83	86	30	34	31
Bifunctional acetaldehyde-alcohol dehydrogenase(adhE)	1.1.1.1/ 1.2.1.10	3198	2940	01310	5426	0423	2402	3925	96	88	88	75	67	76
Acetate kinase	2.7.2.1	2136	4304	03043	4518	1028	1892	1327	93	83	83	70	56	60
Phosphotransacetylase	2.3.1.8	2137	4305	03044	4519	1029	1891	1326	98	89	89	73	53	65
Acetate thiokinase	6.2.1.1	0494	2620	02662	5570	0551	0888	–	93	85	85	78	30	–

Ccel, *C. cellulolyticum* H10; CTER, *C. termitidis* CT1112, DSM 5398; Cthe27405, *C. thermocellum* ATCC 27405; Clocel, *C. cellulovorans* 743B, ATCC 35296; Cphy, *C. phytofermentans* ISDg; CpapDRAFT, *C. papyrosolvens* DSM 2782; T344DRAFT, *C. cellobioparum* ATCC 15832

[a]*C. thermocellum* ATCC 27405 used as a reference

obtained for each pathway were confirmed by the percent amino acids sequence identity, higher score (better alignment), and lowest e-value (most significant hit). In addition, the functionality of each enzyme was verified by conserved domain analysis. Here, we mainly discuss the comparative analysis of genes encoding proteins involved in pyruvate metabolism and end-product synthesis, xylose production, and hydrogen (H_2) production in seven cellulolytic *Clostridium* species.

Glucose is an important carbon source used by bacteria to synthesize a wide-range of metabolic intermediates used in many biosynthetic reactions. During metabolism, glucose can be stored as a polysaccharide, converted to sucrose, oxidized to pyruvate via glycolysis, or oxidized to pentose sugars via the pentose phosphate pathway (Fig. 4a–d). Pyruvic acid can be converted back to glucose via gluconeogenesis, or to acetyl-CoA which is a branch-point precursor for many biosynthesis pathways. Pyruvate can be converted to alanine or to citric acid in the presence of oxygen, whereas it ferments to produce lactic acid using the enzyme lactate dehydrogenase (CTER_2504) and NADH in the absence of oxygen. Alternatively, it is converted to acetaldehyde and then to ethanol in alcoholic fermentation.

Carbon and electrons are distributed between catabolic, anabolic, and energy-producing pathways in bacterial cells. In the glycolytic pathway, carbon and electrons flow from phosphoenolpyruvate (PEP) to the various end-products via different nodes, such as the Phosphoenolpyruvate/Oxaloacetate/Pyruvate node, the Acetyl-CoA/Acetate/Ethanol node, and the Ferredoxin/NAD(P)H/Hydrogen node [71, 72]. Several different enzymes involved in the conversion of intermediate metabolites and the presence of corresponding genes which encode these proteins in *C. termitidis* have been identified in the annotated genome (Fig. 3). Phosphoenolpyruvate (PEP) is converted to pyruvate directly by pyruvate kinase (PPK) or pyruvate phosphate dikinase (PPDK) (CTER_0809). PEP can also be converted to pyruvate via the "malate shunt" [73] using phosphoenolpyruvate carboxykinase (PEPCK; CTER_1146), malate dehydrogenase (MDH; CTER_0412), and malic enzyme (MalE; CTER_0411) or via other reactions using PEPCK and oxaloacetate decarboxylase (OAADC; CTER_0730). Conversion of pyruvate to lactate via NADH-dependent lactate dehydrogenase diverts reducing equivalents away from hydrogen and ethanol. The reducing equivalents NADH and reduced ferredoxin (Fdr) are generated by the oxidative decarboxylation of pyruvate to acetyl-CoA via pyruvate dehydrogenase (PDH) or pyruvate:ferreodoxin oxidoreductase (Pfor), respectively. NADH and Fdr are oxidized to NAD^+ and oxidized ferredoxin (Fdo) during hydrogen and ethanol synthesis. Another important enzyme in pyruvate catabolism is pyruvate formate lyase (PFL; CTER_0038), which converts pyruvate to acetyl-CoA and producing formate during the process.

The presence or absence of these enzymes was determined through comparative genomic analyses in *C. cellulolyticum*, *C. cellobioparum*, *C. phytofermentans*, C. *papyrosolvens*, *C. cellulovorans*, and *C. thermocellum* (summarized in Table 2). All genes associated with glycolytic pathways are observed in the *C. termitidis* genome. Seven copies of cellobiose phosphorylase, two copies of glucose kinase, three copies of 6-phosphofructokinase, and two copies of phosphoglycerate mutase were identified in the *C. termitidis* genome, and the functional domains of these

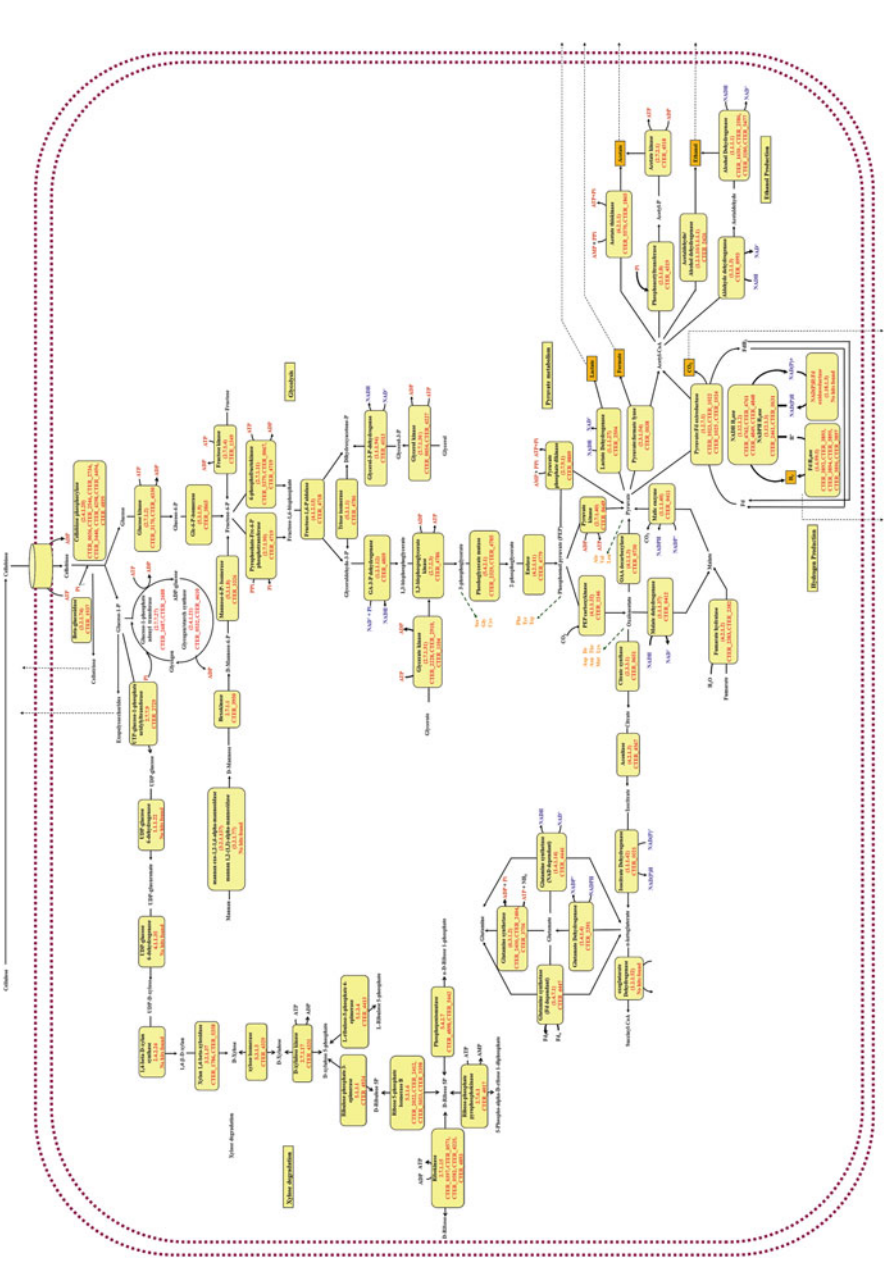

Fig. 3 Schematic representation of the core metabolic pathways in *Clostridium termitidis* indicating the flow of carbon from cellobiose degradation through glycolysis to pyruvate catabolism to synthesis of H_2 and ethanol

enzymes were confirmed by RPS-BLAST using the conserved domain database (CDD) [65].

3.2 Hydrogen Synthesis

Hydrogen (H_2) is an essential component in the metabolism of many microorganisms. *Clostridium* species are the most widely studied microorganisms for fermentative H_2 production [73–75]. Fermentative H_2 production is a process in which reversible reduction of proton to dihydrogen is catalyzed by hydrogenases [76, 77]. Hydrogenases have highly reactive and complex metallocenters and synthesize H_2 more efficiently than nitrogenases [78, 79]. On the basis of the metallocenter, hydrogenases are divided into three major classes: (1) [NiFe] hydrogenases which contain Ni and Fe atoms; (2) [FeFe] hydrogenases, which contain two Fe atoms bound to cysteine residues in their active sites; [FeFe] hydrogenases may be dimeric, trimeric, or even tetrameric enzymes, and size variations were observed because of the presence of additional domain accommodating different number of FeS cluster [80]; and (3) [Fe] hydrogenases with one Fe atom [81]. [FeFe] hydrogenases exist in multiple forms with different modular structures and are mostly observed in the genus *Clostridium*, whereas [NiFe] hydrogenases have been reported in many bacteria, archaea, and in a few *Clostridium* species. [Fe] hydrogenases are mainly reported in methanogenic archaea [82].

Sequences encoding [NiFe] and [FeFe] hydrogenases, which are homologous to *C. cellulolyticum* hydrogenases, have been searched in the genome sequences of seven cellulolytic *Clostridium* species. This analysis revealed the presence of a small subunit of genes encoding a [NiFe] hydrogenase and Cytochrome b5 only in *C. cellulolyticum*, *C. papyrosolvens*, and *C. cellulovorans*. The small subunit of [NiFe] hydrogenases contains FeS clusters in their active sites, which transfers electrons between the catalytic center of enzyme and the electron donors. The large subunit of [NiFe] hydrogenases contains the Ni-Fe cluster in the active site. Genes encoding the [NiFe] hydrogenase large subunit were observed in *C. cellulolyticum*, *C. termitidis*, *C. cellobioparum*, *C. cellulovorans*, *C. phytofermentans*, *C. papyrosolvens*, and *C. thermocellum*.

Cellulolytic *Clostridium* species were also investigated for Ech hydrogenases. The complex Ech hydrogenases have six membrane-bound subunits, and the genes encoding these subunits are organized into operons. Genes corresponding to these membrane-bound subunits were found in six of seven *Clostridium* species investigated and were absent in the genome of *C. cellulovorans*. Homologs of [FeFe] hydrogenases were also investigated in the seven *Clostridium* species. Amino acid sequences of hydrogenases from *C. cellulolyticum* showed highest percent identity with *C. papyrosolvens*, *C. termitidis*, and *C. cellobioparum* [83]. The locus tags of these predicted enzymes are included in Table 3.

This analysis showed that the genome of *C. termitidis* has all the hydrogenase and pyruvate-ferredoxin oxidoreductase encoding genes responsible for H_2 synthesis (Fig. 4c). Ramachandran et al. [22] reported that rate of H_2 synthesis by

Table 3 Comparative analysis of hydrogenases in cellulolytic *Clostridium* species

Remarks	EC no.	Predicted localization	Locus tag							Amino acid identity (%)					
			Ccel_	CpapDRAFT_	CTER_	T344DRAFT_	Cphy_	Cthe27405_	Clocel_	Ccel/ CpapDRAFT	Ccel/ Cter	Ccel/ T344DRAFT	Ccel/ cphy	Ccel/ Cthe	Ccel/ Clocel
[NiFe] HYDROGENASE															
Cytochrome b5	–	U	1069	2861	–	–	–	–	0269	74	–	–	–	–	50
Hydrogenase (NiFe) small SU HydA	1.12.99.6	CM	1070	2860	–	–	–	–	1155	88	–	–	–	–	52
Nickel-dependent hydrogenase large SU	1.6.99.5, 1.12.5.1	C	1071	2859	3893	02875	1734	3020	1156	84	31	32	33	30	51
4Fe-4S ferredoxin, iron-sulfur binding (EchF)	1.6.99.5	C	3366	1309	3892	02874	1735	3019	0104	92	82	82	46	50	40
Ech hydrogenase subunit E	1.6.99.5	C	3367	1310	3893	02875	1734	3020	–	96	90	89	62	62	–
Ech hydrogenase subunit D	1.6.99.5	C	3368	1311	3894	02876	1733	3021	–	91	79	78	37	44	–
Ech hydrogenase subunit C	1.6.99.5	C	3369	1312	3895	02877	1732	3022	–	97	92	92	72	73	–
Ech hydrogenase subunit B	1.6.99.5	CM (10 IH)	3370	1313	3896	02878	1731	3023	–	93	88	88	49	46	–
Ech hydrogenase subunit A	1.6.99.5	CM (18 IH)	3371	1314	3897	02879	1730	3024	–	93	85	84	41	41	–
NADH ubiquinone oxidoreductase 20 kDa SU	1.6.99.5	CM	1686	1634	0018	02571	1732	3022	–	88	73	73	47	47	–
NADH dehydrogenase (ubiquinone) 30 kDa SU	1.6.99.5	C	1687	1635	0657	02570	1734	3020	–	90	78	78	34	35	–
Hydrogenase, membrane SU 2-like protein	1.6.99.5	CM (13 IH)	1688	1636	0658	02569	1730	3024	–	93	78	78	25	26	–
Hydrogenase, membrane SU 2-like protein	1.6.99.5	CM (7 IH)	1689	1637	0659	02568	–	–	–	89	77	77	–	–	–
Respiratory-chain NADH dehydrogenase, SU 1	1.6.99.5	CM (8 IH)	1690	1638	0660	02567	1731	3023	–	88	74	74	24	26	–

(continued)

Table 3 (continued)

Remarks	EC no.	Predicted localization	Locus tag							Amino acid identity (%)					
			Ccel_	CpapDRAFT_	CTER_	T344DRAFT_	Cphy_	Cthe27405_	Clocel_	Ccel/ CpapDRAFT	Ccel/ Cter	Ccel/ T344DRAFT	Ccel/ cphy	Ccel/ Cthe	Ccel/ Clocel
NADH/Ubiquinone/plastoquinone (complex I)	1.6.99.5	CM (18 IH)	1691	1639	0661	02566	1730	3024	–	82	62	62	28	27	–
[FeFe] HYDROGENASE															
Putative PAS/PAC sensor protein	1.12.7.2	C	2300	3820	4846	00036	0092	0426	3813	94	79	79	41	42	26
Putative uncharacterized protein	–	C	2301	3821	4847	00035	0093	0431	2244	94	83	83	44	46	35
Hydrogenase, Fe-only	1.12.7.2	C	2303	3823	4848	00034	0087	0430	2243	94	85	85	57	56	44
NADH dehydrogenase (quinone)	1.6.99.5	C	2304	3824	4849	00033	0088	0429	2588	94	88	89	67	69	56
NADH dehydrogenase (ubiquinone), 24 kDa subunit	1.6.99.5	C	2305	3825	4850	00032	0089	0428	2583	92	85	86	65	68	60
Hydrogenase, Fe-only	1.12.7.2	C	2467	0711	2461	00655	3805	0342	2243	94	89	89	54	63	63
Hydrogenase, Fe-only	1.12.7.2	C	2232	2734	4761	00113	3805	0342	2243	95	87	87	60	79	56
Pyruvate ferredoxin oxidoreductase, delta subunit	1.6.99.5	C	2233	2735	4762	00112	3804	0341	2244	95	90	91	72	82	62
Ferredoxin	1.6.99.3, 1.4.1.13	C(IH)	2234	2736	4763	00111	3803	0340	2244	88	71	70	50	69	30
Ferredoxin	1.6.99.3, 1.4.1.13	C	2545	2719	0631	03446	2934	0372	2971	94	91	91	65	81	52

Ccel, *C. cellulolyticum* H10; CTER, *C. termitidis* CT1112, DSM 5398; Cthe27405, *C. thermocellum* ATCC 27405; Clo1313, *C. thermocellum* LQ8, DSM 1313; Cphy, *C. phytofermentans* ISDg; T344DRAFT, *C. cellobioparum* ATCC 15832; CpapDRAFT, *C. papyrosolvens* DSM 2782; Clocel, *C. cellulovorans* 743B, ATCC 35296; C, cytoplasmic; CM, cytoplasmic membrane (IH, number of internal membrane helices); U, unknown [83]

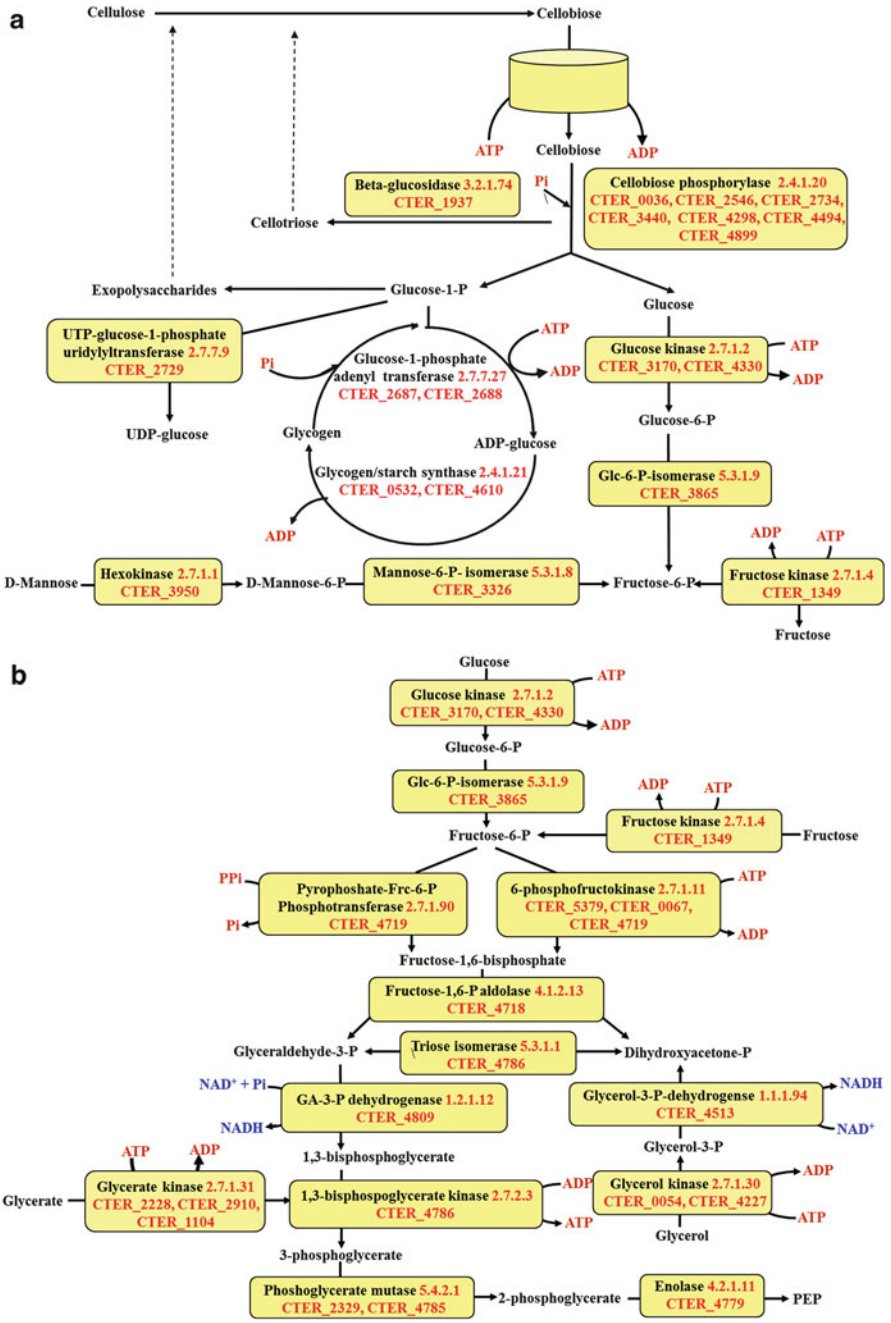

Fig. 4 Metabolic pathways in *Clostridium termitidis*. Detailed schematic representations of: (**a**) cellobiose degradation; (**b**) glycolysis; (**c**) pyruvate catabolism and H_2 synthesis; (**d**) ethanol synthesis. The KEGG database was used as the reference for construction of pathways. EC numbers and locus tags for each enzyme are written in *red*. "No hits found" means genes encoding these enzymes were not observed in *C. termitidis* genome

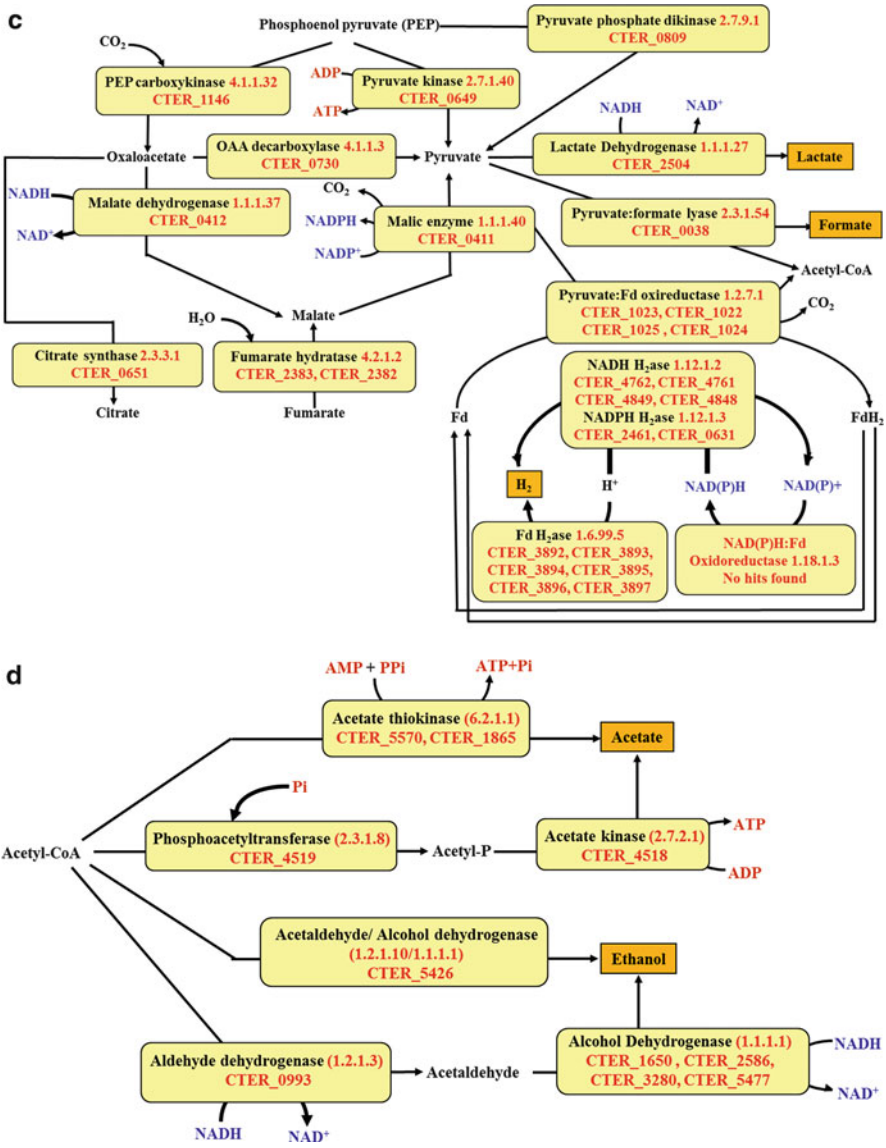

Fig. 4 (continued)

C. termitidis strain CT1112 was comparable to that of the cellulolytic bacterium *C. cellulolyticum*. *C. termitidis* was observed to synthesize greater amounts of H_2 than ethanol when cultured with either cellobiose or α-cellulose, even though the rates of growth of *C. termitidis* with these substrates were vastly different. *C. termitidis* grows faster using cellobiose (doubling-time of 6.5 h) than α-cellulose (doubling-time of 18.9 h). Moreover, a metabolic shift from ethanol to acetate and H_2 synthesis and a trend toward lower H_2:CO_2 ratios were observed when the pH dropped below 6.2 during fermentation.

3.3 Xylose Utilization

Many bacteria in the Phylum *Firmicutes*, including *Clostridium* species, are able to utilize xylose as a carbon source. Depolymerization of xylan and xyloglucan, which are major constituents of hemicellulose in plant cell walls, produces α- and β-xylosides, respectively. These compounds are transported into the cell and converted into D-xylose and then transformed to xylulose 5-phosphate. It has already been reported that *C. termitidis* can utilize different types of simple and complex carbohydrates such as cellulose, cellobiose, xylose, glucose, and mannose and produces acetate, formate, ethanol, H_2, and CO_2 during fermentation [21–23, 72]. Recently, Munir et al. [42] reported that *C. termitidis* is also able to grow well on xylan polymers. A computational search, based on gene homology, was used to determine the complete xylose degradation and pentose phosphate pathway in *C. termitidis* (Fig. 3). This analysis showed that *C. termitidis* contains most D-xylose pathway genes.

In the xylose utilization pathway, the first reaction is the conversion of D-xylose into D-xylulose by xylose isomerase (EC 5.3.1.5) (CTER_4329). In the second reaction, phosphorylation of D-xylulose to D-xylulose 5-phosphate via xylulokinase (EC 2.7.1.17) (CTER_4331) was observed as a key intermediate in the pentose phosphate pathway [84]. D-Xylulose 5-phosphate is finally converted into D-ribose, 5-phospho-α-D-ribose 1-diphosphate, L-ribulose 5-phosphate, and α-D-ribose 1-phosphate via several steps. All locus tags belonging to these enzymes involved in the xylose pathway are present in *C. termitidis* (Fig. 5).

Enzymes involved in glutamine synthesis were also screened and hits for almost all the enzymes were found in *C. termitidis* (Fig. 6). Glutamine synthetase (GS) (EC 6.3.1.2) is an enzyme that plays an important role in the metabolism of nitrogen by catalyzing the condensation of glutamate and ammonia to form glutamine. Glutamine synthetase uses ammonia produced by nitrate reduction, amino acid degradation, and photorespiration [85] The amide group of glutamate is a nitrogen source for the synthesis of glutamine pathway metabolites [86] (Fig. 6).

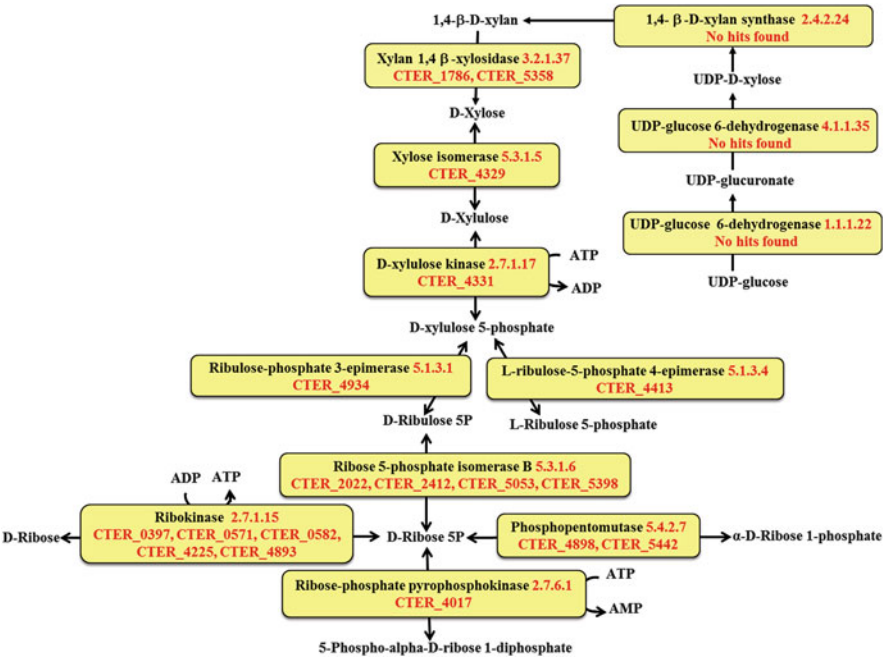

Fig. 5 Detailed schematic representation of xylose catabolism in *C. termitidis*. The KEGG database was used as the reference for construction of pathways. EC numbers and locus tags for each enzyme are written in *red*. "No hits found" means genes encoding these enzymes were not observed in *C. termitidis* genome

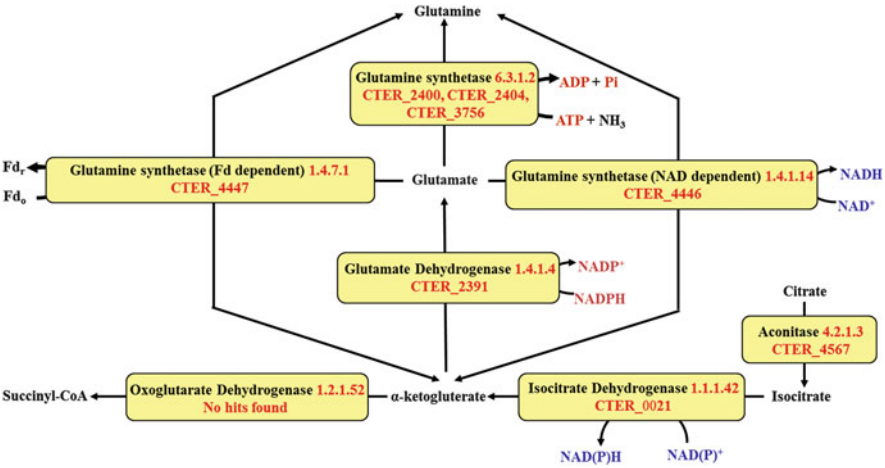

Fig. 6 Detailed schematic representation of glutamine synthesis in *C. termitidis*. The KEGG database was used as the reference for construction of pathways. EC numbers and locus tags for each enzyme are written in *red*. "No hits found" means genes encoding these enzymes were not observed in *C. termitidis* genome

4 Comparative Analysis of Whole Genome of *Clostridium* Species

4.1 Comparative Synteny Dot Plot Analyses of Clostridial Genomes

Comparative synteny dot plot analyses based on protein sequences were carried out using *C. termitidis* and 8 cellulolytic *Clostridium* species (Fig. 7). In this analysis, pairwise comparisons of the *C. termitidis* genome and each of the other genomes

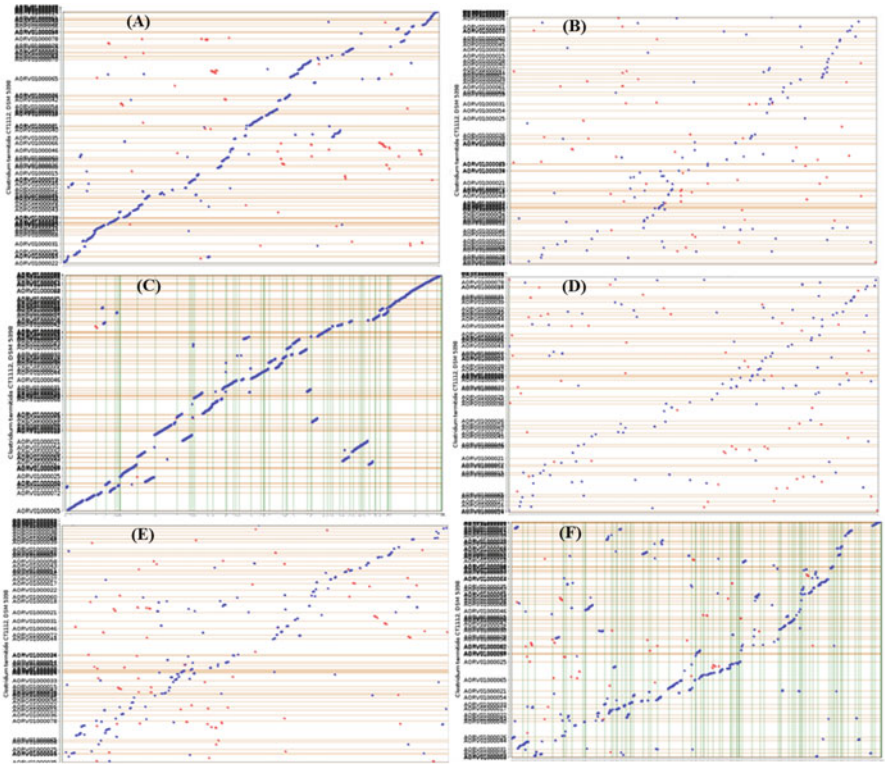

Fig. 7 Comparative synteny *dot* plots between the *C. termitidis* genome and the genomes of other *Clostridiu*m species. The synteny plots reveal orthologous relationships between *C. termitidis* and six *Clostridium* species. (**a**) *C. termitidis* CT1112, DSM 5398 vs *C. cellulolyticum* H10. (**b**) *C. termitidis* CT1112 vs *C. cellulovorans* 743B, ATCC 35296. (**c**) *C. termitidis* CT1112 vs *C. cellobioparum* ATCC 15832. (**d**) *C. termitidis* CT1112 vs *C. phytofermentans* ISDg. (**e**) *C. termitidis* CT1112 vs *C. papyrosolvens* C7. (**f**) *C. termitidis* CT1112 vs *C. thermocellum* ATCC 27405. Comparison was based on amino acid sequences. The *blue points* in the *dot plots* represent the regions of similarity found on parallel strands (fplot) and the *red points* represent the regions of similarity found on antiparallel strands (rplot). The tooltip for each point shows which plot it is and the coordinates and scaffolds of the alignment

was accomplished using Mummer to generate dot plot diagrams [87]. The diagonal line in Fig. 7 show the co-linearity of DNA strands. The blue points in the dot plots represent the regions of similarity found on parallel strands and the red points represent the regions of similarity found on antiparallel strands. On the basis of genome arrangement, *C. termitidis* showed high synteny (co-linearity) with *C. cellobioparum* and *C. cellulolyticum*, but displayed markedly different genome organization from *C. cellulovorans*, *C. papyrosolvens*, *C. phytofermentans*, and *C. thermocellum*.

4.2 *Whole Genome Comparisons to Identify Orthologous Genes*

Genomes can be compared in terms of gene content using the "Phylogenetic Profiler" which allows one to identify genes in a query genome in terms of presence or absence of homologs in other genomes. Whole genome comparisons were conducted to find orthologous genes in 24 genomes of cellulolytic and non-cellulolytic *Clostridium* species using the *C. termitidis* genome as the reference organism. The analysis was based on the selection of "best gene homologs" between the *C. termitidis* genome and other *Clostridium* species genomes at a 60–90% amino acid sequence identity level. Comparative analysis revealed the highest number of orthologs was observed between *C. termitidis* and *C. cellobioparum*. This observation was supported by phylogenetic analyses based on 16S *rRNA*, *cpn60*, and COG functional profiles. No homologs (0) were found with *C. acetobutylicum* ATCC 824 and *C.* cf. *saccharolyticum* K10 at 90% identity level (Table 4).

4.3 *Comparative Analysis of COGs in Eight* Clostridium *Species*

An analysis of Clusters of Orthologous Groups (COGs) was conducted using the "Abundance Profiles" tools in IMG, which provide a comprehensive examination of the functional components of genomes between strains. In other words, "Abundance Profiles" are based on annotation profiles (e.g., COGs, Pfam, Enzyme EC Numbers, and TIGRfams) to compare and contrast the genome content. In this analysis, eight *Clostridium* species – *C. termitidis* CT1112, DSM 5398, *C. cellulolyticum* H10, *C. thermocellum* ATCC 27405, *C. thermocellum* LQ8, *C. acetobutylicum* ATCC 824, *C. cellulovorans* 743B, ATCC 35296, *C. papyrosolvens* DSM 2782, and *C. phytofermentans* ISDg – were used for comparative COG matrix analysis. On the basis of this analysis, 2,362 common COG families were identified in the genomes of the 8 *Clostridium* species. Some

Table 4 Pairwise whole genome comparison for "best" homologs in cellulolytic and non-cellulolytic *Clostridium* species

Cellulolytic and non-cellulolytic *Clostridium* species	Reference genome (Cter[a]) vs query genome	Genome identity (%)			
		60+	70+	80+	90+
		Number of homologs			
Clostridium cellobioparum ATCC 15832	Cter-Ccellobio	4,317	4,268	4,228	4,138
C. papyrosolvens DSM 2782	Cter-Cpap	1,947	1,449	791	190
Clostridium species BNL1100	Cter-Clo1100	1,944	1,474	812	218
C. josui JCM 17888	Cter-Cjos	1,830	1,400	781	214
C. cellulolyticum H10	Cter-Ccel	1,775	1,342	739	165
C. cellulosi CS-4-4	Cter-CcelCS-4-4	849	306	48	1
C. clariflavum EBR 45, DSM 19732	Cter-Cclari	824	441	128	8
C. thermocellum ATCC 27405	Cter-Cthe	775	417	137	6
C. thermocellum LQ8, DSM 1313	Cter-Cthe1313	773	415	137	6
C. alkalicellulosi Z-7026, DSM 17461	Cter-Calkali	747	360	111	11
C. stercorarium stercorarium DSM 8532	Cter-Cster	512	207	36	2
C. cellulovorans 743B, ATCC 35296	Cter-Ccel743B	364	113	19	1
C. phytofermentans ISDg	Cter-Cphy	325	107	26	1
C. saccharolyticum WM1, DSM 2544	Cter-CsacWM1	300	79	15	1
C. saccharoperbutylacetonicum N1-4(HMT)	Cter-CsacN1	448	136	21	2
C. pasteurianum BC1	Cter-Cpast	386	132	19	1
C. intestinale URNW	Cter-Cinte	371	139	21	2
C. kluyveri DSM 555	Cter-Cklu	352	109	17	1
C. carboxidivorans P7, DSM 15243	Cter-Ccarboxi	341	106	18	2
C. ljungdahlii PETC, DSM 13528	Cter-Cljung	320	97	19	2
C. saccharobutylicum DSM 13864	Cter-Csac13864	318	100	16	2
C. acetobutylicum ATCC 824	Cter-Cace	315	95	16	0
C. autoethanogenum DSM 10061	Cter-Cauto	312	98	18	2
C. cf. *saccharolyticum* K10	Cter-CsacK10	179	46	9	0

[a]Cter, *C. termitidis*

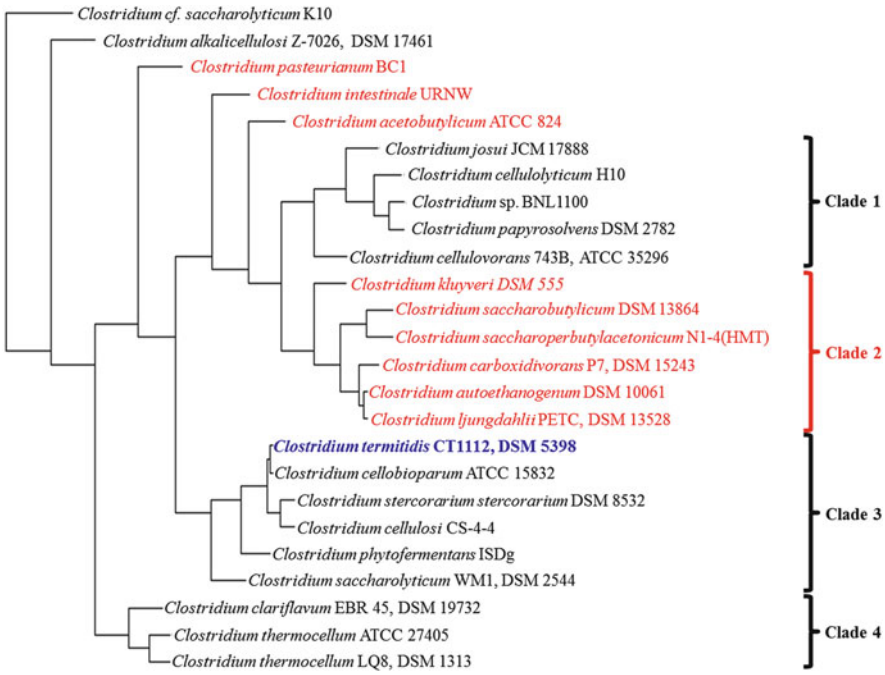

Fig. 8 Phylogenetic tree based on annotated COG functional profiles for 25 sequenced cellulolytic and non-cellulolytic *Clostridium* species genomes. Clustering is based on Cluster 3.0 analysis software [70], and was performed within the IMG-ER platform using the COG profile for each of the genome. Branch lengths correspond to calculated distances between functional profiles. The tree was generated using the Archaeopteryx applet on IMG platform

COGs were present in all species, some were present in more than one copy in some genomes, and some were absent in one or more genomes. Out of 2,362 COG families, 694 COG families were absent in the *C. termitidis* genome. *C. termitidis* genome showed 3,392 protein coding genes with COG families, which was greater than all other *Clostridium* species (Ccel-2036; Cthe27405-1892; Cthe1313-1814; Clocel-2514; Cphy-2497, and Cac-2648).

An analysis of hierarchical clustering based upon COGs was conducted with the genomes of 25 cellulolytic and non-cellulolytic *Clostridium* species to determine whether similar functional profiles among the species could be identified. This analysis revealed four distinct Clades within the genus. Clades 1, 3, and 4 contained the cellulolytic *Clostridium* species, whereas Clade 2 contained non-cellulolytic solventogenic *Clostridium* species (Fig. 8).

5 Conclusions

Although the genus *Clostridium* contains a large number of species with seeming great genetic variation and metabolic capabilities, comparative genomic analyses enabled rapid visualization of the evolutionary relationships among cellulolytic and non-cellulolytic species. Phylogenetic analyses of 16S *rRNA* and *cpn60* genes have suggested a close evolutionary relationship between *C. termitidis* and *C. cellobioparum*. This observation was supported by whole genome comparisons. Genome analyses also revealed that *C. termitidis* has the largest genome (6.42 Mb) [23] of all the cellulolytic *Clostridium* species studied and its genome encodes all enzymes required for pyruvate metabolism and catabolism, xylose utilization, ethanol synthesis, and hydrogen synthesis. Our comparative genomic analyses identified homologs of [NiFe] and [FeFe] hydrogenases in seven cellulolytic *Clostridium* species.

Genomic analyses further determined that the *C. termitidis* genome encodes the greatest number of carbohydrate active enzymes (CAZymes) in comparison to other cellulolytic *Clostridium* species. These CAZymes have the potential ability to degrade a wide variety of complex and simple carbohydrates, such as cellulose, hemicellulose, starch, chitin, fructans, pectin, glucose, cellobiose, and xylose, thus making *C. termitidis* an attractive microorganism for biofuel production through CBP [42]. Although many of these CAZymes have homologs in other bacteria, a multidomain GH5 protein, Cter_2817, seems to be unique to *C. termitidis* because BLAST searches did not give any hits in other *Clostridium* species with this protein. This protein has a modular structure CBM66-CBM66-CBM66-GH5_distGH43-CBM35-CBM66-GH43-SLH-SLH-SLH and would be putatively bound to the cell via the SLH domains. In addition to the non-cellulosomal enzymes, *C. termitidis* has been shown to harbor the genes and express the products of cellulosomal components and CAZymes, suggesting cellulosome assembly [42, 88]. In conclusion, computational approaches and comparative genomic analyses can facilitate deep insight into the genetic basis of metabolic pathways involved in synthesis of various useful products by cellulolytic and non-cellulolytic biofuel and solvent producing *Clostridium* species. Searching pathways using in silico approaches generates valuable information concerning the presence or absence of the genes involved in particular pathways in a much shorter time. On the basis of this information, we can divert or delete the particular pathway to manipulate or engineer any organisms to enhance the production of various bioproducts.

Acknowledgements This work was supported by funds from Genome Canada through the Applied Genomics Research in Bioproducts or Crops (ABC) program for the grant titled "Microbial genomics for biofuels and coproducts from biorefining processes," and the Province of Manitoba from the Manitoba Research Innovation Fund (MRIF).

References

1. Levin DB, Zhu H, Beland M, Cicek N, Holbein BE (2007) Potential for hydrogen and methane production from biomass residues in Canada. Bioresour Technol 98:654–660
2. Zhou C, Ma Q, Mao Z, Yanbin Y, Ying X (2014) New insights into *Clostridia* through comparative analyses of their 40 genomes. BioEnergy Res 7:1481–1492
3. Shoseyov O, Takagi M, Goldstein M, Doi RH (1992) Primary sequence analysis of *Clostridium cellulovorans* cellulose binding protein a (CbpA). Proc Natl Acad Sci U S A 89:3483–3487
4. Tamaru Y, Karita S, Ibrahim A, Chan H, Doi RH (2000) A large gene cluster for the *Clostridium cellulovorans* cellulosome. J Bacteriol 182:5906–5910
5. Tamaru Y, Miyake H, Kuroda K, Nakanishi A, Kawade Y, Yamamoto K, Uemura M, Fujita Y, Doi RH, Ueda M (2010) Genome sequence of the cellulosome-producing mesophilic organism *Clostridium cellulovorans* 743B. J Bacteriol 192:901–902
6. Pagès S, Gal L, Bélaich A, Gaudin C, Tardif C, Bélaich JP (1997) Role of scaffolding protein CipC of *Clostridium cellulolyticum* in cellulose degradation. J Bacteriol 179:2810–2816
7. Petitdemange E, Caillet F, Giallo J, Gaudin C (1984) *Clostridium cellulolyticum* species nov., a cellulolytic, mesophilic species from decayed grass. Int J Syst Bacteriol 34:155–159
8. Murray WD, Khan AW, van den Berg L (1982) *Clostridium saccharolyticum* species nov., a saccharolytic species from sewage sludge. Int J Syst Bacteriol 32:132–135
9. Warnick TA, Methé A, Leschine SB (2002) *Clostridium phytofermentans* species nov., a cellulolytic mesophile from forest soil. Int J Syst Evol Microbiol 52:1155–1160
10. Lamed R, Kenig R, Morgenstern E, Calzada JF, Micheo FD, Bayer EA (1991) Efficient cellulose solubilization by a combined cellulosome β-glucosidase system. Appl Biochem Biotechnol 27:173–183
11. Stevenson DM, Weimer PJ (2005) Expression of 17 genes in *Clostridium thermocellum* ATCC 27405 during fermentation of cellulose or cellobiose in continuous culture. Appl Environ Microbiol 71:4672–4678
12. Feinberg L, Foden J, Barrett T, Davenport KW, Bruce D, Detter C, Tapia R, Han C, Lapidus A, Lucas S, Cheng JF, Pitluck S, Woyke T, Ivanova N, Mikhailova N, Land M, Hauser L, Argyros DA, Goodwin L, Hogsett D, Caiazza N (2011) Complete genome sequence of the cellulolytic thermophile *Clostridium thermocellum* DSM1313. J Bacteriol 193:2906–2907
13. Shiratori H, Sasaya K, Ohiwa H, Ikeno H, Ayame S, Kataoka N, Miya A, Beppu T, Ueda K (2009) *Clostridium clariflavum* species nov. and *Clostridium caenicola* species nov., moderately thermophilic, cellulose-/cellobiose-digesting bacteria isolated from methanogenic sludge. Int J Syst Evol Microbiol 59:1764–1770
14. Izquierdo JA, Goodwin L, Davenport KW, Teshima H, Bruce D, Detter C, Tapia R, Han S, Land M, Hauser L, Jeffries CD, Han J, Pitluck S, Nolan M, Chen A, Huntemann M, Mavromatis K, Mikhailova N, Liolios K, Woyke T, Lynd LR (2012) Complete genome sequence of *Clostridium clariflavum* DSM 19732. Stand Genomic Sci 6:104–115
15. Schwarz W, Bronnenmeier K, Landmann B, Wanner G, Staudenbauer W, Kurose N, Takayama T (1995) Molecular characterization of four strains of the cellulolytic thermophile *Clostridium stercorarium*. Biosci Biotechnol Biochem 59:1661–1665
16. Poehlein A, Hartwich K, Krabben P, Ehrenreich A, Liebl W, Dürre P, Gottschalk G, Daniel R (2013) Complete genome sequence of the solvent producer *Clostridium saccharobutylicum* NCP262 (DSM 13864). Genome Announc 1(6pii):e00997–e001013
17. Schellenberg JJ, Verbeke TJ, McQueen P, Krokhin OV, Zhang X, Alvare GM, Fristensky B, Thallinger GG, Henrissat B, Wilkins JA, Levin DB, Sparling R (2014) Enhanced whole genome sequence and annotation of *Clostridium stercorarium* DSM8532T using RNA-seq transcriptomics and high-throughput. BMC Genomics 15(1):567
18. Li LL, Taghavi S, Izquierdo JA, van der Lelie D (2012) Complete genome sequence of *Clostridium* species strain BNL1100, a cellulolytic mesophile isolated from corn stover. J Bacteriol 194:6982–6983

19. Madden RH, Bryder MJ, Poole NJ (1982) Isolation and characterization of an anaerobic, cellulolytic bacterium *Clostridium papyrosolvens* sp nov. Int J Syst Bacteriol 32:87–91
20. Leschine SB, Canale-Parola E (1983) Mesophilic cellulolytic clostridia from freshwater environments. Appl Environ Microbiol 46:728–737
21. Hethener P, Brauman A, Garcia JL (1992) *Clostridium termitidis* species nov., a cellulolytic bacterium from the gut of the woodfeeding termite, *Nasutitermes lujae*. Syst Appl Microbiol 15:52–58
22. Ramachandran U, Wrana N, Cicek N, Sparling R, Levin DB (2008) Hydrogen production and end-product synthesis patterns by *Clostridium termitidis* strain CT1112 in batch fermentation cultures with cellobiose or α-cellulose. Int J Hydrogen Energy 33:7006–7012
23. Lal S, Ramachandran U, Zhang X, Munir R, Sparling R, Levin DB (2013) Draft genome sequence of the cellulolytic, mesophilic, anaerobic bacterium *Clostridium termitidis* strain CT1112 (DSM 5398). Genome Announc 1(3):e00281–13
24. Yanling H, Youfang D, Yanquan L (1991) Two cellulolytic *Clostridium* species: *Clostridium cellulosi* species nov. and *Clostridium cellulofermentans* species nov. Int J Syst Bacteriol 41:306–309
25. Lamed R, Naimark J, Morgenstern E, Bayer EA (1987) Specialized cell surface structures in cellulolytic bacteria. J Bacteriol 169:3792–3800
26. Zhilina TN, Kevbrin VV, Turova TP, Lysenko AM, Kostrikina NA, Zavarzin GA (2005) *Clostridium alkalicellum* species nov., an obligately alkaliphilic cellulolytic bacterium from a soda lake in the Baikal region. Mikrobiologiia 74:642–653
27. Sukhumavasi J, Ohmiya K, Shimizu S, Ueno K (1988) *Clostridium josui* species nov., a cellulolytic, moderate thermophilic species from Thai compost. Int J Syst Bacteriol 38:179–182
28. Kelly WJ, Asmundson RV, Hopcroft DH (1987) Isolation and characterization of a strictly anaerobic, cellulolytic spore former: *Clostridium chartatabidum* species nov. Arch Microbiol 147:169–173
29. Yang JC, Chynoweth DP, Williams DS, Li A (1990) *Clostridium aldrichii* species nov., a cellulolytic mesophile inhabiting a wood-fermenting anaerobic digester. Int J Syst Bacteriol 40:268–272
30. Palop ML, Valles S, Pinaga F, Flors A (1989) Isolation and characterization of anaerobic, cellulolytic bacterium *Clostridium celerecrescens* species nov. Int J Syst Bacteriol 39:68–71
31. Keis S, Shaheen R, Jones DT (2001) Emended descriptions of *Clostridium acetobutylicum* and *Clostridium beijerinckii*, and descriptions of *Clostridium saccharoperbutylacetonicum* species nov. and *Clostridium saccharobutylicum* species nov. Int J Syst Evol Microbiol 51:2095–2103
32. Poehlein A, Krabben P, Dürre P, Daniel R (2014) Complete genome sequence of the solvent producer *Clostridium saccharoperbutylacetonicum* strain DSM 14923. Genome Announc 2 (5pii):e01056–14
33. Seedorf H, Fricke WF, Veith B, Brüggemann H, Liesegang H, Strittmatter A, Buckel W, Hinderberger J, Hagemeier C, Thauer RK, Gottschalk G (2008) The genome of *Clostridium kluyveri*, a strict anaerobe with unique metabolic features. Proc Natl Acad Sci 105:2128–2133
34. Pyne ME, Utturkar S, Brown SD, Moo-Young M, Chung DA, Chou CP (2014) Improved draft genome sequence of *Clostridium pasteurianum* strain ATCC 6013 (DSM 525) using a hybrid next-generation sequencing approach. Genome Announc 2(4pii):e00790–14
35. Köpke M, Held C, Hujer S, Liesegang H, Wiezer A, Wollherr A, Ehrenreich A, Liebl W, Gottschalk G, Dürre P (2010) *Clostridium ljungdahlii* represents a microbial production platform based on syngas. Proc Natl Acad Sci U S A 107:13087–13092
36. Bruno-Barcena JM, Chinn MS, Grunden AM (2013) Genome sequence of the autotrophic acetogen *Clostridium autoethanogenum* JA1-1 strain DSM 10061, a producer of ethanol from carbon monoxide. Genome Announc 1(4pii):e00628–13
37. Liou JSC, Balkwill DL, Drake GR, Tanner RS (2005) *Clostridium carboxidivorans* species nov., a solvent-producing *Clostridium* isolated from an agricultural settling lagoon,

and reclassification of the acetogen *Clostridium scatologenes* strain SL1 as *Clostridium drakei* species nov. Int J Syst Evol Microbiol 55:2085–2091
38. Bruant G, Lévesque MJ, Peter C, Guiot SR, Masson L (2010) Genomic analysis of carbon monoxide utilization and butanol production by *Clostridium carboxidivorans* strain P7. PLoS One 5(9), e13033
39. Lal S, Ramachandran U, Zhang X, Sparling R, Levin DB (2013) Draft genome sequence of the hydrogen- and ethanol-producing bacterium *Clostridium intestinale* strain URNW. Genome Announc 1(5pii):e00871–13
40. Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC (2009) IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics 25:2271–2278
41. Hemme CL, Mouttaki H, Lee YJ, Zhang G, Goodwin L, Lucas S, Copeland A, Lapidus A, Glavina del Rio T, Tice H, Saunders E, Brettin T, Detter JC, Han CS, Pitluck S, Land ML, Hauser LJ, Kyrpides N, Mikhailova N, He Z, Wu L, Van Nostrand JD, Henrissat B, He Q, Lawson PA, Tanner RS, Lynd LR, Wiegel J, Fields MW, Arkin AP, Schadt CW, Stevenson BS, McInerney MJ, Yang Y, Dong H, Xing D, Ren N, Wang A, Huhnke RL, Mielenz JR, Ding SY, Himmel ME, Taghavi S, van der Lelie D, Rubin EM, Zhou J (2010) Sequencing of multiple clostridial genomes related to biomass conversion and biofuel production. J Bacteriol 192:6494–6496
42. Munir RI, Schellenberg J, Henrissat B, Verbeke TJ, Sparling R, Levin DB (2014) Comparative analysis of carbohydrate active enzymes in *Clostridium termitidis* CT1112 reveals complex carbohydrate degradation ability. PLoS One 9(8), e104260
43. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci 96:2896–2901
44. Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A, Kyrpides NC (2010) GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. Nat Methods 7:455–457
45. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC (2012) IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res 40:D115–D122
46. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA (2012) Artemis: an integrated platform for visualization and analysis of high throughput sequence-based experimental data. Bioinformatics 28:464–469
47. Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986) Microbial ecology and evolution: a ribosomal RNA approach. Annu Rev Microbiol 40:337–365
48. Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. Proc Natl Acad Sci U S A 87:4576–4579
49. Cole JR, Chai B, Marsh TL, Farris RJ, Wang Q, Kulam SA, Chandra S, McGarrell DM, Schmidt TM, Garrity GM, Tiedje JM (2003) The ribosomal database project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res 31:442–443
50. Hill JE, Penny SL, Crowell KG, Goh SH, Hemmingsen SM (2004) CpnDB: a chaperonin sequence database. Genome Res 14:1669–1675
51. Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbo CL, Case RJ, Doolittle WF (2003) Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet 37:283–328
52. Jian W, Zhu L, Dong X (2001) New approach to phylogenetic analysis of the genus Bifidobacterium based on partial HSP60 gene sequences. Int J Syst Evol Microbiol 51:1633–1638
53. Hill JE, Seipp RP, Betts M, Hawkins L, Van Kessel AG, Crosby WL, Hemmingsen SM (2002) Extensive profiling of a complex microbial community by high-throughput sequencing. Appl Environ Microbiol 68:3055–3066

54. Verbeke TJ, Sparling R, Hill JE, Links MG, Levin D, Dumonceaux TJ (2011) Predicting relatedness of bacterial genomes using the chaperonin-60 universal target (cpn60 UT): application to *Thermoanaerobacter* species. Syst Appl Microbiol 34:171–179
55. Verbeke TJ, Dumonceaux TJ, Wushke S, Cicek N, Levin DB, Sparling R (2011) Isolates of *Thermoanaerobacter thermohydrosulfuricus* from decaying wood compost display genetic and phenotypic microdiversity. FEMS Microbiol Ecol 78:473–487
56. Brousseau R, Hill JE, Préfontaine G, Goh SH, Harel J, Hemmingsen SM (2001) *Streptococcus suis* serotypes characterized by analysis of chaperonin 60 gene sequences. Appl Environ Microbiol 67:4828–4833
57. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM (2014) Ribosomal database project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42(Database issue):D633–D642
58. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
59. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner A, Huang J, Woyke T, Huntemann M, Anderson I, Billis K, Varghese N, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res 42:D560–D567
60. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98
61. Felsenstein J (1993) Phylip (Phylogeny Inference Package) version 3.57c. Department of Genetics, University of Washington, Seattle. http://evolution.genetics.washington.edu/phylip.html
62. Saitou N, Nei M (1987) The neighbor-joining method – a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425
63. Page RDM (1996) TreeView: an application to display phylogenetic trees on personal computer. Comput Appl Biosci 12:357–358
64. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH (2015) CDD: NCBI's conserved domain database. Nucleic Acids Res 43(D):222–226
65. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K et al (2010) The Pfam protein families database. Nucleic Acids Res 38: D211–D222
66. Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36
67. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36:D480–D484
68. Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families. Nucleic Acids Res 31:371–373
69. de Hoon MJL, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20:1453–1454
70. Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acids Res 30:42–46
71. Sauer U, Eikmanns BJ (2005) The PEP-pyruvate-oxaloacetate node as the switch point for carbon flux distribution in bacteria. FEMS Microbiol Rev 29:765–794
72. Carere CR, Rydzak T, Verbeke TJ, Cicek N, Levin DB, Sparling R (2012) Linking genome content to biofuel production yields: a meta-analysis of major catabolic pathways among select H_2 and ethanol-producing bacteria. BMC Microbiol 12:295
73. Li Y, Tschaplinski TJ, Engle NL, Hamilton CY, Rodriguez M Jr, Liao JC, Schadt CW, Guss AM, Yang Y, Graham DE (2012) Combined inactivation of the *Clostridium cellulolyticum*

lactate and malate dehydrogenase genes substantially increases ethanol yield from cellulose and switchgrass fermentations. Biotechnol Biofuels 5(1):2
74. Levin DB, Pitt L, Love M (2004) Biohydrogen production prospects and limitations to practical application. Int J Hydrogen Energy 29:173–185
75. Levin DB, Islam R, Cicek N, Sparling R (2006) Hydrogen production by *Clostridium thermocellum* 27405 from cellulosic biomass substrates. Int J Hydrogen Energy 31:1496–1503
76. Kalia VC, Lal S, Ghai R, Mandal M, Chauhan A (2003) Mining genomic databases to identify novel hydrogen producers. Trends Biotechnol 21:152–156
77. Kim DH, Kim MS (2001) Hydrogenases for biological hydrogen production. Bioresour Technol 102:8423–8431
78. Masukawa H, Mochimaru M, Sakurai H (2002) Disruption of the uptake hydrogenase gene, but not of the bidirectional hydrogenase gene, leads to enhanced photobiological hydrogen production by the nitrogen-fixing cyanobacterium *Anabaena* species PCC 7120. Appl Microbiol Biotechnol 58:618–624
79. Hallenbeck PC, Benemann JR (2002) Biological hydrogen production; fundamentals and limiting processes. Int J Hydrogen Energy 27:1185–1193
80. Calusinska M, Happe T, Joris B, Wilmotte A (2010) The surprising diversity of clostridial hydrogenases: a comparative genomic perspective. Microbiology 156:1575–1588
81. Vignais PM, Magnin JP, Willison JC (2006) Increasing biohydrogen production by metabolic engineering. Int J Hydrogen Energy 31:1478–1483
82. Pilak O, Mamat B, Vogt S, Hagemeier CH, Thauer RK, Shima S, Vonrhein C, Warkentin E, Ermler U (2006) The crystal structure of the apoenzyme of the iron-sulfur cluster-free hydrogenase. J Mol Biol 358:798–809
83. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FSL (2005) PSORTb v. 2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21:617–623
84. Gu Y, Ding Y, Ren C, Sun Z, Rodionov DA, Zhang W, Yang S, Yang C, Jiang W (2010) Reconstruction of xylose utilization pathway and regulons in *Firmicutes*. BMC Genomics 11:255
85. Liaw SH, Kuo I, Eisenberg D (1995) Discovery of the ammonium substrate site on glutamine synthetase, a third cation binding site. Protein Sci 4:2358–2365
86. Liaw SH, Pan C, Eisenberg D (1993) Feedback inhibition of fully unadenylylated glutamine synthetase from *Salmonella typhimurium* by glycine, alanine, and serine. Proc Natl Acad Sci U S A 90:4996–5000
87. Huang Y, Zhang L (2004) Rapid and sensitive dot matrix methods for genome analysis. Bioinformatics 20:460–466
88. Munir RI, Spicer V, Shamshurin D, Krokhin OV, Wilkins J, Ramachandran U, Sparling R, Levin DB (2015) Quantitative proteomic analysis of the cellulolytic system of *Clostridium termitidis* CT1112 reveals distinct protein expression profiles upon growth on α-cellulose and cellobiose. J Proteomics 125:41–53