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

B.J. Kunath[1#], F. Delogu[1#], M.Ø. Arntzen[1], V.G.H. Eijsink[1], T.R. Hvidsten[1], P.B. Pope[1*]

1. Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1432 Ås, NORWAY.

[#] **Equal contributors**

***Corresponding Author:** Phillip B. Pope
Faculty of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences
Post Office Box 5003
1432, Ås, Norway
Phone: +47 6496 6232
Email: phil.pope@nmbu.no

RUNNING TITLE:

The genetic plasticity of *Coprothermobacter*

COMPETING INTERESTS

The authors declare there are no competing financial interests in relation to the work described.

KEYWORDS

CAZymes; Horizontal gene transfer; strain heterogeneity; Metatranscriptomics

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, whilst 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 various carbohydrate-active enzymes (CAZymes), presumably derived from horizontal gene transfer (HGT) events involving *C. thermocellum*- or Thermotogae-affiliated populations that are historically co-located. HGT material included whole saccharolytic operons and dockerin-encoding enzymatic subunits that are synonymous with cellulosomes. 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 towards co-degradation of cellulose and hemicellulose*.* We show that *C. proteolyticus*, a ubiquitous keystone population, consists of closely related strains that have adapted via HGT to 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 whilst generating hydrogen and acetate, key intermediate metabolites for biogas production (Tandishabo *et al.,* 2012). 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 (**Figure 1**). Despite their promiscuous distribution, global abundance and key role in biogas production, only two species have been described: *Coprothermobacter platensis* (Etchebehere *et al.,* 1998) and *Coprothermobacter proteolyticus* (Ollivier *et al.,* 1985). 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 (Hagen *et al.,* 2017).

Frequently in nature, microbial populations are composed of multiple strains with genetic heterogeneity (Kashtan *et al.,* 2014, Schloissnig *et al.,* 2013). Studies of strain-level populations have been predominately performed with the human microbiome and especially the gut microbiota (Bron *et al.,* 2012, Spanogiannopoulos *et al.,* 2016). The reasons for strain diversification and their coexistence remain largely unknown (Ellegaard and Engel 2016), however several mechanisms have been hypothesized, such as: micro-niche selection (Hunt *et al.,* 2008, Kashtan *et al.,* 2014), host selection (McLoughlin *et al.,* 2016), cross-feed interactions (Rosenzweig *et al.,* 1994, Zelezniak *et al.,* 2015) and phage selection (Rodriguez-Valera *et al.,* 2009). Studies of isolated strains have shown that isolates can differ in a multitude of ways, including virulence and drug resistance (Gill *et al.,* 2005, Sharon *et al.,* 2013, Solheim *et al.,* 2009), motility (Zunino *et al.,* 1994) and nutrient utilization (Siezen *et* *al.,* 2010). Strain-level genomic variations typically consist of single-nucleotide variants (SNVs) as well as acquisition/loss of genomic elements such as genes, operons or plasmids via horizontal gene transfer (HGT) (Koskella and Vos 2015, Tettelin *et al.,* 2005, Treangen and Rocha 2011). Variability in gene content caused by HGT is typically attributed to phage-related genes and other genes of unknown function (Ochman *et al.,* 2000), and can give rise to ecological adaptation, niche differentiation and eventually speciation (Bendall *et al.,* 2016, Biller *et al.,* 2015, Shapiro *et al.,* 2012). 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 (Truong *et al.,* 2017). Since closely-related strains can also differ in gene expression (González-Torres *et al.,* 2015), 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 to the only available type strain, *C. proteolyticus* DSM 5265. Notable genomic differences included the acquisition of carbohydrate-active enzymes (CAZymes), which inferred that *C. proteolyticus* has adapted to take advantage of lignocellulosic polysaccharides. 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.

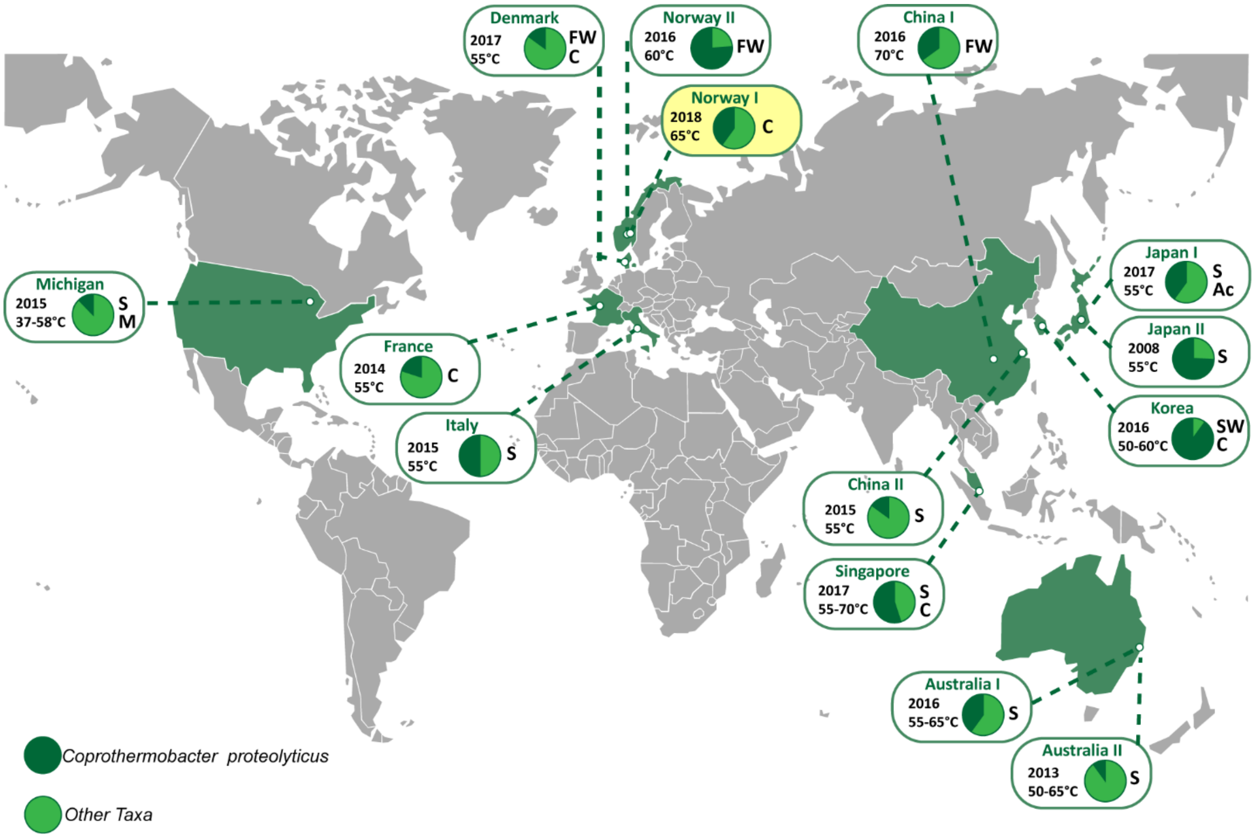

Figure 1. Global distribution of *Coprothermobacter proteolyticus*-affiliated populations in anaerobic biogas reactors. Charts indicate relative 16S rRNA gene abundance of OTUs affiliated to *C. proteolyticus* (dark green), in comparison to the total community (light green). The year of publication, 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.

MATERIALS AND METHODS

Origin of samples and generation of the SEM1b consortium

An inoculum (100µl) 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 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 (Hagen *et al.*, 2017) and the TD bioreactor (Zamanzadeh *et al.*, 2016), which provided a detailed understanding of the original microbial community. The inoculum was transferred for serial dilution and enrichment to an anaerobic serum bottle incubated at 65°C and containing the rich ATCC medium 1943, with cellobiose substituted for 10g/L of cellulose in the form of

BALI treated Norway spruce (Rødsrud *et al.,* 2012). After an initial growth cycle, an aliquot was removed and serially diluted to extinction. Briefly, a 100µl sample was transferred to a new bottle of anaerobic medium, mixed and 100µl 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 generation 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 generations older than D1B and was used to leverage improvements in metagenome assembly and binning. From 6ml of culture, cells were pelleted by centrifugation at 14000 x *g* for 5 minutes 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 (Kunath *et al.,* 2017). 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'-CCTACGGGNBGCASCAG-3' / 5'-GACTACNVGGGTATCTAATCC-3' (Takahashi *et al.,* 2014). The PCR was performed as previously described (Zamanzadeh *et al.,* 2016) and the sequencing library was prepared using Nextera XT Index kit according to Illumina's instructions for the MiSeq system (Illumina Inc.). MiSeq sequencing (2x300bp with paired-ends) was conducted using the MiSeq Reagent Kit v3. The reads were quality filtered (Phred ≥ Q20) and USEARCH61 (Edgar 2010) 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 (Caporaso *et al.,* 2010) using the Greengenes database (gg_13_8). The resulting OTU table was corrected based on the predicted number of *rrs* operons for each taxon (Stoddard *et al.,* 2015).

D1B and D2B were also subjected to metagenomic shotgun sequencing using the Illumina HiSeq3000 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 (2x125bp) on four lanes (two lanes per sample). Quality trimming of the raw reads was performed using cutadapt (Martin 2011), removing all bases on the 3' end with a Phred score lower than 20 (if any present) and excluding all reads shorter than 100nt, 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 BWA-MEM algorithm requiring 100% identity (Li 2013). Reads that mapped the strains were removed from the metagenomic data and the remaining reads were co-assembled using MetaSpades v3.10.0 (Nurk *et al.,* 2017) with default parameters and k-mer sizes of 21, 33, 55 and 77. The subsequent contigs were binned with Metabat v0.26.3 (Kang *et al.,* 2015) 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 (Parks *et al.,* 2015) with default parameters.

Isolation of *C. proteolyticus* strains

Strains were isolated using the Hungate method (Hungate 1969). In brief: Hungate tubes were anaerobically prepared with the DSMZ medium 481 with and without agar (15g/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, 100µl were transferred to a new tube and mixed. From this new tube, 100µl was directly transferred to 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 (Schumann 1991): 5'-AGAGTTTGATCMTGGCTCAG-3' / 5'-TACGGYTACCTTGTTACGACTT-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 Clean-up 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 Norwegian Sequencing Center (NSC, Oslo, Norway). Samples were prepared with the TrueSeq DNA PCR-free preparation and sequenced using paired-ends (2x300bp) on a MiSeq system (Illumina Inc). Quality trimming, filtering and assembly were performed as described in the aforementioned metagenomic assembly section. All contigs from the strains and the Metagenome Assembled Genomes (MAGs) were submitted to the Integrated Microbial Genomes and Microbiomes (IMG/M) system (Chen *et al.,* 2017) for genomic feature prediction and annotation (pipeline version 4.15.1). Resulting annotated open reading frames (ORFs) were retrieved, further annotated for carbohydrate-active enzymes (CAZymes) using dbCAN HMMs v5.0 (Yin *et al.,* 2012), and subsequently used as a reference database for the metatranscriptomics. The genomes for both Strains and MAGs corresponding to *C. proteolyticus* were compared to the reference genome from *C. proteolyticus* DSM 5265. Using the BRIG tool (Alikhan *et al.,* 2011) for mapping and visualization, the different genomes were mapped against their pan genome generated using Roary (Page *et al.,* 2015).

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) (Hug *et al.,* 2016). The dataset was augmented with metagenomic sequences retrieved from our previous research on the original FREVAR reactor (Hagen *et al.,* 2017) and with sequences from reference genomes identified during the 16S rRNA analysis. Each gene set was individually aligned using MUSCLE v3.8.31 (Edgar 2004) 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 (Kumar *et al.,* 2016) with 1000 bootstrap replicates. The radial tree was visualized using iTOL (Letunic and Bork 2016). Additionally, an average nucleotide identity (ANI) comparison was performed between each MAG and their closest relative using the ANI calculator (Rodriguez-R and Konstantinidis 2016).

Temporal meta-omic analyses of SEM1b

A "meta-omic" time series analysis was conducted over the lifetime span of the SEM1b consortium (≈45hours). A collection of 27 replicate bottles containing ATCC medium 1943 with 10g/L of cellulose, 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 hours) 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 16000 x *g* for 5 minutes and the supernatants were filtered with 0.2μm sterile filters and boiled for

15 minutes before being stored at -20°C until processing. Solubilized sugars released during microbial hydrolysis were identified and quantified by high-performance anion exchange chromatography (HPAEC) with pulsed amperiometric detection (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-minute isocratic run at 1 mM KOH at 30°C. For quantification, peaks were compared to 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 proteins measurements were taken to estimate SEM1b growth rate. Proteins were extracted following a previously described method (Hagen *et al.,* 2017) with a few modifications. Briefly, 30ml culture aliquots were centrifuged at 500 x *g* for 5 minutes to remove the substrate and the supernatant was centrifuged at 9000 x *g* for 15 minutes to pellet the cells. Cell lysis was performed by resuspending the cells in 1ml of lysis buffer (50 mM Tris-HCl, 0.1% (v/v) Triton X-100, 200 mM NaCl, 1 mM DTT, 2mM EDTA) and keeping them on ice for 30 minutes. Cells were disrupted in 3 x 60 second cycles using a FastPrep24 (MP Biomedicals, USA) and the debris were removed by centrifugation at 16000 x *g* for 15 minutes. Supernatants containing proteins were transferred into low bind protein tubes and the proteins were quantified using Bradford's method (Bradford 1976).

Because 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 x *g* for 5 minutes and the supernatant was discarded. The resulting pellets were boiled under acidic conditions as previously described (Zhou *et al.,* 2014) and the dried weights, corresponding to the remaining cellulose, were measured.

mRNA extraction was performed in triplicate on time points T2 to T8, using previously described methods (Gifford *et al.,* 2011) 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 994nt 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 (25ng) 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 ThermoFisher 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 Norwegian Sequencing Center (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 (2x125bp) 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 (Bolger *et al.,* 2014). Subsequently, a quality filtering was applied with an average Phred threshold of 30 over a 10nt window and a minimum read length of 100nt. rRNA and tRNA were removed using SortMeRNA v.2.1b (Kopylova *et al.,* 2012). 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 (Langmead 2012) with default parameters and the number of reads per transcript was quantified. The remaining reads were pseudoaligned against the metagenomic dataset, augmented with the annotated strains, using Kallisto pseudo –pseudobam (Bray *et al.,* 2016). The resulting output was used to generate mapping files with bam2hits, which were used for expression quantification with mmseq (Turro *et al.,* 2011). Of the 40126 ORFs identified from the assembled SEM1b metagenome and two *C. proteolyticus* strains, 16060 (40%) were not found to be expressed, whereas 21482 (54%) were expressed and could be reliably quantified due to unique hits (reads mapping unambiguously against one unique ORF) (**Figure S1A**). The remaining 2584 ORFs (6%) 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**). Since having unique hits improves the expression estimation accuracy, the ORFs were grouped using mmseq in order to improve the precision of expression estimates, with only a small reduction in biological resolution (Turro *et al.,* 2014). The process first collapses ORFs into expression 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 39242 expression groups of which 38535 (98%) were singletons (groups composed of single ORF) and 707 (2%) were groups containing more than one homologous ORF. From the initial 2584 low-information ORFs, 1333 became part of an expression group containing unique hits, 116 became part of ambiguous group (no unique hits) and 1135 remained singletons (without unique hits). All expression groups without unique hits were then excluded from the subsequent analysis. A total of 21482 singletons and 594 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 (Gifford *et al.,* 2011). The number of reads mapping on the defined region of the internal standard molecule were calculated to be 2.2x10^7 +/- 2.2x10^6 reads per sample out of 6.2x10^9 molecules added. Using this information, the estimated number of transcript molecules per sample was computed to be 5.1x10^12 +/- 3.7x10^12 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 (PCA) 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 (Langfelder *et al.,* 2008) with hybrid mode, deepsplit=1 and minClusterSize=7. Eigengenes were computed for the clusters and clusters with a Pearson Correlation Coefficient (PCC) greater than 0.8 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 0.01. Expression groups composed of multiple MAGs/strains were included in several enrichment tests.

RESULTS AND DISCUSSION

The SEM1b consortium is a simplistic community, co-dominated by Clostridium 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 (**Figure 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 *Clostridium (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 (Hagen *et al.,* 2017). 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 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.

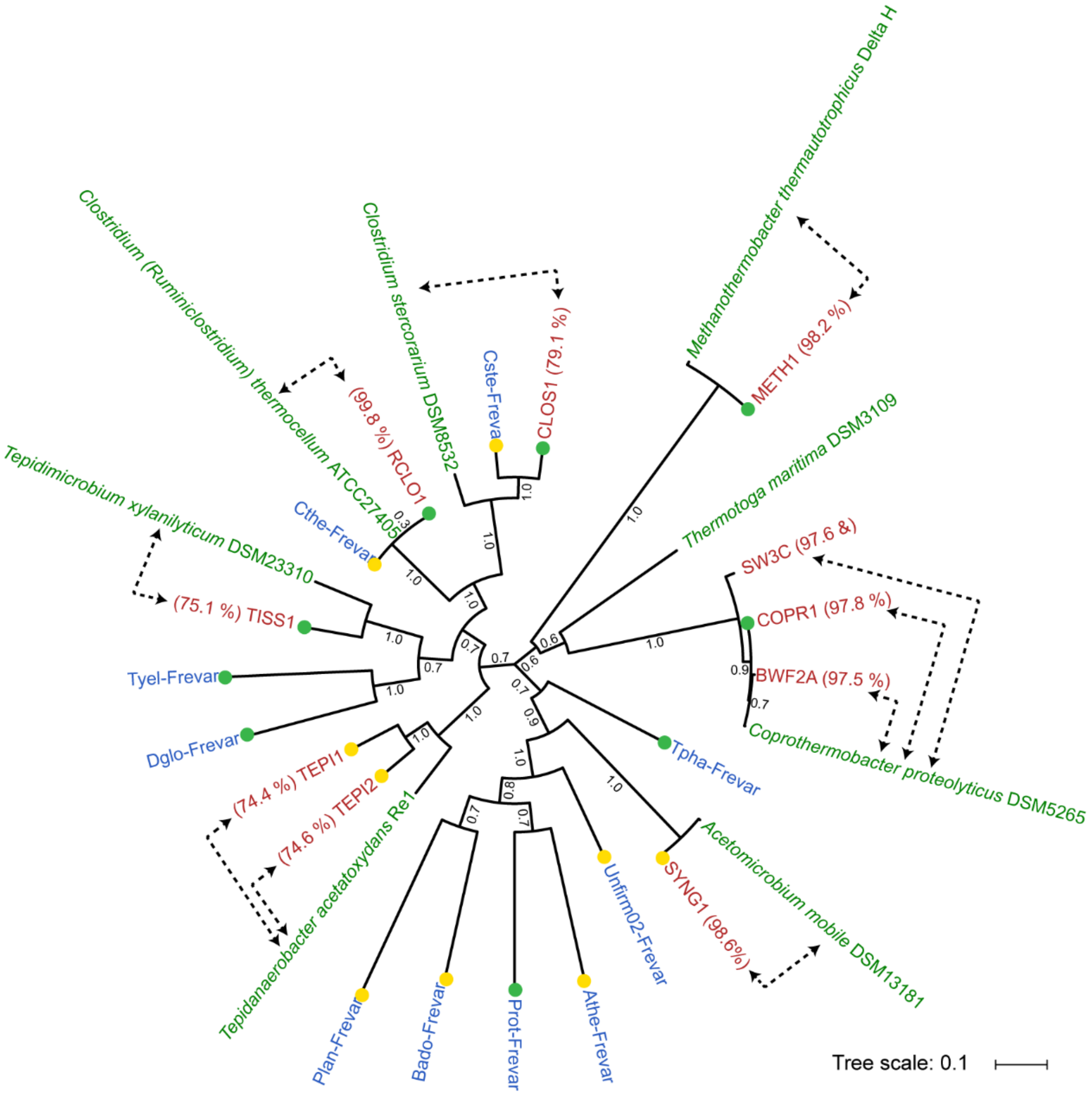
Methanothermobacter thermautotrophicus Delta H
METH1 (98.2 %)
Thermotoga maritima DSM3109
SW3C (97.6 &)
COPR1 (97.8 %)
BWF2A (97.5 %)
Coprothermobacter proteolyticus DSM5265
Tpha-Frevar
Acetomicrobium mobile DSM13181
SYNG1 (98.6%)
Unfirm02-Frevar
Athe-Frevar
Prot-Frevar
Bado-Frevar
Plan-Frevar
Tepidanaerobacter acetatoxydans Re1
(74.6 %) TEPI2
(74.4 %) TEPI1
Dglo-Frevar
Tyel-Frevar
(75.1 %) TISS1
Tepidimicrobium xylanilyticum DSM23310
Cthe-Freva
Clostridium (Ruminiclostridium) thermocellum ATCC27405
(99.8 %) RCLO1
Clostridium stercorarium DSM8532
Cste-Freva
CLOS1 (79.1 %)
Tree scale: 0.1

Figure 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, Hagen et al., 2017) 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.

Shotgun metagenome sequencing of two SEM1b samples (D1B and D2B), generated 290Gb (502M paired-end reads) and 264Gb (457M paired-end reads) of data, respectively. Co-assembly of both datasets using strain-depleted reads with Metaspades produced 20760 contigs totalizing 27Mbp with a maximum contig length of 603Kbp. Functional annotation resulted in 36292 annotated ORFs and taxonomic binning revealed 11 MAGs and a community structure similar to the one observed by 16S analysis (**Figure 2**, **Table S3**). A total of eight MAGs exhibited good 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 average nucleotide identities (ANI) and a phylogenomic tree was constructed via analysis of 16 concatenated ribosomal proteins (**Figure 2**). The COPR1 MAG was observed to cluster together with *C. proteolyticus* DSM 5265 and the two strains *BWF2A* and *SW3C.* 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 290 and 160 carbohydrate-active enzymes (CAZymes), respectively (**Table S4**). RCLO1 in particular encoded cellulolytic (e.g. glycosyl hydrolase (GH) families GH5, GH9, GH48) and cellulosomal features (dockerins and cohesins), whereas CLOS1 appears more specialized towards hemicellulose degradation (e.g. GH3, GH10, GH26, GH43, GH51, GH130). Surprisingly, CAZymes were also identified in COPR1 (n=30) and both *BWF2A* (n=37) and *SW3C* (n=33) at levels higher than what has previously been observed in *C. proteolyticus* DSM 5265 (n=14) (**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).

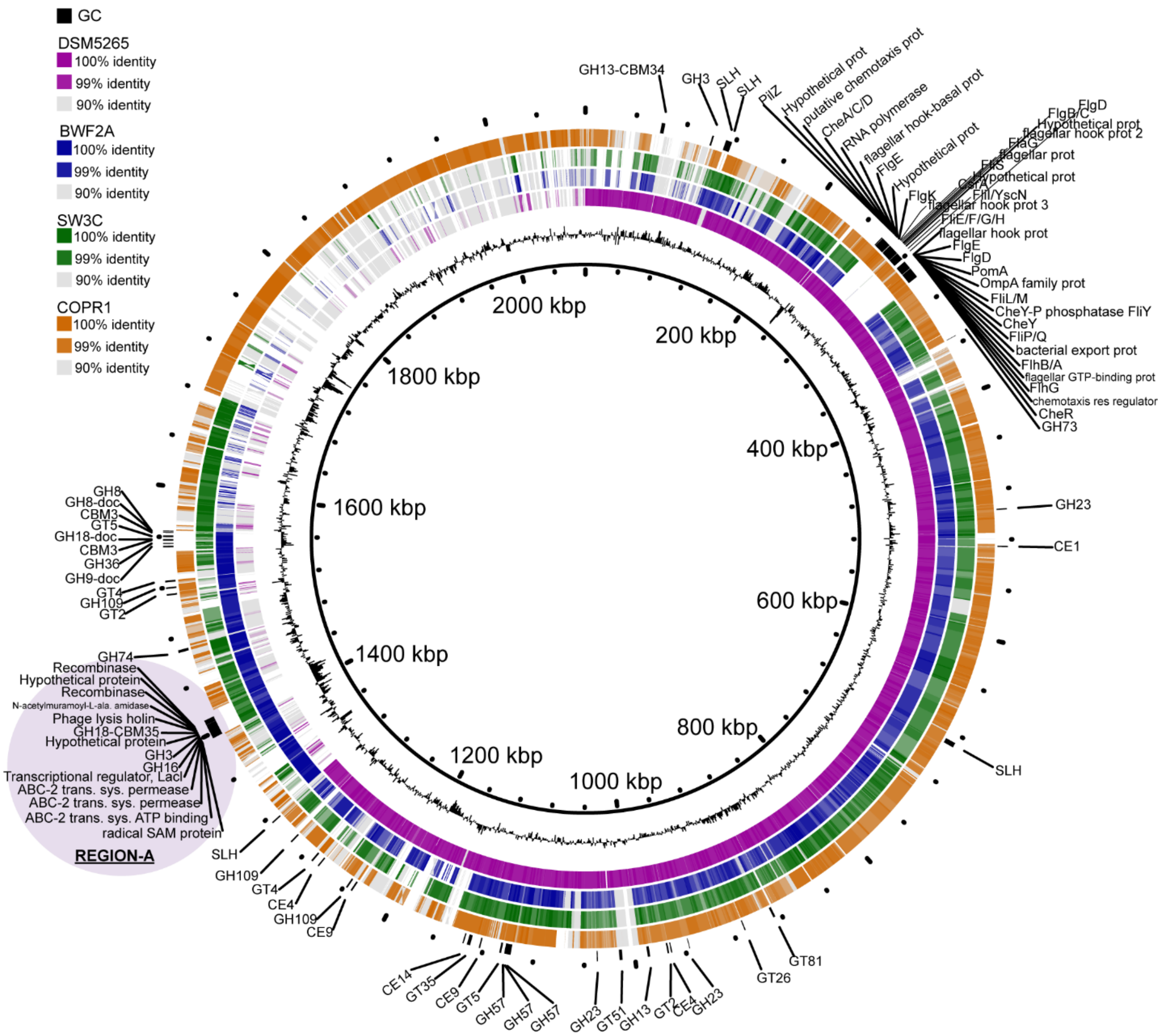

Figure 3. Comparative genome content of *C. proteolyticus* representatives including isolated strains, a recovered MAG (COPR1) and the reference strain DSM 5265. The innermost ring corresponds to the pangenome of the three *C. proteolyticus* spp. genomes and one MAG as produced by Roary (Page et al., 2015) 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.

Novel strains of C. proteolyticus reveal acquisition of carbohydrate-active enzymes

Genome annotation of COPR1, *BWF2A* and *SW3C* identified both insertions and deletions in comparison to the only available reference genome, sequenced from the type strain DSM 5265 (**Figure 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, while in contrast, they both acquired numerous extra CAZymes. Although both strains encoded a slightly different collection of CAZymes, they both contained a particular genomic region that encoded a cluster of three CAZymes: GH16, GH3 and GH18-CBM35 (region-A, **Figure 3**). Other notable CAZy families found throughout the genome included GH36, GH74, GH109, CBM3, CE1 and CE10 as well as GH9, GH8 and GH18 domains co-located with a dockerin. The putative function of these GHs, suggests that both *BWF2A* and *SW3C* are capable of hydrolyzing amorphous forms of cellulose (GH9: endoglucanase, CBM3: cellulose-binding) and/or different hemicellulosic substrates (GH16: endo-1,3(4)-β-glucanase / xyloglucanase; GH3: β-glucosidase / 1,4-β-xylosidase; GH36: α-galactosidase) (Álvarez *et al.,* 2016), which are significant fractions of the spruce-derived cellulose substrate (glucan: 88.3%, xylan: 4.5%, mannan: 4.8%, (Chylenski *et al.,* 2017)). Regarding the putative GH18s encoded in both strains (2 ORFs, with and without a dockerin), it could play a role in bacterial cell wall recycling (Johnson *et al.,* 2013) 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 (Lü *et al.,* 2014).

Taking a closer look, the region-A of CAZymes (GH16, GH3, GH18-CBM35) in *BWF2A* and *SW3C* was located on the same chromosomal cassette but organized onto two different operons with opposite directions (**Figure S2A**). Comparison of the genes and their organization, revealed a high percentage of gene similarity and synteny with genome representatives of *Fervidobacterium nodosum* (Phylum: Thermotogae) and *C. thermocellum* (**Table S5**). Both populations were previously identified in the original Frevar reactor (Hagen *et al.,* 2017), and *C. thermocellum* representatives are also found in SEM1b (RCLO1). 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 (**Figure 3, Figure S2A**). Further comparisons with *F. nodosum* and *C. thermocellum* illustrated that only the later encodes the same prophage elements, together with an additional terminase and more phage component proteins on the 5' region (**Figure**

S2B). Aside from region-A, most CAZymes encoded within *BWF2A* and *SW3C* genomes also exhibited high similarity to representatives from *C. thermocellum* (**Table S5**). Considering the high sequence homology and presence of phage-genes, it could be hypothesized that genomic regions encoding CAZymes in *BWF2A* and *SW3C* are the result of HGT originating from *C. thermocellum* and *F. nodosum*. HGT within anaerobic digesters has been reported for antibiotic resistance genes (Miller *et al.,* 2016), whereas HGT of CAZymes have been detected previously among gut microbiota (Hehemann *et al.,* 2010, Ricard *et al.,* 2006, Song *et al.,* 2016). Since many microbes express only a specific array of carbohydrate-degrading capabilities, bacteria that acquire CAZymes from phages may gain additional capacities and consequently, a selective growth advantage (Modi *et al.,* 2013).

Interestingly, several *BWF2A* and *SW3C* CAZymes were found to encode a GH domain coupled with a dockerin (GH-doc) (**Table S5**), which are typical building block of cellulosomes. Cellulosomes are multi-enzyme complexes produced by anaerobic cellulolytic bacteria for the degradation of lignocellulosic biomass, and were first discovered in *C. thermocellum* (Lamed et al., 1983). These complexes are composed of two main types of building blocks: dockerin-containing enzymatic subunits and cohesin-containing structural proteins called scaffoldins (Artzi *et al.,* 2017). Surprisingly, no cohesins were detected in *BWF2A* or *SW3C*. Dockerin-only containing genomes have previously been observed, however they remain poorly understood (Dassa *et al.,* 2014). In addition to cellulosome building blocks, examination of the genome regions flanking a GH8 encoded in *BWF2A* (GH8, IMG geneID: 2731989313), revealed both anti-sigma and RNA polymerase sigma factors (**Figure S3**). Previous studies have shown that cellulosomal genes are regulated by anti-sigma factors and alternative sigma factors in a substrate-dependent way (Artzi *et al.,* 2017, Nataf *et al.,* 2010), and are unique to *C. thermocellum (Kahel-Raifer et al., 2010)*. The anti-sigma system encoded upstream to the *BWF2A* GH8 exhibited 99% sequence identity to the corresponding *C. thermocellum* genes that harbor the same gene organization (**Figure S3**). Together, the high gene similarity, and presence of GH-doc domains and the anti-sigma factors supports our previous assertions that many of the GHs present in *BWF2A* and *SW3C* originated from *C. thermocellum*. Since the anti-sigma factor has an exocellular CBM-like component to detect the presence of a particular substrate, it is tempting to speculate that

GHs from *C. proteolyticus* could be expressed and bind the complementary cohesin modules produced by *C. thermocellum.*

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

To better understand the role(s) played by *C. proteolyticus* in a saccharolytic consortium, a temporal metatranscriptomic analyses of SEM1b over a complete life cycle was performed. 16S rRNA gene analysis of eight time points (T1-8) over a 43hr period reaffirmed that *C. thermocellum*- and *C. proteolyticus*-affiliated populations dominate SEM1b over time (**Figure 4A**). 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 408 singleton CAZyme expression groups and 13 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 S6**). In several instances, expressed CAZymes from *BWF2A* and *SW3C* could be distinguished from their original sources (i.e. *C. thermocellum* and/or *F. nodosum*), but 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. In contrast, the GH9-doc and GH8-doc ORFs were unambiguously expressed and could not be resolved between *BWF2A* and *SW3C* and the RCLO1 MAG, whereas GH8, GH18 and CBM3 ORFs were expressed by at least one of the *C. proteolyticus* strains but could not be resolved further.

From the CAZymes subset of expression groups, a cluster analysis was performed to reveal eight expression clusters (I-VIII, **Figure 4B**). Clusters I and II comprised 13 and 10 expression groups (respectively) and followed a similar profile over time (**Figure 4C**), increasing at earlier stages (T2-3) and again at later stationary/death stages (T6-8). Both clusters were enriched for *C. proteolyticus*-affiliated MAGs and isolated strains and predominately consisted of CAZymes targeting linkages associated with N-acetylglucosamine (CE9, CE14), peptidoglycan (GH23, GH18, GH73) and chitosan (GH8), suggesting a role in bacterial cell wall hydrolysis (**Table S6**). 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 (**Figure 4A**). 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, towards stationary phase (T6-8) when dead cell debris is expected to be increasing, expression levels in clusters I and II were maintained at high levels (**Figure 4B**), which was consistent with high *C. proteolyticus* 16S rRNA gene abundance at the same time-points.

Cluster IV, which was the second largest with 161 expression groups, was enriched with the RCLO1 MAG that was closely related to *C. thermocellum*. As expected, numerous expressed genes in cluster IV were inferred in cellulosome assembly (via cohesin and dockerin domains) as well as cellulose (e.g. GH5, GH9, GH44, GH48, CBM3) and hemicellulose (e.g. GH10, GH26, GH43, GH74) hydrolysis (**Table S6**). This cluster was increasing throughout the consortium's exponential phase (time points T1-4, **Figure 4A**), whilst 16S rRNA data also shows *C. thermocellum*-affiliated populations at high levels during the same stages (**Figure 4A**). Interestingly, the *BWF2A* and *SW3C* GH18-doc was also found in cluster IV. A GH9-doc and GH8-doc encoded in *BWF2A* and *SW3C* were also expressed, however they exhibited 99-100% identity to *C. thermocellum* representatives (**Table S5**) and formed expression groups without unique hits, hence they were not part of the clustering (**Table S6**). Since cellulosomal genes are seemingly expressed collectively in order to facilitate coordinated assembly, finding the GH18 in the same cluster is not overly surprising. However, the fact that the gene is expressed by a different bacterium to the one creating the bulk of the cellulosome machinery (i.e. RCLO1) makes it extremely interesting and supports the possibility that "multi-species" cellulosomes putatively exist. Although species-specific high-affinity cohesin-dockerin interactions are required for cellulosome assembly (Pagès *et al.,* 1997), the *C. thermocellum*-origin and high homology of the *C. proteolyticus* GH18 gene lends itself to the hypothesis that once expressed, the GH18 will bind to a RCLO1 cohesin domain and be part of the resulting cellulosomes. Obviously, much more experimental validation is required to confirm physical interactions between *C. proteolyticus* GH-doc representatives and cellulosome assemblies.

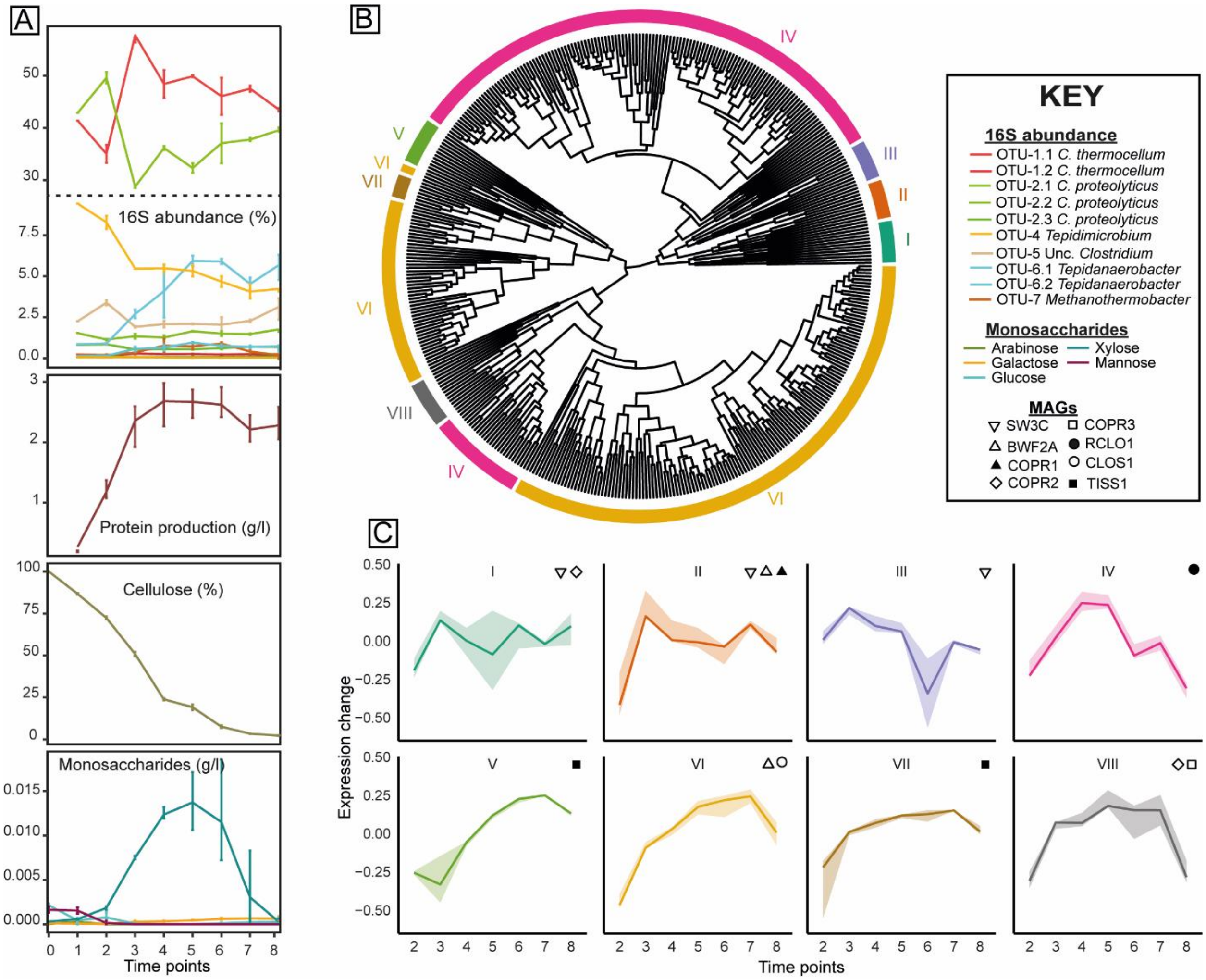

Figure 4. Temporal meta-analysis of the SEM1b consortium. (A) 16S rRNA gene amplicon and metadata analysis was performed over a 43-hour period, which was segmented into 9 time-points. OTU IDs are detailed in **Table S2**. Cellulose degradation rate, monosaccharide accumulation and growth rate (estimated by total protein concentration) is presented. (B) Gene expression dendrogram and clustering of CAZymes from BWF2A, SW3C and MAGs: RCLO1, CLOS1, COPR1-3, and TISS1. Eight expression clusters (I-VIII) are displayed in different colors on the outer ring. (C) Clusters I-VIII 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 < 0.01) in the clusters are displayed in the subpanels.

ORFs from both *BWF2A* and *SW3C* as well as CLOS1 were enriched in cluster VI, which was determined as the largest with 193 expression groups. CLOS1 in particular expressed many genes involved in hemicellulose deconstruction (e.g. GH3, GH5, GH10, GH29, GH31, GH43 and GH130) and carbohydrate deacetylation (e.g. CE1, CE4, CE7, CE9, CE12) (**Table S6**). Expressed cluster VI genes from *C. proteolyticus* -affiliated genomes/MAGs were also inferred in hemicellulose-degradation, including GH3, GH16, GH74, CE1, CE4, CE9 and CE10 (**Table S6**). In particular, the GH16 and GH3-encoding ORFs from region-A within *BWF2A* and *SW3C* were detected in cluster VI, which reaffirms our earlier predictions that certain *C. proteolyticus* populations in SEM1b are capable of degrading hemicellulosic substrates. The expression profile of cluster VI over time was observed to slightly lag after cluster IV (**Figure 4**), suggesting that hemicellulases in cluster VI genes are expressed once the hydrolytic effects of the RCLO1-cellulosome (expressed in cluster IV) have liberated hemicellulosic substrates (Zverlov *et al.,* 2005b). Although *C. thermocellum* cannot readily utilize other carbohydrates besides cellodextrins (Demain *et al.,* 2005), the cellulosome is composed of a number of hemicellulolytic enzymes such as GH10 endoxylanases, GH26 mannanases and GH74 xyloglucanases (Zverlov *et al.,* 2005a), which are involved in the deconstruction of the underlying cellulose-hemicellulose matrix (Zverlov *et al.,* 2005b). Representatives of GH10, GH26 and GH74 from RCLO1 were all expressed in cluster IV and are presumably acting on the hemicellulose fraction present in the spruce-derived cellulose (Chylenski *et al.,* 2017). Furthermore, detection of hydrolysis products (**Figure 4A**), revealed that xylose increased significantly at T5-7, indicating that hemicellulosic polymers containing beta-1-4-xylan were likely available at these stages. In addition to cluster VI, clusters V, VII and VIII also exhibited expression profiles that gradually increased after the initial peak of cluster IV. These clusters were all found to contain many enzymes putatively targeting medium and short length carbohydrate chains, including those derived from xylan sources (**Table S6**).

All in all, the SEM1b expression data shows sequential community progression that co-ordinates hydrolysis of cellulose and hemicellulose as well as carbohydrates that are found in the microbial cell wall. In particular, *C. proteolyticus* populations in SEM1b were suspected to play key roles degrading microbial cell wall carbohydrates as well as hemicellulosic substrates, possibly in cooperation or in parallel to other clostridium populations at the later stages of the SEM1b growth cycle. The detection and expression of *C. proteolyticus* -affiliated

GH-doc enzymes also raises intriguing questions regarding the possibility of multi-species cellulosomes and their potential role in saccharolytic consortia.

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 and time-resolved metatranscriptomics to describe a novel *C. proteolyticus* population that is comprised of closely related strains that have acquired sets of CAZymes via HGT and putatively evolved to incorporate a saccharolytic lifestyle. The co-expression patterns of *C. proteolyticus* CAZymes in clusters I and II 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 hemicellulases by *C. proteolyticus*, and their expression in cluster VI 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 S7**. All microbial genomes are publicly available on JGI under the analysis project numbers listed in **Table S7**.

ACKNOWLEDGEMENTS

We are grateful for support from The Research Council of Norway (FRIPRO program, PBP: 250479 / NorZymeD, VGHE: 221568), as well as the European Research Commission Starting Grant Fellowship (awarded to PBP; 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.

COMPETING INTERESTS

The authors declare there are no competing financial interests in relation to the work described.

REFERENCES

Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA (2011). BLAST Ring Image Generator (BRIG): Simple prokaryote genome comparisons. *BMC Genomics* **12**.

Álvarez C, Reyes-Sosa FM, Díez B (2016). Enzymatic hydrolysis of biomass from wood. *Microbial Biotechnology* **9:** 149-156.

Artzi L, Bayer EA, Moraïs S (2017). Cellulosomes: Bacterial nanomachines for dismantling plant polysaccharides. *Nature Reviews Microbiology* **15:** 83-95.

Bendall ML, Stevens SLR, Chan LK, Malfatti S, Schwientek P, Tremblay J *et al* (2016). Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME Journal* **10:** 1589-1601.

Biller SJ, Berube PM, Lindell D, Chisholm SW (2015). Prochlorococcus: The structure and function of collective diversity. *Nature Reviews Microbiology* **13:** 13-27.

Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114-2120.

Bradford MM (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry* **72:** 248-254.

Bray NL, Pimentel H, Melsted P, Pachter L (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34:** 525-527.

Bron PA, Van Baarlen P, Kleerebezem M (2012). Emerging molecular insights into the interaction between probiotics and the host intestinal mucosa. *Nature Reviews Microbiology* **10:** 66-78.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7:** 335-336.

Chen IMA, Markowitz VM, Chu K, Palaniappan K, Szeto E, Pillay M *et al* (2017). IMG/M: Integrated genome and metagenome comparative data analysis system. *Nucleic Acids Research* **45:** D507-D516.

Chylenski P, Petrović DM, Müller G, Dahlström M, Bengtsson O, Lersch M *et al* (2017). Enzymatic degradation of sulfite-pulped softwoods and the role of LPMOs. *Biotechnology for Biofuels* **10:** 1-13.

Dassa B, Borovok I, Ruimy-Israeli V, Lamed R, Flint HJ, Duncan SH *et al* (2014). Rumen cellulosomics: Divergent fiber-degrading strategies revealed by comparative genome-wide analysis of six ruminococcal strains. *PLoS ONE* **9**.

Demain AL, Newcomb M, Wu JHD, Demain AL, Newcomb M, Wu JHD (2005). Cellulase, Clostridia, and Ethanol. *Microbiology and Molecular Biology Reviews* **69:** 124-154.

Edgar RC (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32:** 1792-1797.

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460-2461.

Ellegaard KM, Engel P (2016). Beyond 16S rRNA community profiling: Intra-species diversity in the gut microbiota. *Frontiers in Microbiology* **7:** 1-16.

Etchebehere C, Pavan ME, Zorzópulos J, Soubes M, Muxí L (1998). Coprothermobacter platensis sp. nov., a new anaerobic proteolytic thermophilic bacterium isolated from an anaerobic mesophilic sludge. *International journal of systematic bacteriology* **48:** 1297-1304.

Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA (2011). Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME Journal* **5:** 461-472.

Gill SR, Fouts DE, Archer GL, Mongodin EF, Deboy RT, Ravel J *et al* (2005). 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* **187:** 2426-2438.

González-Torres P, Pryszcz LP, Santos F, Martínez-García M, Gabaldón T, Antón J (2015). Interactions between closely related bacterial strains are revealed by deep transcriptome sequencing. *Applied and Environmental Microbiology* **81:** 8445-8456.

Hagen LH, Frank JA, Zamanzadeh M, Eijsink VGH, Pope PB, Horn SJ *et al* (2017). Quantitative metaproteomics highlight the metabolic contributions of uncultured phylotypes in a thermophilic anaerobic digester. *Applied and Environmental Microbiology* **83**.

Hehemann JH, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G (2010). Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464:** 908-912.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ *et al* (2016). A new view of the tree of life. *Nature Microbiology* **1:** 1-6.

Hungate RE (1969). Chapter IV A Roll Tube Method for Cultivation of Strict Anaerobes. In: Norris JR, Ribbons DWBTMiM (eds). *Methods in Microbiology*. Academic Press. pp 117-132.

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF (2008). Resource Partitioning and Sympatric Differentiation Among Closely Related Bacterioplankton. *Science* **320:** 1081 LP-1085.

Johnson JW, Fisher JF, Mobashery S (2013). Bacterial cell wall recycling. *Annals of the new york academy* **1277:** 54-75.

Kahel-Raifer H, Jindou S, Bahari L, Nataf Y, Shoham Y, Bayer EA *et al* (2010). The unique set of putative membrane-associated anti-σ factors in Clostridium thermocellum suggests a novel extracellular carbohydrate-sensing mechanism involved in gene regulation. *FEMS Microbiology Letters* **308:** 84-93.

Kang DD, Froula J, Egan R, Wang Z (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3:** e1165-e1165.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A *et al* (2014). Single-Cell Genomics Reveals Hundreds of coexisting subpopulations in wild Prochlorococcus. *Science (New York, NY)* **344:** 416-420.

Kopylova E, Noé L, Touzet H (2012). SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28:** 3211-3217.

Koskella B, Vos M (2015). Adaptation in Natural Microbial Populations. *Annual Review of Ecology, Evolution, and Systematics* **46:** 503-522.

Kumar S, Stecher G, Tamura K (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular biology and evolution* **33:** 1870-1874.

Kunath BJ, Bremges A, Weimann A, McHardy AC, Pope PB (2017). Metagenomics and CAZyme Discovery. In: Abbott DW, Lammerts van Bueren A (eds). *Protein-Carbohydrate Interactions: Methods and Protocols*. Springer New York: New York, NY. pp 255-277.

Lamed R, Setter E, Bayer EA (1983). Characterization of a cellulose-binding, cellulase-containing complex in Clostridium thermocellum. *Journal of Bacteriology* **156:** 828-836.

Langfelder P, Zhang B, Horvath S (2008). Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **24:** 719-720.

Langmead (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods* **9:** 357-359.

Letunic I, Bork P (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research* **44:** W242-W245.

Li H (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arxiv* **00:** 1-3.

Lü F, Bize A, Guillot A, Monnet V, Madigou C, Chapleur O *et al* (2014). Metaproteomics of cellulose methanisation under thermophilic conditions reveals a surprisingly high proteolytic activity. *ISME Journal* **8:** 88-102.

Martin M (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17:** 10-10.

McLoughlin K, Schluter J, Rakoff-Nahoum S, Smith AL, Foster KR (2016). Host Selection of Microbiota via Differential Adhesion. *Cell Host and Microbe* **19:** 550-559.

Miller JH, Novak JT, Knocke WR, Pruden A (2016). Survival of antibiotic resistant bacteria and horizontal gene transfer control antibiotic resistance gene content in anaerobic digesters. *Frontiers in Microbiology* **7:** 1-11.

Modi SR, Lee HH, Spina CS, Collins JJ (2013). Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499:** 219-222.

Nataf Y, Bahari L, Kahel-Raifer H, Borovok I, Lamed R, Bayer EA *et al* (2010). Clostridium thermocellum cellulosomal genes are regulated by extracytoplasmic polysaccharides via alternative sigma factors. *Proceedings of the National Academy of Sciences* **107:** 18646-18651.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA (2017). MetaSPAdes: A new versatile metagenomic assembler. *Genome Research* **27:** 824-834.

Ochman H, Lawrence JG, Grolsman EA (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405:** 299-304.

Ollivier BM, Mah Ra, Ferguson TJ, Boone DR, Garcia JL, Robinson R (1985). Emendation of the Genus Thermobacteroides: Thermobacteroides proteolyticus sp. nov., a proteolytic acetogen from a methanogenic enrichment. *International Journal of Systematic Bacteriology* **35:** 425-428.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG *et al* (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31:** 3691-3693.

Pagès S, Bélaïch A, Bélaïch J-P, Morag E, Lamed R, Shoham Y *et al* (1997). Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: Prediction of specificity determinants of the dockerin domain. *Proteins: Structure, Function, and Bioinformatics* **29:** 517-527.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25:** 1043-1055.

Ricard G, McEwan NR, Dutilh BE, Jouany JP, Macheboeuf D, Mitsumori M *et al* (2006). Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics* **7:** 1-13.

Rodriguez-R LM, Konstantinidis KT (2016). The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints* **4:** e1900v1901.

Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F *et al* (2009). Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology* **7:** 828-828.

Rosenzweig RF, Sharp RR, Treves DS, Adams J (1994). Microbial evolution in a simple unstructured environment: Genetic differentiation in Escherichia coli. *Genetics* **137:** 903-917.

Rødsrud G, Lersch M, Sjöde A (2012). History and future of world's most advanced biorefinery in operation. *Biomass and Bioenergy* **46:** 46-59.

Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A *et al* (2013). Genomic variation landscape of the human gut microbiome. *Nature* **493:** 45-50.

Schumann P (1991). Nucleic Acid Techniques in Bacterial Systematics (Modern Microbiological Methods). *Journal of Basic Microbiology* **31:** 479-480.

Shapiro BJ, Timberlake SC, Szabó G, Polz MF, Alm EJ (2012). Population Genomics of Early Differentiation of Bacteria. *Science* **336:** 48-51.

Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Research* **23:** 111-120.

Siezen RJ, Tzeneva VA, Castioni A, Wels M, Phan HTK, Rademaker JLW *et al* (2010). Phenotypic and genomic diversity of Lactobacillus plantarum strains isolated from various environmental niches. *Environmental Microbiology* **12:** 758-773.

Solheim M, Aakra Å, Snipen LG, Brede DA, Nes IF (2009). Comparative genomics of Enterococcus faecalis from healthy Norwegian infants. *BMC Genomics* **10:** 1-11.

Song T, Xu H, Wei C, Jiang T, Qin S, Zhang W *et al* (2016). Horizontal Transfer of a Novel Soil Agarase Gene from Marine Bacteria to Soil Bacteria via Human Microbiota. *Scientific Reports* **6:** 1-10.

Spanogiannopoulos P, Bess EN, Carmody RN, Turnbaugh PJ (2016). The microbial pharmacists within us: A metagenomic view of xenobiotic metabolism. *Nature Reviews Microbiology* **14:** 273-287.

Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM (2015). rrnDB: Improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research* **43:** D593-D598.

Takahashi S, Tomita J, Nishioka K, Hisada T, Nishijima M (2014). Development of a prokaryotic universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing. *PLoS ONE* **9**.

Tandishabo K, Nakamura K, Umetsu K, Takamizawa K (2012). Distribution and role of Coprothermobacter spp. in anaerobic digesters. *Journal of Bioscience and Bioengineering* **114:** 518-520.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL *et al* (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences* **102:** 13950-13955.

Treangen TJ, Rocha EPC (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genetics* **7**.

Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N (2017). Microbial strain-level population structure & genetic diversity from metagenomes. *Genome Research* **27:** 626-638.

Turro E, Su SY, Gonçalves Â, Coin LJM, Richardson S, Lewin A (2011). Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology* **12:** 1-15.

Turro E, Astle WJ, Tavaré S (2014). Flexible analysis of RNA-seq data using mixed effects models. *Bioinformatics* **30:** 180-188.

Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y (2012). DbCAN: A web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* **40:** 445-451.

Zamanzadeh M, Hagen LH, Svensson K, Linjordet R, Horn SJ (2016). Anaerobic digestion of food waste - Effect of recirculation and temperature on performance and microbiology. *Water Research* **96:** 246-254.

Zelezniak A, Andrejev S, Ponomarova O, Mende DR, Bork P, Patil KR (2015). Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proceedings of the National Academy of Sciences* **112:** 6449-6454.

Zhou Y, Pope PB, Li S, Wen B, Tan F, Cheng S *et al* (2014). Omics-based interpretation of synergism in a soil-derived cellulose-degrading microbial community. *Scientific Reports* **4:** 1-6.

Zunino P, Piccini C, Legnani-Fajardo C (1994). Flagellate and non-flagellate Proteus mirabilis in the development of experimental urinary tract infection. *Microbial Pathogenesis* **16:** 379-385.

Zverlov VV, Kellermann J, Schwarz WH (2005a). Functional subgenomics of Clostridium thermocellum cellulosomal genes: Identification of the major catalytic components in the extracellular complex and detection of three new enzymes. *Proteomics* **5:** 3646-3653.

Zverlov VV, Schantz N, Schmitt-Kopplin P, Schwarz WH (2005b). Two new major subunits in the cellusome of Clostridium thermocellum: Xyloglucanase Xgh74A and endoxylanase Xyn10D. *Microbiology* **151:** 3395-3401.