From proteins to polysaccharides: lifestyle and genetic evolution of Coprothermobacter proteolyticus

Benoit J Kunath, Francesco Delogu, Adrian E Naas, Magnus Ø. Arntzen, Vincent Eijsink, Bernard Henrissat, Torgeir Hvidsten, Philippe B. Pope

▶ **To cite this version:**

Benoit J Kunath, Francesco Delogu, Adrian E Naas, Magnus Ø. Arntzen, Vincent Eijsink, et al.. From proteins to polysaccharides: lifestyle and genetic evolution of Coprothermobacter proteolyticus. The International Society of Microbiologial Ecology Journal, 2019, 13 (3), pp.603-617. 10.1038/s41396-018-0290-y . hal-02611222

HAL Id: hal-02611222
https://hal.science/hal-02611222

Submitted on 18 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

The ISME Journal (2018) 13:603–617
https://doi.org/10.1038/s41396-018-0290-y

ARTICLE

From proteins to polysaccharides: lifestyle and genetic evolution of *Coprothermobacter proteolyticus*

Benoit J. Kunath[1] · Francesco Delogu[1] · Adrian E. Naas[1] · Magnus Ø. Arntzen[1] · Vincent G. H. Eijsink[1] · Bernard Henrissat[2] · Torgeir R. Hvidsten[1] · Phillip B. Pope[1]

Received: 13 March 2018 / Revised: 11 July 2018 / Accepted: 19 September 2018 / Published online: 12 October 2018
© The Author(s) 2018. This article is published with open access

Abstract

Microbial communities that degrade lignocellulosic biomass are typified by high levels of species- and strain-level complexity, as well as synergistic interactions between both cellulolytic and non-cellulolytic microorganisms. *Coprothermobacter proteolyticus* frequently dominates thermophilic, lignocellulose-degrading communities with wide geographical distribution, which is in contrast to reports that it ferments proteinaceous substrates and is incapable of polysaccharide hydrolysis. Here we deconvolute a highly efficient cellulose-degrading consortium (SEM1b) that is co-dominated by *Clostridium (Ruminiclostridium) thermocellum* and multiple heterogenic strains affiliated to *C. proteolyticus*. Metagenomic analysis of SEM1b recovered metagenome-assembled genomes (MAGs) for each constituent population, whereas in parallel two novel strains of *C. proteolyticus* were successfully isolated and sequenced. Annotation of all *C. proteolyticus* genotypes (two strains and one MAG) revealed their genetic acquisition of carbohydrate-active enzymes (CAZymes), presumably derived from horizontal gene transfer (HGT) events involving polysaccharide-degrading Firmicutes or Thermotogae-affiliated populations that are historically co-located. HGT material included a saccharolytic operon, from which a CAZyme was biochemically characterized and demonstrated hydrolysis of multiple hemicellulose polysaccharides. Finally, temporal genome-resolved metatranscriptomic analysis of SEM1b revealed expression of *C. proteolyticus* CAZymes at different SEM1b life stages as well as co-expression of CAZymes from multiple SEM1b populations, inferring deeper microbial interactions that are dedicated toward community degradation of cellulose and hemicellulose. We show that *C. proteolyticus*, a ubiquitous population, consists of closely related strains that have adapted via HGT to presumably degrade both oligo- and longer polysaccharides present in decaying plants and microbial cell walls, thus explaining its dominance in thermophilic anaerobic digesters on a global scale.

Introduction

The anaerobic digestion of plant biomass profoundly shapes innumerable ecosystems, ranging from the gastrointestinal tracts of humans and other mammals to those that drive industrial applications such as biofuel generation. Biogas reactors are one of the most commonly studied anaerobic systems, yet many keystone microbial populations and their metabolic processes are poorly understood due to a lack of cultured or genome sampled representatives. *Coprothermobacter* spp. are frequently observed in high abundance in thermophilic anaerobic systems, where they are believed to exert strong protease activity, while generating hydrogen and acetate, key intermediate metabolites for biogas production [1]. Molecular techniques have shown that their levels range from 10% to 90% of the total microbial community, irrespective of bioreactors being operated on lignocellulose- or protein-rich substrates (Fig. 1). Despite their promiscuous distribution, global abundance and key role in biogas production, only two species have been described: *Coprothermobacter platensis* [2] and *Coprothermobacter*

These authors contributed equally: B. J. Kunath, F. Delogu.

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41396-018-0290-y) contains Supplementary Material, which is available to authorized users.

✉ Phillip B. Pope
phil.pope@nmbu.no

[1] Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås 1432, Norway

[2] Architecture et Fonction des Macromolécules Biologiques, CNRS, Aix-Marseille Université, Marseille F-13288, France

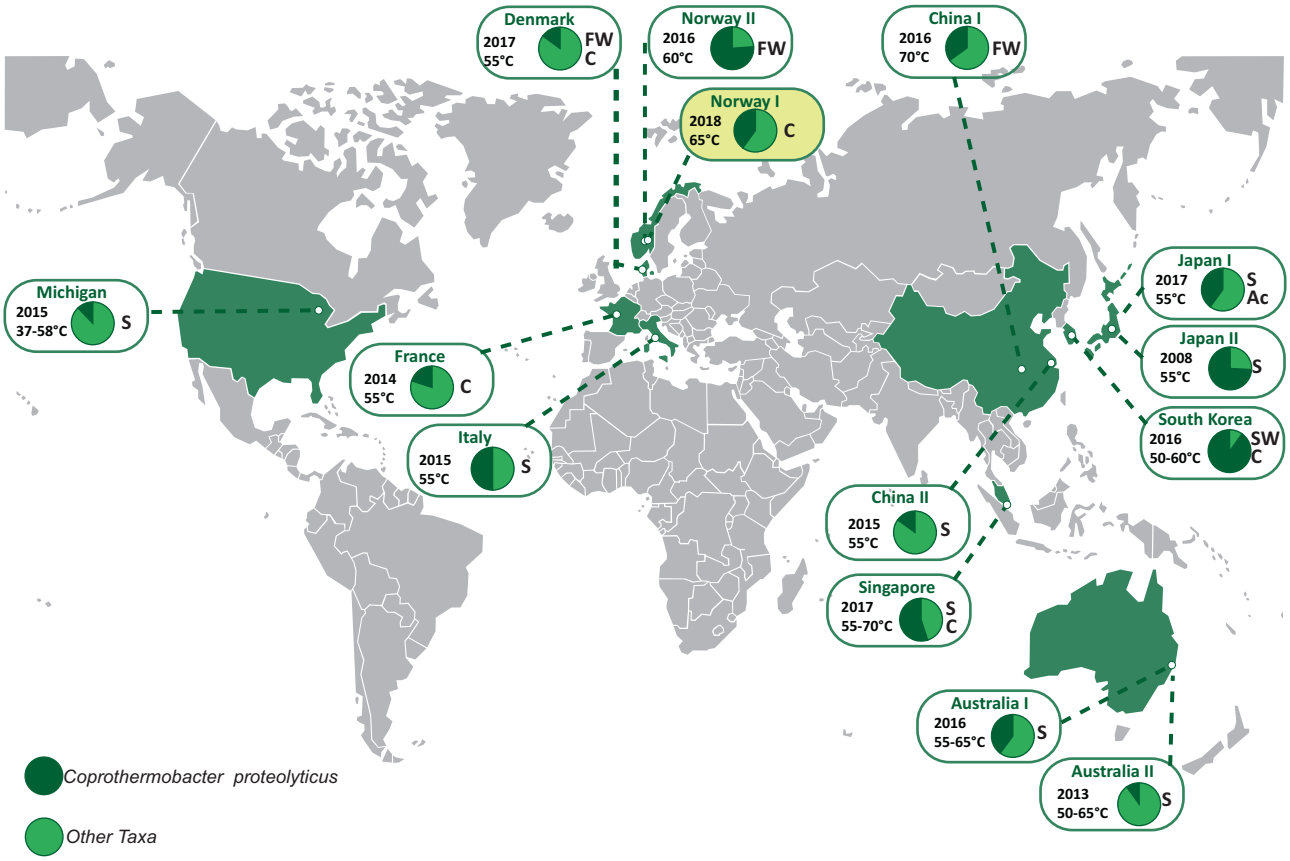

Fig. 1 Global distribution of *C. proteolyticus*-affiliated populations in anaerobic biogas reactors. Charts indicate relative 16S rRNA gene abundance of OTUs affiliated to *C. proteolyticus* (dark green), in comparison with the total community (light green). The year of publication, reactor temperature, and substrate (C cellulose, FW food waste, S sludge, SW Seaweed, Ac acetate) is indicated (details in Table S1). The SEM1b consortium analyzed in this study is highlighted in yellow

proteolyticus [3]. These two species and their inherent phenotypes have formed the predictive basis for the majority of *Coprothermobacter*-dominated systems described to date. Recent studies have illustrated that *C. proteolyticus* populations in anaerobic biogas reactors form cosmopolitan assemblages of closely related strains that are hitherto unresolved [4].

Frequently in nature, microbial populations are composed of multiple strains with genetic heterogeneity [5, 6]. Studies of strain-level populations have been predominately performed with the human microbiome and especially the gut microbiota [7, 8]. The reasons for strain diversification and their coexistence remain largely unknown [9]; however, several mechanisms have been hypothesized, such as micro-niche selection [5, 10], host selection [11], cross-feed interactions [12, 13], and phage selection [14]. Studies of axenic strains have shown that isolates can differ in a multitude of ways, including virulence and drug resistance [15–17], motility [18], and nutrient utilization [19]. Strain-level genomic variations typically consist of single-nucleotide variants, as well as acquisition/loss of genomic elements such as genes, operons, or plasmids via horizontal gene transfer (HGT) [20–22]. Variability in gene content caused by HGT is typically attributed to phage-related genes and other genes of unknown function [23], and can give rise to ecological adaptation, niche differentiation, and eventually, speciation [24–26]. Although differences in genomic features can be accurately characterized in isolated strains, it has been difficult to capture such information using culture-independent approaches such as metagenomics. Advances in bioinformatics have improved taxonomic profiling of microbial communities from phylum to species level but it remains difficult to profile similar strains from metagenomes and compare them with the same level of resolution obtained by comparison of isolate genomes [27]. As closely related strains can also differ in gene expression [28], being able to distinguish the expression profiles of individual strains in a broader ecological context is elemental to understanding the influence they exert towards the overall community function.

In this study, a novel population of *C. proteolyticus* that included multiple closely related strains was observed within a simplistic biogas-producing consortium enriched on cellulose (hereafter referred to as SEM1b). Using a

combined metagenomic and culture-dependent approach, two strains and a metagenome-assembled genome (MAG) affiliated to *C. proteolyticus* were recovered and genetically compared with the only available type strain, *C. proteolyticus* DSM 5265 [29]. Notable genomic differences included the acquisition of an operon (region-A) encoding carbohydrate-active enzymes (CAZymes), which inferred that *C. proteolyticus* has adapted to take advantage of longer polysaccharides. Enzymology was used to further support our hypothesis that the CAZymes within region-A are functionally active. We further examined the saccharolytic potential of our recovered *C. proteolyticus* population in a broader community context, by examining genome-resolved temporal metatranscriptomic data generated from the SEM1b consortium. Collective analysis highlighted the time-specific polysaccharide-degrading activity that *C. proteolyticus* exerts in a cellulolytic microbial community.

Materials and methods

Generation of the SEM1b consortium

An inoculum (100 µl) was collected from a lab-scale biogas reactor (Reactor TD) fed with manure and food waste and run at 55 °C. The TD reactor originated itself from a thermophilic (60 °C) biogas plant (Frevar) fed with food waste and manure in Fredrikstad, Norway. Our research groups have previously studied the microbial communities in both the Frevar plant [4] and the TD bioreactor [30], which provided a detailed understanding of the original microbial community. The inoculum was transferred for serial dilution and enrichment to an anaerobic serum bottle and containing the rich ATCC medium 1943, with cellobiose substituted for 10 g/L of cellulose in the form of Borregaard Advanced Lignin technology (BALI^{TM})-treated Norway spruce [31]. Our enrichment was incubated at 65 °C with the lesser objective to study community biomass conversion at the upper temperature limits of methanogenesis. After an initial growth cycle, an aliquot was removed and used for a serial dilution to extinction experiment. Briefly, a 100 µl sample was transferred to a new 100 ml bottle containing 60 ml of anaerobic medium, mixed, and 100 µl was directly transferred again to a new one (six serial transfers in total). The consortium at maximum dilution that retained the cellulose-degrading capability (SEM1b) was retained for the present work and aliquots were stored at − 80 °C with glycerol (15% v/v). In parallel, continuous SEM1b cultures were maintained via regular transfers into fresh media (each recultivation incubated for ~2–3 days).

Metagenomic analysis

Two different samples (D1B and D2B) were taken from a continuous SEM1b culture and were used for shotgun metagenomic analysis. D2B was 15 recultivations older than D1B and was used to leverage improvements in metagenome assembly and binning. From 6 ml of culture, cells were pelleted by centrifugation at 14,000 × *g* for 5 min and were kept frozen at − 20 °C until processing. Non-invasive DNA extraction methods were used to extract high molecular weight DNA as previously described [32]. The DNA was quantified using a Qubit™ fluorimeter and the Quant-iT™ dsDNA BR Assay Kit (Invitrogen, USA), and the quality was assessed with a NanoDrop 2000 (Thermo Fisher Scientific, USA).

16S rRNA gene analysis was performed on both D1B and D2B samples. The V3–V4 hyper-variable regions of bacterial and archaeal 16S rRNA genes were amplified using the 341F/805R primer set: 5′-CCTACGGGNBGC ASCAG-3′/5′-GACTACNVGGGTATCTAATCC-3′ [33]. The PCR was performed as previously described [30] and the sequencing library was prepared using Nextera XT Index kit according to Illumina's instructions for the MiSeq system (Illumina, Inc.). MiSeq sequencing (2 × 300 bp with paired ends) was conducted using the MiSeq Reagent Kit v3. The reads were quality filtered (Phred ≥ Q20) and USEARCH61 [34] was used for detection and removal of chimeric sequences. Resulting sequences were clustered at 97% similarity into operational taxonomic units (OTUs) and taxonomically annotated with the pick_closed_reference_otus.py script from the QIIME v1.8.0 toolkit [35] using the Greengenes database (gg_13_8). The resulting OTU table was corrected based on the predicted number of *rrs* operons for each taxon [36].

D1B and D2B were also subjected to metagenomic shotgun sequencing using the Illumina HiSeq 3000 platform (Illumina, Inc.) at the Norwegian Sequencing Center (NSC, Oslo, Norway). Samples were prepared with the TrueSeq DNA PCR-free preparation, and sequenced with paired ends (2 × 125 bp) on four lanes (two lanes per sample). Quality trimming of the raw reads was performed using cutadapt [37], removing all bases on the 3′-end with a Phred score lower than 20 (if any present) and excluding all reads shorter than 100 nt, followed by a quality filtering using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Reads with a minimum Phred score of 30 over 90% of the read length were retained. In addition, genomes from two isolated *C. proteolyticus* strains (see below) were used to decrease the data complexity and to improve the metagenomic assembly and binning. The quality-filtered metagenomic reads were mapped against the assembled strains using the Burrows-Wheeler Aligner with maximal exact

matches (BWA-MEM) algorithm requiring 100% identity [38]. Reads that mapped the strains were removed from the metagenomic data and the remaining reads were co-assembled using MetaSpades v3.10.0 [39] with default parameters and k-mer sizes of 21, 33, 55, and 77. The subsequent contigs were binned with Metabat v0.26.3 [40] in "very sensitive mode", using the coverage information from D1B and D2B. The quality (completeness, contamination, and strain heterogeneity) of the bins (hereafter referred to as MAGs) was assessed by CheckM v1.0.7 [41] with default parameters.

Isolation of *C. proteolyticus* strains

Strains were isolated using the Hungate method [42]. In brief Hungate tubes were anaerobically prepared with the DSMZ medium 481 with and without agar (15 g/L). Directly after being autoclaved, Hungate tubes containing agar were cooled down to 65 °C and sodium sulfide nonahydrate was added. From the SEM1b culture used for D1B, 100 µl were transferred to a new tube and mixed. From this new tube, 100 µl was directly transferred to 10 ml of fresh medium, mixed, and transferred again (six transfers in total). Tubes were then cooled to 60 °C for the agar to solidify and then kept at the same temperature. After growth, single colonies were picked and transferred to liquid medium.

DNA was extracted using the aforementioned method for metagenomic DNA, with one amendment: extracted DNA was subsequently purified with DNeasy PowerClean Pro Cleanup Kit (Qiagen, USA) following manufacturer's instructions. To insure the purity of the *C. proteolyticus* colonies, visual confirmation was performed using light microscopy and long 16S rRNA genes were amplified using the primers pair 27F/1492R [43]: 5′-AGAGTTTG ATCMTGGCTCAG-3′/5′-TACGGYTACCTTGTTACGA CTT-3′ and sequenced using Sanger technology. The PCR consisted of an initial denaturation step at 94 °C for 5 min and 30 cycles of denaturation at 94 °C for 1 min, annealing at 55 °C for 1 min, and extension at 72 °C for 1 min, and a final elongation at 72 °C for 10 min. PCR products were purified using the NucleoSpin Gel and PCR Cleanup kit (Macherey-Nagel, Germany) and sent to GATC Biotech for Sanger sequencing.

The genomes of two isolated *C. proteolyticus* strains (hereafter referred to as *BWF2A* and *SW3C*) were sequenced at the NSC (Oslo, Norway). Samples were prepared with the TrueSeq DNA PCR-free preparation and sequenced using paired ends (2 × 300 bp) on a MiSeq system (Illumina, Inc). Quality trimming, filtering, and assembly were performed as described in the aforementioned metagenomic assembly section. The raw reads were additionally mapped on assembled contigs using bowtie2 (–very-sensitive -X 1000 -I 350) and the coverage was retrieved for every nucleotide with samtool depth –a. All the contigs with an average coverage higher than 100 were selected and individually inspected for coverage discontinuity. All the contigs selected with the average coverage criterion (BWF2A: 11, SW3C: 13) looked continuous in coverage and, together with the MAGs, they were submitted to the Integrated Microbial Genomes and Microbiomes system [44] for genomic feature prediction and annotation (pipeline version 4.15.1). Resulting annotated open reading frames (ORFs) were retrieved, further annotated for CAZymes using the CAZy annotation pipeline [45], and subsequently used as a reference database for the metatranscriptomics (with exception of glycosyltransferases). The genomes from both strains and MAGs corresponding to *C. proteolyticus* were compared with the reference genome from *C. proteolyticus* DSM 5265. Using the BRIG tool [46] for mapping and visualization, the different genomes were mapped against their pan genome generated using Roary [47].

Phylogenetic analysis

A concatenated ribosomal protein phylogeny was performed on the MAGs and the isolated strains using 16 ribosomal proteins chosen as single-copy phylogenetic marker genes (RpL2, 3, 4, 5, 6, 14, 15, 16, 18, 22, and 24, and RpS3, 8, 10, 17, and 19) [48]. The dataset was augmented with metagenomic sequences retrieved from our previous research on the original FREVAR reactor [4] and with sequences from reference genomes identified during the 16S rRNA analysis. Each gene set was individually aligned using MUSCLE v3.8.31 [49] and then manually curated to remove end gaps and ambiguously aligned terminal regions. The curated alignments were concatenated and a maximum likelihood phylogeny was obtained using MEGA7 [50] with 1000 bootstrap replicates. The radial tree was visualized using iTOL [51]. In addition, an average nucleotide identity (ANI) comparison was performed between each MAG and their closest relative using the ANI calculator [52].

Heterologous expression and purification of the GH16 enzyme

The *C. proteolyticus* BWF2A Ga0187557_1002 gene-sequence without predicted signal peptide [53] was cloned from isolated genomic DNA using the following primers; GH16_Fwd: 5′-TTAAGAAGGAGATATACTATGCTCG GCGTGAATGTGATG-AATATAAGTGA-3′; GH16_rev: 5′-AATGGTGGTGATGATGGTGCGCCTCATTTTCAA GCTTGTATA-CACGGACATAATC-3′, and cloned into the pNIC-CH plasmid in *Escherichia coli* TOP10 by ligation-independent cloning [54]. The transformant's sequence was verified by sequencing before transformation

into OneShot® *E. coli* BL21 Star™ cells (Thermo Fischer Scientific, Waltham, MA, USA) for expression, where 200 ml Luria-broth containing 50 µg/ml kanamycin was inoculated with 2 ml overnight culture and incubated at 37 °C, 200 r.p.m. Expression was induced when the culture reached an OD600 of 0.6, by addition of isopropyl-β-D-1-thiogalactopyranoside. The culture was incubated at 22 °C, 200 r.p.m. for 16 h, before collection by centrifugation (5000 × *g*, 10 min) and storage of the pellet at − 80 °C. The frozen pellet was transferred to 20 mL buffer A (20 mM Tris-HCL pH 8.0, 200 mM NaCl, 5 mM imidazole) containing 1 × BugBuster (Merck Millipore, Berlington, MA, USA) and stirred for 20 min at room temperature to lyse the cells. Cell debris was removed by centrifugation (30,000 × *g*, 20 min) and the protein was purified by immobilized metal-ion chromatography using a 5 ml HisTrap FF column (GE-Healthcare, Little Chalfont, UK) pre-equilibriated with buffer A. The protein was eluted using a linear gradient to Buffer B (Buffer A with 500 mM imidazole). The purity of the eluted fractions were assessed by SDS-polyacrylamide gel electrophoresis and the imidazole was removed from the buffer by repeated concentration and dilution using a Vivaspin (Sartorius, Göttingen, Germany) concentrator with a 10 kDa cutoff. The protein concentration was determined by measured A280 and the calculated extinction coefficient.

Biochemical characterization of the GH16 enzyme

Assays were performed in triplicate in 96-well plates and contained 1 mg/ml substrate, 20 mM BisTris, pH 5.8 (50 °C), and 1 µM enzyme in a volume of 100 µl. The reactions were pre-heated to 50 °C before addition of enzyme and were sealed before incubation for 1 h in a Thermomixer C incubator with heated lid (Eppendorf, Hamburg, Germany). The substrates used were as follows: barley β-glucan, carboxymethyl-curdlan, carboxymethyl-pachyman, carob galactomannan, tamarind xyloglucan, wheat arabinoxylan, larch arabinogalactan (all from Megazyme, Bray, Co. Wicklow, Ireland), and laminarin from Laminaria digitate (Sigma-Aldrich, St. Louis, MO, USA). Reactions were stopped by addition of DNS reagent (100 µl, 10 g/l 3,5-dinitrosalicylic acid, 300 g/L potassium sodium tartrate, 10 g/L NaOH [55] for quantification, or NaOH to a final concentration of 0.1 M for product analysis. Reducing ends were quantified against a standard curve of glucose, where reactions with DNS reagent were incubated at 95 °C for 20 min before cooling on ice and the absorbance was measured at 540 nm. For product analysis, the reactions containing NaOH were further diluted 1:10 in water, before analysis by high-performance anion-exchange chromatography with pulsed amperometric detection (HPAEC-PAD), using a Dionex ICS3000 system with a CarboPac PA1 column (Sunnyvale, CA, USA). Oligosaccharides were eluted using a multi-step gradient, going from 0.1 M NaOH to 0.1 M NaOH–0.3 M sodium acetate (NaOAc) over 35 min, to 0.1 M NaOH–1.0 M NaOAc over 5 min, before going back to 0.1 M NaOH over 1 min, and reconditioning for 9 min at 0.1 M NaOH.

Temporal meta-omic analyses of SEM1b

A "meta-omic" time series analysis was conducted over the lifetime span of the SEM1b consortium (≈45 h). A collection of 27 replicate bottles containing ATCC medium 1943 with 10 g/L of cellulose (60 ml total volume) were inoculated from the same SEM1b culture and incubated at 65 °C in parallel. For each sample time point, three culture-containing bottles were removed from the collection and processed in triplicate. Sampling occurred over nine time points (at 0, 8, 13, 18, 23, 28, 33, 38, and 43 h) during the SEM1b life cycle and are hereafter referred as T0, T1, T2, T3, T4, T5, T6, T7, and T8, respectively. DNA for 16S rRNA gene analysis was extracted (as above) from T1 to T8 and kept at − 20 °C until amplification and sequencing, and the analysis was performed using the protocol described above. Due to low cell biomass at the initial growth stages, sampling for metatranscriptomics was performed from T2 to T8. Sample aliquots (6 ml) were treated with RNAprotect Bacteria Reagent (Qiagen, USA) following the manufacturer's instructions and the treated cell pellets were kept at − 80 °C until RNA extraction.

In parallel, metadata measurements including cellulose degradation rate, monosaccharide production, and protein concentration were performed over all the nine time points (T0–T8). For monosaccharide detection, 2 ml samples were taken in triplicates, centrifuged at 16,000 × *g* for 5 min and the supernatants were filtered with 0.2 µm sterile filters and boiled for 15 min before being stored at − 20 °C until processing. Solubilized sugars released during microbial hydrolysis were identified and quantified by HPAEC with PAD. A Dionex ICS3000 system (Dionex, Sunnyvale, CA, USA) equipped with a CarboPac PA1 column (2 × 250 mm; Dionex, Sunnyvale, CA, USA) and connected to a guard of the same type (2 × 50 mm) was used. Separation of products was achieved using a flow rate of 0.25 mL/min in a 30 min isocratic run at 1 mM KOH at 30 °C. For quantification, peaks were compared with linear standard curves generated with known concentrations of selected monosaccharides (glucose, xylose, mannose, arabinose, and galactose) in the range of 0.001–0.1 g/L.

Total protein measurements were taken to estimate SEM1b growth rate. Proteins were extracted following a previously described method [4] with a few modifications. Briefly, 30 ml culture aliquots were centrifuged at 500 × *g* for 5 min to remove the substrate and the supernatant was centrifuged at 9000 × *g* for 15 min to pellet the cells. Cell lysis was performed by resuspending the cells in 1 ml of

lysis buffer (50 mM Tris-HCl, 0.1% (v/v) Triton X-100, 200 mM NaCl, 1 mM dithiothreitol, 2 mM EDTA) and keeping them on ice for 30 min. Cells were disrupted in 3 × 60 s cycles using a FastPrep24 (MP Biomedicals, USA) and the debris were removed by centrifugation at 16,000 × *g* for 15 min. Supernatants containing proteins were transferred into low bind protein tubes and the proteins were quantified using Bradford's method [56].

As estimation of cellulose degradation requires analyzing the total content of a sample to be accurate, the measurements were performed on individual cultures that were prepared separately. A collection of 18 bottles (9 time points in duplicate) were prepared using the same inoculum described above and grown in parallel with the 27-bottle collection used for the meta-omic analyses. For each time point, the entire sample was recovered, centrifuged at 5000 × *g* for 5 min, and the supernatant was discarded. The resulting pellets were boiled under acidic conditions as previously described [57] and the dried weights, corresponding to the remaining cellulose, were measured.

mRNA extraction was performed in triplicate on time points T2–T8, using previously described methods [58] with the following modifications in the processing of the RNA. The extraction of the mRNA included the addition of an in vitro-transcribed RNA as an internal standard to estimate the number of transcripts in the natural sample compared with the number of transcripts sequenced. The standard was produced by the linearization of a pGem-3Z plasmid (Promega, USA) with ScaI (Roche, Germany). The linear plasmid was purified with a phenol/chloroform/isoamyl alcohol extraction and digestion of the plasmid was assessed by agarose gel electrophoresis. The DNA fragment was transcribed into a 994 nt-long RNA fragment with the Riboprobe in vitro Transcription System (Promega, USA) following the manufacturer's protocol. Residual DNA was removed using the Turbo DNA Free kit (Applied Biosystems, USA). The quantity and the size of the RNA standard was measured with a 2100 bioanalyzer instrument (Agilent).

Total RNA was extracted using enzymatic lysis and mechanical disruption of the cells and purified with the RNeasy mini kit following the manufacturer's protocol (Protocol 2, Qiagen, USA). The RNA standard (25 ng) was added at the beginning of the extraction in every sample. After purification, residual DNA was removed using the Turbo DNA Free kit, and free nucleotides and small RNAs such as tRNAs were cleaned off with a lithium chloride precipitation solution according to Thermo Fisher Scientific's recommendations. To reduce the amount of rRNAs, samples were treated to enrich for mRNAs using the MICROBExpress kit (Applied Biosystems, USA). Successful rRNA depletion was confirmed by analyzing both pre- and post-treated samples on a 2100 bioanalyzer instrument. Enriched mRNA was amplified with the MessageAmp II-Bacteria Kit (Applied Biosystems, USA) following manufacturer's instruction and sent for sequencing at the NSC (Oslo, Norway). Samples were subjected to the TruSeq stranded RNA sample preparation, which included the production of a cDNA library, and sequenced with paired-end technology (2 × 125 bp) on one lane of a HiSeq 3000 system.

RNA reads were assessed for overrepresented features (adapters/primers) using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/) and ends with detected features and/or a Phred score lower than 20 were trimmed using Trimmomatic v.0.36 [59]. Subsequently, a quality filtering was applied with an average Phred threshold of 30 over a 10 nt window and a minimum read length of 100 nt. rRNA and tRNA were removed using SortMeRNA v.2.1b [60]. SortMeRNA was also used to isolate the reads originating from the pGem-3Z plasmid. These reads were mapped against the specific portion of the plasmid containing the Ampr gene using Bowtie2 [61] with default parameters and the number of reads per transcript was quantified and scaled to match the length of the standard (x5.08). The remaining reads were pseudoaligned against the metagenomic dataset, augmented with the annotated strains, using Kallisto pseudo –pseudobam [62]. The resulting output was used to generate mapping files with bam2hits, which were used for expression quantification with mmseq [63], and the results were scaled to match the initial volume of the samples (x 10). Of the 40,046 ORFs identified from the assembled SEM1b metagenome and 2 *C. proteolyticus* strains, 17,598 (44%) were not found to be expressed, whereas 21,480 (54%) were expressed and could be reliably quantified due to unique hits (reads mapping unambiguously against one unique ORF) (Figure S1A). The remaining 968 ORFs (2%) were expressed but identified only with shared hits (reads mapping ambiguously against more than one ORF, resulting in an unreliable quantification of the expression of each ORF) (Figure S1B). As having unique hits improves the expression estimation accuracy, the ORFs were grouped using mmcollapse, in order to improve the precision of expression estimates, with only a small reduction in biological resolution [64]. The process first collapses ORFs into homologous groups if they have 100% sequence identity and then further collapses ORFs (or expression groups) if they acquire unique hits as a group (Figure S1C). This process generated 39,146 expression groups of which 38,428 (98%) were singletons (groups composed of single ORF) and 718 (2%) were groups containing more than one homologous ORF. From the initial 968 low-information ORFs, 661 (68%) became part of an expression group containing unique hits, 77 (8%) became part of ambiguous group (no unique hits), and 230 (24%) remained singletons (without unique hits). All expression groups without unique hits were then excluded from the subsequent analysis. A total of 21,480 singletons and 605

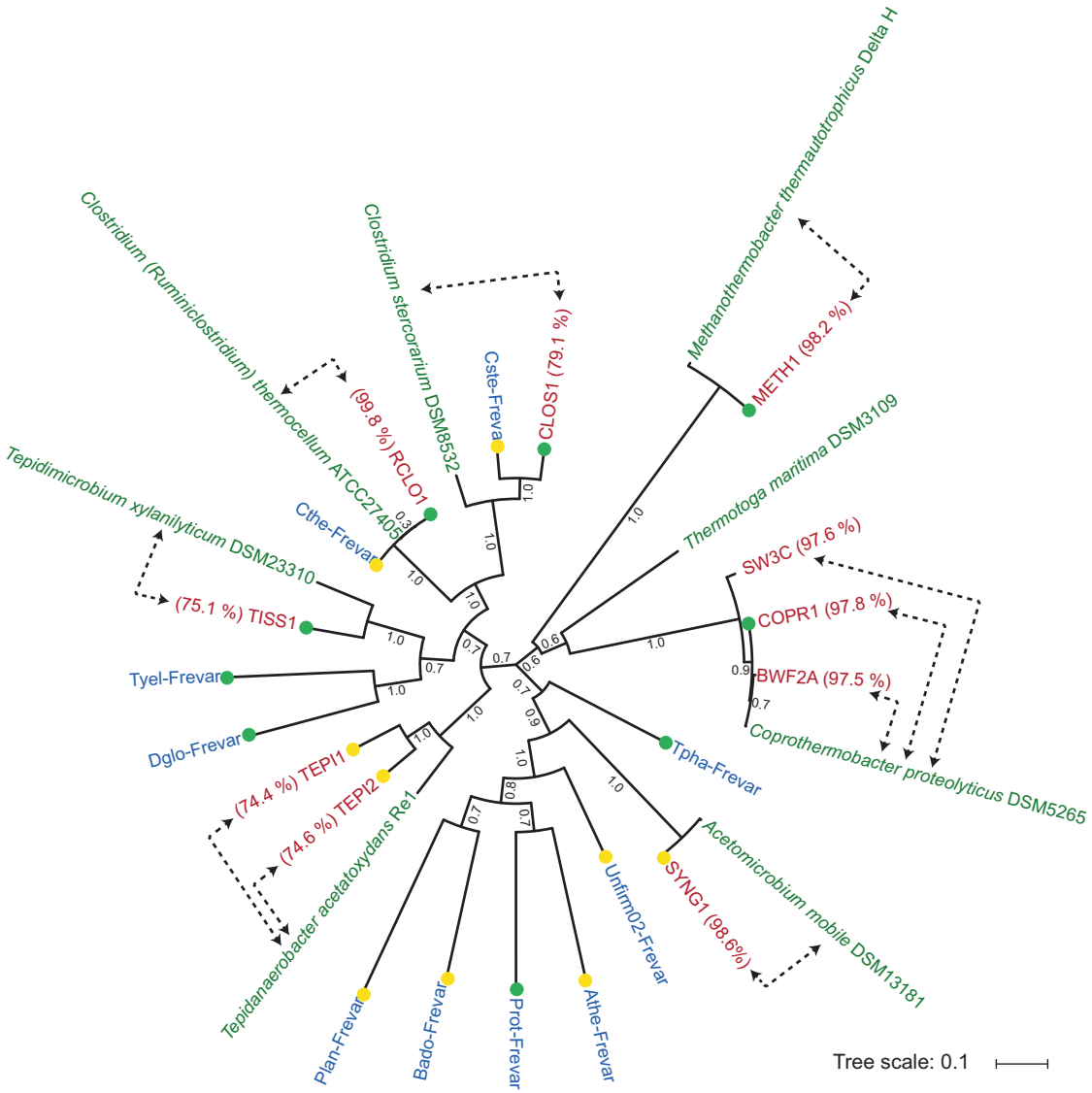

Fig. 2 Phylogeny of *C. proteolyticus* strains and other MAGs recovered from the SEM1b consortium. Concatenated ribosomal protein tree of reference isolate genomes (green), MAGs from the previous Frevar study (blue [4]), and MAGs and isolate genomes recovered in this study (red). Average nucleotide identities (percentage indicated in parenthesis) were generated between SEM1b MAGs and their closest relative (indicated by dotted arrows). Bootstrap values are based on 1000 bootstrap replicates and the completeness of the MAGs are indicated by green (> 90 %) and yellow (> 80 %) colored dots

multiple homologous expression groups were reliably quantified between *BWF2A*, *SW3C*, and the SEM1b metatranscriptome (Figure S1C).

In order to normalize the expression estimates, sample sizes were calculated using added internal standards, as described previously [58]. The number of reads generated from the internal standard molecule were calculated to be 2.4×10^4 +/− 2.1×10^4 reads per sample out of 6.2×10^9 molecules added. Using this information, the estimated number of transcript molecules per sample was computed to be 1.0×10^{13} +/− 7.3×10^{12} transcripts. The resulting estimates for the sample sizes were used to scale the expression estimates from mmseq collapse and to obtain absolute expression values. During initial screening the sample T7C (time point T7, replicate C) was identified as an outlier using principle component analysis and removed from downstream analysis.

The expression groups were clustered using hierarchical clustering with Euclidean distance. Clusters were identified using the Dynamic Tree Cut algorithm [65] with hybrid mode, deepsplit = 1, and minClusterSize = 7. Eigengenes were computed for the clusters and clusters with a Pearson's correlation coefficient > 0.9 were merged. The MAG/strain enrichment of the clusters was assessed using the BiasedUrn R package. The *p*-values were corrected with the Benjamini–Hochberg procedure and the significance threshold was set to 0.05. Expression groups composed of multiple MAGs/strains were included in several enrichment tests.

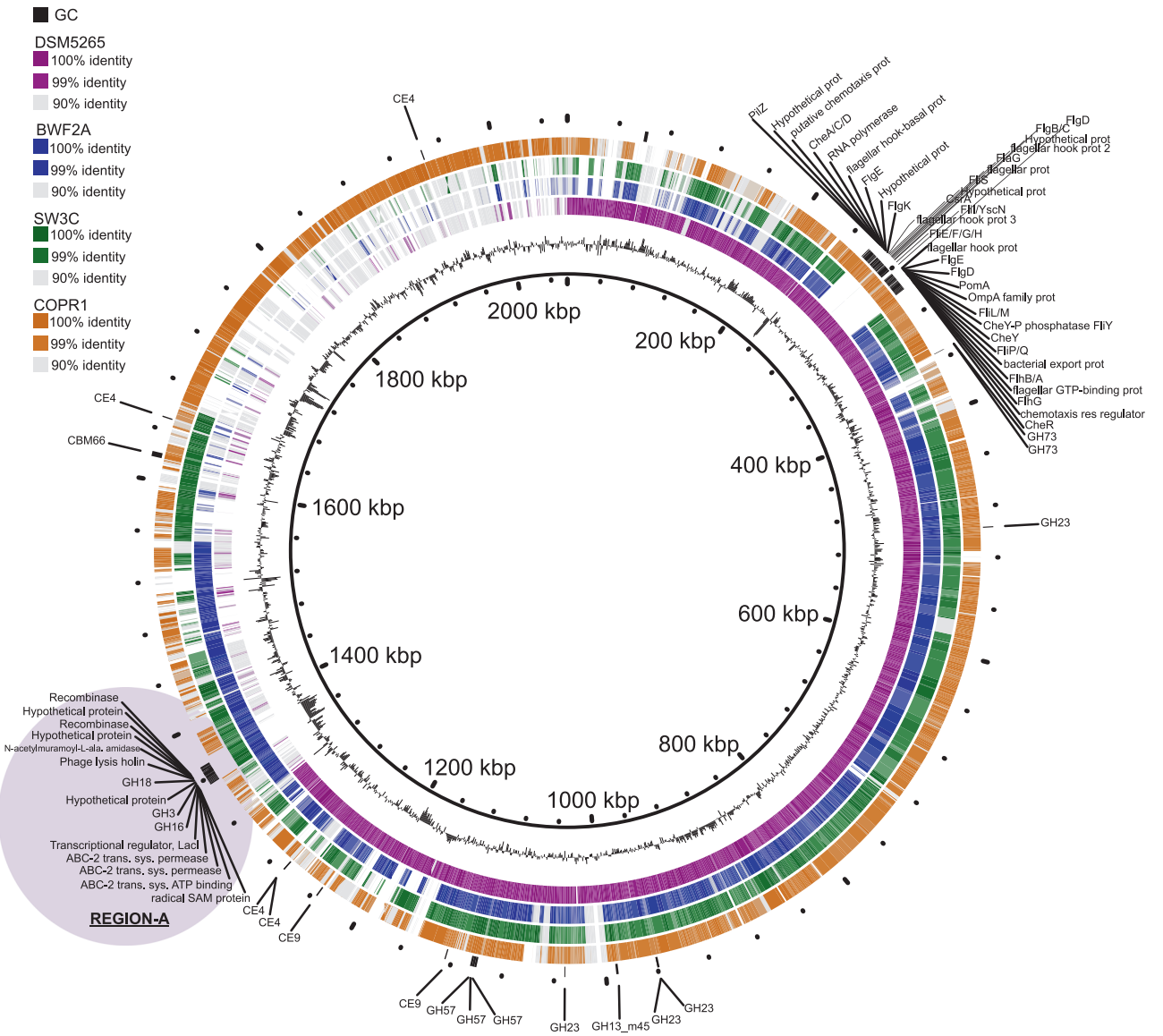

Fig. 3 Comparative genome content of *C. proteolyticus* representatives including isolated strains, a recovered MAG (COPR1), and the reference strain DSM 5265. The innermost ring corresponds to the pan genome of the three *C. proteolyticus* spp. genomes and one MAG as produced by Roary [47], and the second innermost ring represents the GC content. Outer rings represent the reference strain DSM 5265 (purple), the isolated strains *BWF2A* (blue) and *SW3C* (green), and the recovered COPR1 MAG (orange). Genes coding for carbohydrate-active enzymes (CAZymes) and flagellar proteins are indicted in black on the outermost ring. Genomic region-A is indicated by purple shading

Results and discussion

The SEM1b consortium is a simplistic community, co-dominated by *Clostridium (Ruminiclostridium) thermocellum* and heterogeneic *C. proteolyticus* strains

Molecular analysis of a reproducible, cellulose-degrading, and biogas-producing consortium (SEM1b) revealed a stable and simplistic population structure that contained approximately seven populations, several of which consisted of multiple strains (Fig. 2, Table S2–S3). 16S rRNA gene analysis showed that the SEM1b consortium was co-dominated by OTUs affiliated to the genera *Clostridium* (52%) and *Coprothermobacter* (41%), with closest representatives identified as *C. (Ruminiclostridium) thermocellum*, an uncharacterized *Clostridium spp.* and three *Coprothermobacter* phylotypes (Table S2). Previous meta-omic analysis on the parent Frevar reactor, revealed a multitude of numerically dominant *C. proteolyticus* strains, which created significant assembly and binning related issues [4]. In this study, multiple oligotypes of *C. proteolyticus* were also found (Table S2). We therefore sought to isolate and recover axenic representatives to complement our meta-omic approaches, and using traditional anaerobic isolation

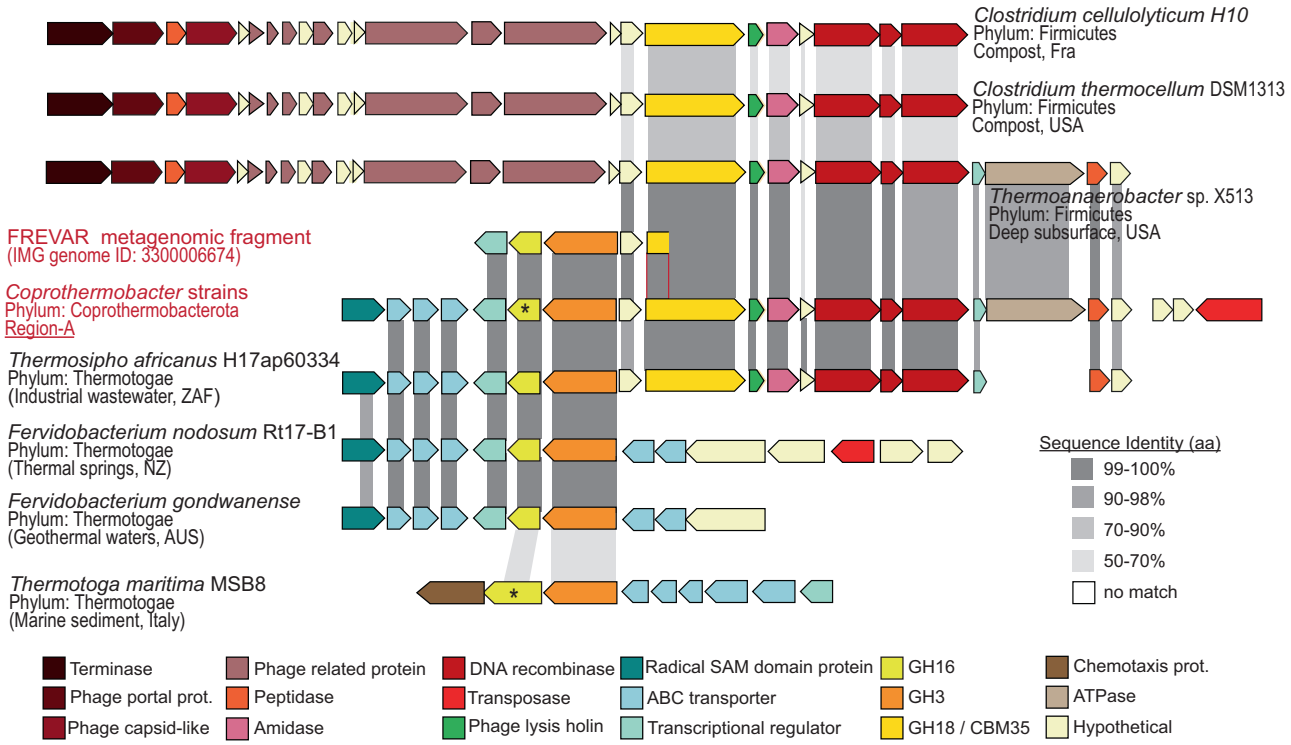

Fig. 4 Gene synteny of CAZymes within region-A encoded in *BWF2A* and *SW3C* genomes. The gene organization of CAZymes within region-A encoded in *BWF2A* and *SW3C* (see Fig. 3), as well as highly similar operons found in the original Frevar metagenome and isolated representatives from both phyla Firmicutes (*Thermoanaerobacter*, *C. cellulolyticum*, *C. thermocellum*) and Thermotogae (*T. africanus*, *F. nodosum*, *F. gondwanense*, and *Thermotoga maritima*). Grey shading between individual ORFs indicates amino acid sequence identity calculated between each query ORF (Frevar metagenome and isolates) and the reference ORF encoded in region-A from *BWF2A* and *SW3C* (identical in both strains). Asterisk denotes biochemically characterized GH16 enzymes, including the *C. proteolyticus* representative from this study and a laminarinase from *Thermotoa maritima* MSB8 that has previously been reported [79]

techniques, we were successful in recovering two novel axenic strains (hereafter referred to as *BWF2A* and *SW3C*). The genomes of *BWF2A* and *SW3C* were sequenced and assembled, and subsequently incorporated into our metagenomic and metatranscriptomic analysis below.

Shotgun metagenome sequencing of two SEM1b samples (D1B and D2B) generated 290 Gb (502 M paired-end reads) and 264 Gb (457 M paired-end reads) of data, respectively. Co-assembly of both datasets using strain-depleted reads with Metaspades produced 20,760 contigs totalizing 27 Mbp with a maximum contig length of 603 Kbp. Taxonomic binning revealed 11 MAGs and a community structure similar to the one observed by 16S analysis (Fig. 2, Table S3). A total of eight MAGs exhibited high completeness (> 80%) and a low level of contamination (< 10%). Three MAGs, COPR2, COPR3, and SYNG2, corresponded to small and incomplete MAGs, although Blastp analysis suggest COPR2 and COPR3 likely represent *Coprothermobacter*-affiliated strain elements.

All near-complete MAGs (> 80%), as well as *BWF2A* and *SW3C*, were phylogenetically compared against their closest relatives using ANIs and a phylogenomic tree was constructed via analysis of 16 concatenated ribosomal proteins (Fig. 2). One MAG was observed to cluster together with *C. proteolyticus* DSM 5265 and the two strains *BWF2A* and *SW3C*, and was defined as COPR1. Two MAGs (RCLO1-CLOS1) clustered together within the *Clostridium*; RCLO1 with the well-known *C. thermocellum*, whereas CLOS1 grouped together with another *Clostridium* MAG generated from the Frevar dataset and the isolate *C. stercorarium* (ANI: 79.1%). Both RCLO1 and CLOS1 encoded broad plant polysaccharide-degrading capabilities, containing 297 and 139 CAZymes, respectively (Table S4). RCLO1 in particular encoded cellulolytic (e.g., glycosyl hydrolase (GH) families GH5, GH9, and GH48) and cellulosomal features (dockerins and cohesins), whereas CLOS1 appears more specialized toward hemicellulose degradation (e.g., GH3, GH10, GH26, GH43, GH51, and GH130). Surprisingly, several CAZymes were also identified in COPR1 ($n = 65$), and both *BWF2A* ($n \times = 37$) and *SW3C* ($n = 34$) at levels higher than what has previously been observed in *C. proteolyticus* DSM 5265 ($n = 29$) (Table S4). Several MAGs were also affiliated with other known lineages associated with biogas processes, including *Tepidanaerobacter* (TEPI1-2), *Synergistales* (SYNG1-2), *Tissierellales* (TISS1), and *Methanothermobacter* (METH1).

Novel strains of *C. proteolyticus* reveal acquisition of CAZymes

Genome annotation of COPR1, *BWF2A*, and *SW3C* identified both insertions and deletions in comparison with the only available reference genome, sequenced from the type strain DSM 5265 (Fig. 3). Functional annotation showed that most of the genomic differences were sporadic and are predicted not to affect the metabolism of the strains. However, several notable differences were observed, which might represent a significant change in the lifestyle of the isolates. Both isolated strains lost the genes encoding flagellar proteins, although it is debatable that these genes originally conferred mobility in the type strain, as it has been previously reported as non-motile [3, 66]. Interestingly, both strains acquired extra CAZymes including a particular genomic region that encoded a cluster of three CAZymes: GH16, GH3, and GH18-CBM35 (region-A, Fig. 3). The putative function of these GHs suggests that both *BWF2A* and *SW3C* are capable of hydrolyzing various β-glucan linkages that are found in different hemicellulosic substrates (GH16: endo-β-1,3-1,4-glucanase; GH3: β-glucosidase). Regarding the putative GH18 encoded in both strains, it could have a role in bacterial cell wall recycling [67] as an endo-β-*N*-acetylglucosaminidase. Indeed, *C. proteolyticus* has previously been considered to be a scavenger of dead cells, even though this feature was mainly highlighted in term of proteolytic activities [68].

Taking a closer look, the region-A of CAZymes (GH16, GH3, and GH18-CBM35) in *BWF2A* and *SW3C* was located on the same chromosomal cassette but organized onto two different operons with opposite directions (Fig. 4). Comparison of the genes and their organization revealed a high percentage of gene similarity and synteny with genome representatives from both phyla Firmicutes (*Thermoanaerobacter*, *Clostridium cellulolyticum*, and *C. thermocellum*) and Thermotogae (*Thermosipho africanus*, *Fervidobacterium nodosum*, and *F. gondwanense*). Both *C. thermocellum* and *Fervidobacterium* populations were previously identified in the original Frevar reactor [4]. Moreover, a truncated contig from the Frevar metagenome (Scaffold Id:Ga0101770_1036339) exhibited 99.9 % nucleotide identity to the *BWF2A* and *SW3C* genomes spanning 4.7 Kb across the CAZymes and genomic sections from both phyla (Fig. 4), suggesting the acquirement of region-A preceded the SEM1b enrichment.

Examination of the flanking regions surrounding the CAZymes in region-A reveals the presence of an incomplete prophage composed of a phage lysis holin and two recombinases located downstream (Figs. 3, 4). Further comparisons revealed that only the Firmicutes lineages encoded the same prophage together with an additional terminase, phage-capsid-like proteins, and more phage-related components on the 5′-region (Fig. 4). Because of the high sequence homology and the presence of phage-genes in the surrounding, we hypothesized that the origin of region-A in *BWF2A* and *SW3C* is the result of phage-mediated HGT. Most likely, the operon from Firmicutes-affiliated lineages (e.g., *Thermoanaerobacter* and *C. thermocellum*) was transferred first due to the presence of its complete phage and generated a hotspot for further HGT for the GH16-GH3-encoding operon originating from Thermotogae-affiliated lineages (Fig. 4). Interestingly, *T. africanus* also encoded a syntenous region that covered Region-A in both *BWF2A* and *SW3C* almost in its entirety (Fig. 4), creating an alternative possibility that vertical gene transfer may also have had a role toward the evolution of this operon in *Coprothermobacter*. Gene transfer within anaerobic digesters has been reported for antibiotic resistance genes [69], whereas HGT of CAZymes have been detected previously among gut microbiota [70–72]. As many microbes express only a specific array of carbohydrate-degrading capabilities, bacteria that acquire CAZymes from gene transfer events may gain additional capacities and, consequently, a selective growth advantage [73].

In response to our discovery of *C. proteolyticus* CAZyme acquisition, we attempted to cultivate our axenic strains in minimal media containing only hemicellulosic substrates (pachyman, curdlan, barley β-glucan) as a sole carbon source. However, no growth was observed for either *BWF2A* or *SW3C* in polysaccharide-supplemented media that was without yeast extract. These results were consistent with the few available studies on type strain DSM 5265, which have shown weak and slow growth on proteins and monomeric sugars, and only in the presence of pluralistic organic compounds found in yeast extract and rumen fluid [3, 66]. Growth was observed in *BWF2A*/*SW3C* cultures with both yeast extract and polysaccharide substrates; however, we detected no increased levels of growth, indicating that in isolation our *C. proteolyticus* strains may require specific undefined cofactor(s) or collaborative microbial partners to support the activity encoded by their acquired CAZymes.

In lieu of axenic *C. proteolyticus* cultivation data to support a saccharolytic lifestyle, we biochemically interrogated the GH16 encoded in region-A (Fig. 4). The catalytic domain was synthesized and expressed in *E. coli*, followed by protein purification. As expected the GH16 demonstrated endoglucanase activity on β-1,3 (pachyman, curdlan, laminarin) and β-1,3-1,4 (Barley) substrates (Figure S2A), which supports our hypothesis that the CAZymes in region-A have transferred the ability of *BWF2A* or *SW3C* to degrade polysaccharides. Against all β-glucan substrates, GH16 hydrolysis generated a large fraction of glucose (Figure S2B), which has been shown to be readily fermented by *C. proteolyticus* [3, 66].

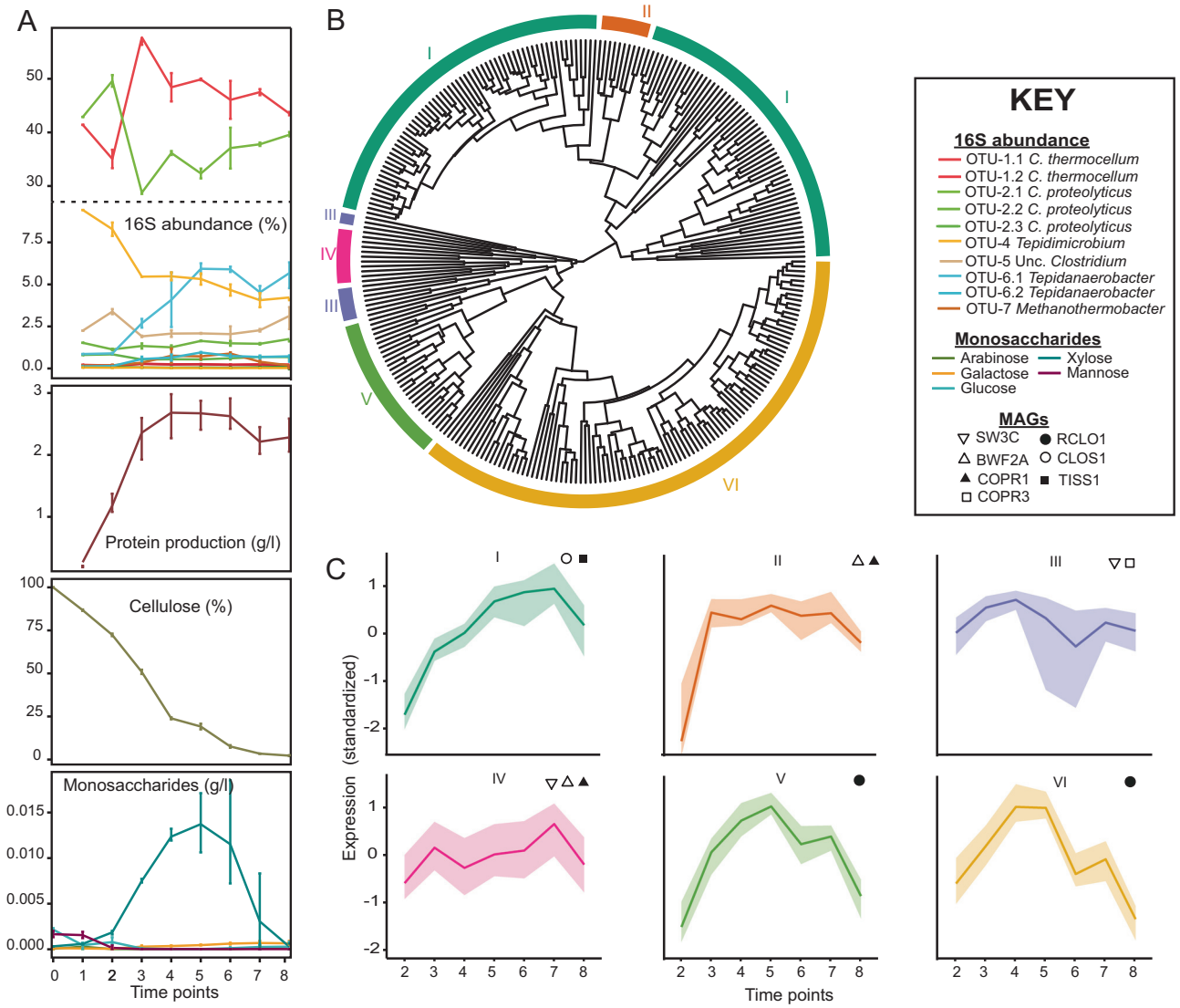

Fig. 5 Temporal meta-analysis of the SEM1b consortium. **a** 16S rRNA gene amplicon and metadata analysis was performed over a 43 h period, which was segmented into nine time points. OTU IDs are detailed in Table S2. Cellulose degradation rate, monosaccharide accumulation, and growth rate (estimated by total protein concentration) are presented. **b** Gene expression dendrogram and clustering of CAZymes from *BWF2A*, *SW3C*, and MAGs: RCLO1, CLOS1, COPR1-3, and TISS1. Six expression clusters (I–VI) are displayed in different colors on the outer ring. **c** Clusters I–VI show characteristic behaviors over time summarized by the median (solid line) and the shaded area between the first and third quartile of the standardized expression. Bacteria that are statistically enriched (p-value < 0.05) in the clusters are displayed in the subpanels

C. proteolyticus expresses CAZymes and is implicit in collaborative polysaccharide degradation within the SEM1b consortium

Although we confirmed that the acquired *C. proteolyticus* GH16 is functionally active, we also sought to better understand the role(s) had by it and other *C. proteolyticus* CAZymes in a saccharolytic consortium, by analyzing the temporal metatranscriptome of SEM1b over a complete life cycle. 16S rRNA gene analysis of eight time points (T1–8) over a 43 h period reaffirmed that *C. thermocellum*- and *C. proteolyticus*-affiliated populations dominate SEM1b over time (Fig. 5a). Highly similar genes from different MAGs/genomes were grouped together, in order to obtain "expression groups" with discernable expression profiles (see Methods and Figure S1A/B). A total of 274 singleton CAZyme expression groups and 8 multiple ORF groups were collectively detected in the two *C. proteolyticus* strains and MAGs suspected of contributing to polysaccharide degradation (RCLO1, CLOS1, COPR1-3, and TISS1, Figure S1D, Table S5). In several instances, expressed CAZymes from *BWF2A* and *SW3C* could not be resolved between the two strains and/or the COPR1 MAG. For example, all GHs within region-A could be identified as expressed by at least one of the isolated strains but could not be resolved further between the strains.

From the CAZymes subset of expression groups, a cluster analysis was performed to reveal six expression clusters (I–VI, Fig. 5b). Clusters II, III, and IV were enriched with *C. proteolyticus*-affiliated MAGs and isolated strains. Clusters III and IV comprised 10 and 11 expression groups, respectively, and followed a similar profile over time (Fig. 5c), increasing at earlier stages (T2–3) and again at later stationary/death stages (T6–8). Cluster II (10 expression groups) was slightly variant and increased more rapidly at T2 and sustained high levels over the course of SEM1b. All three clusters consisted of CAZymes targeting linkages associated with *N*-acetylglucosamine (CE9) and peptidoglycan (CE4, GH23, and GH73), suggesting a role in bacterial cell wall hydrolysis (Table S5). This hypothesis was supported by 16S rRNA gene data, which illustrated that *C. proteolyticus*-affiliated populations (OTU2), were high at initial stages of the SEM1b life cycle when cell debris was likely present in the inoculum that was sourced from the preceding culture at stationary phase (Fig. 5a). At T2, the abundance of *C. thermocellum*-affiliated populations (OTU-1) was observed to outrank *C. proteolyticus* as the community predictably shifted to cellulose utilization. However, toward stationary phase (T6–8) when dead cell debris is expected to be increasing, expression levels in clusters II, III, and IV were maintained at high levels (Fig. 5b), which was consistent with high *C. proteolyticus* 16S rRNA gene abundance at the same time points.

Clusters V and VI comprised 28 and 101 expression groups (respectively), and were enriched with the RCLO1 MAG that was closely related to *C. thermocellum*. As expected, numerous expressed genes in cluster V and VI were inferred in cellulosome assembly (via dockerin domains) as well as cellulose (e.g., GH5, GH9, GH44, GH48, CBM3) and hemicellulose (e.g., GH10, GH11, GH26, GH43, GH74) hydrolysis (Table S5). Both clusters increased throughout the consortium's exponential phase (time points T1–4, Fig. 5a), whereas 16S rRNA data also shows *C. thermocellum*-affiliated populations at high levels during the same stages (Fig. 5a).

Cluster I was determined as the largest with 121 expression groups and was particularity enriched with CLOS1, which expressed many genes involved in hemicellulose deconstruction (e.g., GH3, GH10, GH29, GH31, GH43, and GH130) and carbohydrate deacetylation (e.g., CE4, CE7, CE8, CE9, CE12, and CE15) (Table S5). Genes encoding CAZymes from both *BWF2A* and *SW3C* were also expressed in cluster I including the functionally active GH16- and GH3-encoding ORFs from region-A, which reaffirms our earlier predictions that certain *C. proteolyticus* populations in SEM1b are capable of degrading hemicellulosic substrates. The expression profile of cluster I over time was observed to slightly lag after cluster V and VI (Fig. 5), suggesting that genes encoding hemicellulases in cluster I are expressed once the hydrolytic effects of the RCLO1 cellulosome (expressed in cluster V and VI) have liberated hemicellulosic substrates [74]. Although *C. thermocellum* cannot readily utilize other carbohydrates besides glucose and longer glucans [75], the cellulosome is composed of a number of hemicellulolytic enzymes such as GH10 and GH11 endoxylanases, GH26 mannanases, GH74 xyloglucanases, and GH43 arabinanases/xylosidases [76], which are involved in the deconstruction of the underlying cellulose–hemicellulose matrix [74]. Interestingly, RCLO1 representatives of GH10, GH11, GH5, GH9, GH16, and GH43 were all expressed in the additional RCLO1-enriched cluster V and are presumably acting on the hemicellulose fraction present in the spruce-derived cellulose [77]. Furthermore, detection of hydrolysis products (Fig. 5a) revealed that xylose increased significantly at T5–7, indicating that hemicellulosic polymers containing β-1-4-xylan were likely available at these stages. Cluster V exhibited a similar profile to the other RCLO1-enriched cluster (Cluster VI), however its high expression levels were extended to T7, consistent with our observed levels of xylose release (Fig. 5c).

An additional GH16 from RCLO1 was also expressed in SEM1b cluster V, which has 99.5% amino acid sequence identity to Lic16A, a biochemically characterized endoglucanase that exerts specific β-1,3 activity similar to the *BWF2A*/*SW3C* GH16 that we report here. Notably, Lic16A is a cell wall anchored, non-cellulosomal CAZyme that is believed to enable *C. thermocellum* to grow exclusively on β-1,3-glucans [78]. All in all, the SEM1b expression data shows sequential community progression that co-ordinates putative hydrolysis of cellulose and hemicellulosic substrates as well as carbohydrates that are found in the microbial cell wall. In particular, *C. proteolyticus* populations in SEM1b were suspected to have key roles degrading microbial cell wall carbohydrates and hemicellulosic substrates, possibly in cooperation or in parallel to other clostridium populations at the later stages of the SEM1b growth cycle.

Conclusions

Unraveling the interactions occurring in a complex microbial community composed of closely related species or strains is an arduous task. Here we have leveraged culturing techniques, metagenomics, time-resolved metatranscriptomics, and enzymology to describe a novel *C. proteolyticus* population that comprised closely related strains that have acquired CAZymes via HGT and putatively evolved to incorporate a saccharolytic lifestyle. The co-expression patterns of *C. proteolyticus* CAZymes in clusters II, III, and IV supports the adaptable role of this

bacterium as a scavenger that is able to hydrolyze cell wall polysaccharides during initial phases of growth and in the stationary/death phase, when available sugars are low. Moreover, the acquisition of biochemically verified hemicellulases by *C. proteolyticus* and their co-expression in cluster I at time points when hemicellulose is available further enhances its metabolic versatility and provides substantial evidence as to why this population dominates thermophilic reactors on a global scale, even when substrates are poor in protein.

Data availability

All sequencing reads have been deposited in the sequence read archive (SRP134228), with specific numbers listed in Table S6. All microbial genomes are publicly available on JGI under the analysis project numbers listed in Table S6. The code used to perform the computational analysis is available at: https://github.com/fdelogu/SEM1b-CAZymes.git

Acknowledgements We are grateful for support from The Research Council of Norway (FRIPRO program, P.B.P.: 250479/NorZymeD, V. G.H.E.: 221568), as well as the European Research Commission Starting Grant Fellowship (awarded to P.B.P.; 336355—MicroDE). The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Regional Health Authorities.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. If you remix, transform, or build upon this article or a part thereof, you must distribute your contributions under the same license as the original. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Tandishabo K, Nakamura K, Umetsu K, Takamizawa K. Distribution and role of Coprothermobacter spp. in anaerobic digesters. J Biosci Bioeng. 2012;114:518–20.
2. Etchebehere C, Pavan ME, Zorzópulos J, Soubes M, Muxí L. *Coprothermobacter platensis* sp. nov., a new anaerobic proteolytic thermophilic bacterium isolated from an anaerobic mesophilic sludge. Int J Syst Bacteriol. 1998;48:1297–304.
3. Ollivier BM, Mah RA, Ferguson TJ, Boone DR, Garcia JL, Robinson R. Emendation of the Genus *Thermobacteroides*: *Thermobacteroides proteolyticus* sp. nov., a proteolytic acetogen from a methanogenic enrichment. Int J Syst Bacteriol. 1985;35:425–8.
4. Hagen LH, Frank JA, Zamanzadeh M, Eijsink VGH, Pope PB, Horn SJ, et al. Quantitative metaproteomics highlight the metabolic contributions of uncultured phylotypes in a thermophilic anaerobic digester. Appl Environ Microbiol. 2016;83:pii: e01955–16.
5. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science (New Y, NY). 2014;344:416–20.
6. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013;493:45–50.
7. Spanogiannopoulos P, Bess EN, Carmody RN, Turnbaugh PJ. The microbial pharmacists within us: a metagenomic view of xenobiotic metabolism. Nat Rev Microbiol. 2016;14:273–87.
8. Bron PA, Van Baarlen P, Kleerebezem M. Emerging molecular insights into the interaction between probiotics and the host intestinal mucosa. Nat Rev Microbiol. 2012;10:66–78.
9. Ellegaard KM, Engel P. Beyond 16S rRNA community profiling: Intra-species diversity in the gut microbiota. Front Microbiol. 2016;7:1–16.
10. Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science. 2008;320:1081–5.
11. McLoughlin K, Schluter J, Rakoff-Nahoum S, Smith AL, Foster KR. Host selection of microbiota via differential adhesion. Cell Host Microbe. 2016;19:550–9.
12. Rosenzweig RF, Sharp RR, Treves DS, Adams J. Microbial evolution in a simple unstructured environment: genetic differentiation in *Escherichia coli*. Genetics. 1994;137:903–17.
13. Zelezniak A, Andrejev S, Ponomarova O, Mende DR, Bork P, Patil KR. Metabolic dependencies drive species co-occurrence in diverse microbial communities. Proc Natl Acad Sci USA. 2015;112:6449–54.
14. Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F, et al. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7:828–828.
15. Gill SR, Fouts DE, Archer GL, Mongodin EF, Deboy RT, Ravel J, et al. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. J Bacteriol. 2005;187:2426–38.
16. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. 2013;23:111–20.
17. Solheim M, Aakra Å, Snipen LG, Brede DA, Nes IF. Comparative genomics of *Enterococcus faecalis* from healthy Norwegian infants. BMC Genom. 2009;10:1–11.
18. Zunino P, Piccini C, Legnani-Fajardo C. Flagellate and non-flagellate *Proteus mirabilis* in the development of experimental urinary tract infection. Microb Pathog. 1994;16:379–85.
19. Siezen RJ, Tzeneva VA, Castioni A, Wels M, Phan HTK, Rademaker JLW, et al. Phenotypic and genomic diversity of

Lactobacillus plantarum strains isolated from various environmental niches. Environ Microbiol. 2010;12:758–73.
20. Koskella B, Vos M. Adaptation in natural microbial populations. Annu Rev Ecol, Evol, Syst. 2015;46:503–22.
21. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci USA. 2005;102:13950–55.
22. Treangen TJ, Rocha EPC. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 2011;7:e1001284.
23. Ochman H, Lawrence JG, Grolsman EA. Lateral gene transfer and the nature of bacterial innovation. Nature. 2000;405:299–304.
24. Bendall ML, Stevens SLR, Chan LK, Malfatti S, Schwientek P, Tremblay J, et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 2016;10:1589–601.
25. Biller SJ, Berube PM, Lindell D, Chisholm SW. Prochlorococcus: The structure and function of collective diversity. Nat Rev Microbiol. 2015;13:13–27.
26. Shapiro BJ, Timberlake SC, Szabó G, Polz MF, Alm EJ. Population genomics of early differentiation of bacteria. Science. 2012;336:48–51.
27. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure & genetic diversity from metagenomes. Genome Res. 2017;27:626–38.
28. González-Torres P, Pryszcz LP, Santos F, Martínez-García M, Gabaldón T, Antón J. Interactions between closely related bacterial strains are revealed by deep transcriptome Sequencing. Appl Environ Microbiol. 2015;81:8445–56.
29. Alexiev A, Coil DA, Badger JH, Enticknap J, Ward N, Robb FT, et al. Complete genome sequence of *Coprothermobacter proteolyticus* DSM 5265. Genome Announ. 2014;2: pii: e00470–14.
30. Zamanzadeh M, Hagen LH, Svensson K, Linjordet R, Horn SJ. Anaerobic digestion of food waste - effect of recirculation and temperature on performance and microbiology. Water Res. 2016;96:246–54.
31. Rødsrud G, Lersch M, Sjöde A. History and future of world's most advanced biorefinery in operation. Biomass Bioenergy. 2012;46:46–59.
32. Kunath BJ, Bremges A, Weimann A, McHardy AC, Pope PB. Metagenomics and CAZyme Discovery. In: Abbott DW, Lammerts van Bueren A, editors. Protein-carbohydrate interactions: methods and protocols. New York, NY: Springer New York; 2017. p. 255–77.
33. Takahashi S, Tomita J, Nishioka K, Hisada T, Nishijima M. Development of a prokaryotic universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing. PLoS ONE. 2014;9:e105592.
34. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460–1.
35. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335–6.
36. Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res. 2015;43:D593–8.
37. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17:10–10.
38. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arxiv. 2013;00:1–3.
39. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. MetaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27:824–34.
40. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165–e1165.
41. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
42. Hungate RE (1969). Chapter IV A roll tube method for cultivation of strict anaerobes. In: Norris JR, Ribbons DWBTMiM, editors. Methods in microbiology. Chapter IV. Academic Press. p. 117–32.
43. Schumann P. Nucleic acid techniques in bacterial systematics (modern microbiological methods). J Basic Microbiol. 1991;31:479–80.
44. Chen IMA, Markowitz VM, Chu K, Palaniappan K, Szeto E, Pillay M, et al. IMG/M: Integrated genome and metagenome comparative data analysis system. Nucleic Acids Res. 2017;45:D507–16.
45. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42:D490–5.
46. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics. 2011;12:402.
47. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3.
48. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016;1:1–6.
49. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
50. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33:1870–4.
51. Letunic I, Bork P. Interactive tree of life (iTOL)v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44:W242–5.
52. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Prepr. 2016;4:e1900–1.
53. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785.
54. Aslanidis C, de Jong PJ. Ligation-independent cloning of PCR products (LIC-PCR). Nucleic Acids Res. 1990;18:6069–74.
55. Miller GL. Use of dinitrosalicylic acid reagent for determination of reducing sugar. Anal Chem. 1959;31:426–8.
56. Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem. 1976;72:248–54.
57. Zhou Y, Pope PB, Li S, Wen B, Tan F, Cheng S, et al. Omics-based interpretation of synergism in a soil-derived cellulose-degrading microbial community. Sci Rep. 2014;4:1–6.
58. Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA. Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. ISME J. 2011;5:461–72.
59. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
60. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28:3211–7.
61. Langmead. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

62. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
63. Turro E, Su SY, Gonçalves Â, Coin LJM, Richardson S, Lewin A. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biol. 2011;12:1–15.
64. Turro E, Astle WJ, Tavaré S. Flexible analysis of RNA-seq data using mixed effects models. Bioinformatics. 2014;30:180–8.
65. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics. 2008;24:719–20.
66. Kersters I, Maestrojuan GM, Torck U, Vancanneyt M, Kersters K, Verstraete W. Isolation of *Coprothermobacter proteolyticus* from an anaerobic digest and further characterization of the species. Syst Appl Microbiol. 1994;17:289–95.
67. Johnson JW, Fisher JF, Mobashery S. Bacterial cell wall recycling. Ann New Y Acad. 2013;1277:54–75.
68. Lü F, Bize A, Guillot A, Monnet V, Madigou C, Chapleur O, et al. Metaproteomics of cellulose methanisation under thermophilic conditions reveals a surprisingly high proteolytic activity. ISME J. 2014;8:88–102.
69. Miller JH, Novak JT, Knocke WR, Pruden A. Survival of antibiotic resistant bacteria and horizontal gene transfer control antibiotic resistance gene content in anaerobic digesters. Front Microbiol. 2016;7:1–11.
70. Hehemann JH, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature. 2010;464:908–12.
71. Ricard G, McEwan NR, Dutilh BE, Jouany JP, Macheboeuf D, Mitsumori M, et al. Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. BMC Genomics. 2006;7:1–13.
72. Song T, Xu H, Wei C, Jiang T, Qin S, Zhang W, et al. Horizontal transfer of a novel soil agarase gene from marine bacteria to soil bacteria via human microbiota. Sci Rep. 2016;6:1–10.
73. Modi SR, Lee HH, Spina CS, Collins JJ. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. Nature. 2013;499:219–22.
74. Zverlov VV, Schantz N, Schmitt-Kopplin P, Schwarz WH. Two new major subunits in the cellusome of *Clostridium thermocellum*: xyloglucanase Xgh74A and endoxylanase Xyn10D. Microbiology. 2005b;151:3395–401.
75. Demain AL, Newcomb M, Wu JHD, Demain AL, Newcomb M, Wu JHD. Cellulase, clostridia, and ethanol. Microbiol Mol Biol Rev. 2005;69:124–54.
76. Zverlov VV, Kellermann J, Schwarz WH. Functional subgenomics of *Clostridium thermocellum* cellulosomal genes: identification of the major catalytic components in the extracellular complex and detection of three new enzymes. Proteomics. 2005a;5:3646–53.
77. Chylenski P, Petrović DM, Müller G, Dahlström M, Bengtsson O, Lersch M, et al. Enzymatic degradation of sulfite-pulped softwoods and the role of LPMOs. Biotechnol Biofuels. 2017;10:1–13.
78. Fuchs K-P, Zverlov VV, Velikodvorskaya GA, Lottspeich F, Schwarz WH. Lic16A of *Clostridium thermocellum*, a non-cellulosomal, highly complex endo-β-1,3-glucanase bound to the outer cell surface. Microbiology. 2003;149:1021–31.
79. Jeng W-Y, Wang N-C, Lin C-T, Shyur L-F, Wang AHJ. Crystal structures of the laminarinase catalytic domain from *Thermotoga maritima* MSB8 in complex with inhibitors: essential residues for β-1,3 and β-1,4 glucan selection. J Biol Chem. 2011;286:45030–40.