Whole genome sequence of the anaerobic isosaccharinic acid degrading isolate, Macellibacteroides fermentans strain HH-ZS.

Simon P. Rout, Zohier B. Salah, Christopher J. Charles and Paul N. Humphreys*

Department of Biological Sciences, University of Huddersfield, Queensgate Campus, Huddersfield, United Kingdom, HD1 3DH.

***Corresponding author**: E-mail: p.n.humphreys@hud.ac.uk

Data deposition: The draft genome sequence of *Macellibacteroides fermentans* strain HH-ZS is deposited within BioProject PRJNA324192 under GenBank ID LZEK01000002.2

Keywords: Macellibacteroides fermentans, obligate anaerobe, alkalitolerant, whole genome sequence, Bacteroidetes, isosaccharinic acid.

Abstract

The ability of micro-organisms to degrade isosaccharinic acids (ISAs) whilst tolerating hyperalkaline conditions is pivotal to our understanding of the biogeochemistry associated within these environs, but also in scenarios pertaining to the cementitious disposal of radioactive wastes. An alkalitolerant, ISA degrading micro-organism was isolated from the hyperalkaline soils resulting from lime depositions. Here we report the first whole genome sequence, ISA degradation profile and carbohydrate preoteome of a *Macellibacteroides fermentans* strain HH-ZS, 4.08Mb in size, coding 3,241 proteins, 64 tRNA and one rRNA.

Introduction

Alkaline environments have consistently been studied for the isolation of novel alkaliphilic and alkalitolerant microbial species (Duckworth et al., 1996), with particular focus on biotechnological applications of the enzymes produced by these organisms (Mamo and Mattiasson, 2016). These alkaline environments are either natural, such as soda lakes or anthropogenic as a result of land contamination. One example of an anthropogenic alkaline environment can be found in Harpur Hill, Derbyshire, UK (Milodowski et al., 2015). Here, historical in-valley deposition of lime wastes generated through lime kiln workings has resulted in soil pH values of >11.0 following interactions between the wastes with rainwaters resulting in the generation of $Ca(OH)_2$ (Burke et al., 2012).

The conditions are somewhat analogous to those expected within the deep geological disposal concept for the UK radioactive waste inventory, where anaerobic, alkaline conditions dominate (NDA, 2010, Rout et al., 2015b) Under these conditions isosaccharinic acids (ISAs) are generated and are a key consideration for safety assessments due to their ability to mobilise radioelements such as plutonium (Greenfield et al., 1997) . Microbial degradation of ISAs is therefore a component of this analysis. To date, mixed cultures have demonstrated ISA degradation (Bassil et al., 2014, Kuippers et al., 2015, Rout et al., 2015a), despite this,

degradation by pure cultures has received limited attention and has been limited to historical descriptions of aerobic isolates (Pekarovičová and Mikulášová, 1991, Strand et al., 1984).

The isolation of an ISA degrading strain is therefore of significant interest and the availability of a whole genome sequence of such a strain would allow for the biochemistry of ISA degradation to be further investigated. A novel strain, designated HH-ZS, was isolated from the soil at Harpur Hill and identified via 16S rRNA profiling as *Macellibacteroides fermentans,* distinct from strain LIND7H (Jabari et al., 2012) , of the Phylum Bacteroidetes. This study describes the ISA degradation profile and draft whole genome sequence of this organism.

Materials and Methods

Culturing, Phenotypic and Genomic DNA Isolation

Hyperalkaline soil was obtained from Harpur Hill, Derbyshire and diluted in anoxic mineral media (B.S.I, 2005) with ISA containing cellulose degradation products (as per (Rout et al., 2014)) used as a carbon source. After anoxic incubation within this microcosm for 4 weeks at 25°C, 10µl of the reaction fluid was sub cultured onto fastidious anaerobic agar (LabM, UK) adjusted to pH 9 using 4M NaOH. Growth of single colonies were observed following incubation at 25°C in an anaerobic workstation ((10% H2:10% CO2: 80%N2; DW Scientific) for 3 days. Single colonies were further purified by sub-culture. Cell morphology was determined via Gram staining and SEM via a Quanta FEG 250 scanning electron microscope. Biochemical capabilities of the strain were determined using API 20A kit (bioMérieux, US), and through culture in mineral media at pH 9 containing 4mM $Ca(ISA)_2$, with ISA concentration and volatile fatty acid production monitored using HPAEC-PAD and GC as previously described (Rout et al., 2014). The pH profile of the isolate was determined via BioscreenC technology (Growth Curves USA, US) with optical density of the isolate measured at 450-600nm in fastidious anaerobic broth adjusted to a pH range of 4-12 in 1 pH unit

increments using 4M HCl or 4M NaOH accordingly. Replicates were prepared in 100 well plates which were then sealed using gas proof tape and incubated for 62 hours at 30°C. Genomic DNA was extracted from harvested cells using an UltraClean Microbial Isolation Kit. The concentration and purity of the extracted DNA was then assessed using a NanoDrop ND1000 spectrophotometer. This genomic DNA was then used to prepare a paired-end library.

Genome Sequencing, Assembly and Annotation

Paired end sequence reads were prepared using an Illumina HiSeq 2500 system (Illumina, US) with sequencing carried out by BaseClear, NL. FASTQ sequence files were generated using the Illumina Casava pipeline version 1.8.3. Initial quality assessment was based on data passing the Illumina Chastity filtering. PhiX control signal associated reads were removed using an in house filtering protocol. Reads containing (partial) adapter sequence were trimmed up to a minimum read length of 50bp. FASTQC quality control tool version 0.10.0 was used to provide a second quality assessment of the remaining reads. Low quality bases were trimmed to enhance the quality of FASTQ sequences using the 'trim sequences' option of CLC Genomics Workbench version 8.5.1. Contiguous sequences were then assembled using the 'De novo assembly' option of the CLC genomics workbench and mis-assemblies corrected using Pilon (Walker et al., 2014). Orientation of contigs within scaffolds was performed using SSPACE premium scaffolder v2.3 (Boetzer et al., 2011) and any gaps within these scaffolds were closed using GapFiller v1.10 (Boetzer and Pirovano, 2012). These gaps were then filled using Sanger sequencing technology (MWG-Eurofins, Germany), primers were designed using NCBI Primer-Blast (Ye et al., 2012) and can be seen in Table S1.

Phylogenetic, Genome and Proteome Analysis

In addition to whole genome sequence, a partial fragment of the 16S rRNA gene was amplified by PCR using universal primers 8F and 1512R primers, with the fragment purified using a Qiagen QIAquick PCR purification kit (Qiagen, US) prior to Sanger sequencing (Eurofins,

Germany). The partial sequence was used to identify the isolate using the EzTaxon-e database (Kim et al., 2012) then used for phylogenetic comparison with other members of the Bacteroidales. These sequences were extracted from the National Centre for Biotechnology Information (NCBI) and aligned using the MUSCLE module of MEGA 7.0.14 (Kumar et al., 2016) and a maximum likelihood tree generated in the same package (model: Tamura-Nei; bootstrap; 1000).

For initial annotations sequences were submitted to the NCBI Prokaryotic Genome Annotation Pipeline tools (Tatusova et al., 2013) then further annotated on the RAST server (Overbeek et al., 2014). COG category assignment was performed using the WebMGA server (Wu et al., 2011) and Pfam domain predictions carried out using the NCBI conserved domain database (CDD) (Marchler-Bauer et al., 2014). For the prediction of signal peptides and transmembrane domains, SignalP 4.1 server (Petersen et al., 2011) and the TMHMM server v. 2.0 (Krogh et al., 2001) were used respectively. Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) were found online using the CRISPRfinder tool searching against a CRISPR database (Grissa et al., 2007).

Results and Discussion

Macellibacteroides fermentans strain HH-ZS was a found to be a Gram negative rod, with rods being ~1µm in length (Figure S1). The isolate was aerotolerant, but only capable of growth in anaerobic conditions at pH 5.0-10.0 with an optimum of 8.0 (Figure S2). When grown in minimal medium with $Ca(ISA)_2$ as the sole carbon source, there was a subsequent generation of acetic acid, coupled to a reduction in pH to 8.0 and increase in OD at 620nm (Figure 1). This degradation presented a first order rate of 5.0×10^{-2} ($\pm$ 1.5×10^{-3}) day^{-1}, and is the first rate presented for any microbial species for the degradation of ISA. The isolate was also capable of fermenting a range of other sugars (Table S2). Inference of phylogeny using partial 16S rRNA sequencing placed strain HH-ZS within the Bacteroidales showing homology to

Macellibacteroides fermentans strain LIND7H (Figure S3), with a 99.89% similarity observed using EzTaxon-e identification. This coupled to the alkaliphilitolerant nature of the isolate marked it as being a novel strain of *Macellibacteorides fermentans.*

The assembly was based upon 2,264,150 reads yielding 524MB of data with an average phred quality score of 37.67 following Cassava and FastQC pipeline analysis. The final assembly consisted of 4,081,835 bp within 67 scaffolds, representing a coverage of 127x and GC content of 41.71% (Figure 2). Following assembly, 5 gaps were remaining, resulting in a draft of the genome sequence. The genome contained a total of 3,345 genes, of which 3,241 were protein coding, 69 were RNA coding and 35 were pseudogenes and one CRISPR repeat. The majority of protein-coding genes (82.3 %) were assigned a putative function, 2,318 were found to have Pfam domains, 757 to transmembrane helicases and 422 signal peptides. The distribution of genes into COGs functions shown in Table S3. COG assignment associated 219 genes with carbohydrate metabolism and transport and further annotation using the CAZyme analysis toolkit (Park et al., 2010) identified a further 69 genes and placed all of these into CAZY domains (Figure S4).

When compared with other available whole genome sequences available for *Parabacteroides* sp, *Macellibacteroides fermentans* HHZS was found to have a greater number of proteins associated with the glycoside hydrolase family. With respect to the organic material present in the soil, the genome also contained a number of carbohydrate binding module families, in particular CBM4, are calcium dependent and are implicated in the degradation of a number of cellulosic structures, with the exception of crystalline cellulose (Kataeva et al., 2001). This may be compensated by the presence of a predicted family-30 CBM, which has been shown to bind less available cellulose (Malburg et al., 1997, Raut et al., 2015).

Conclusions

Here we present the first genome of *Macellibacteroides. fermetans* (strain HHZS), isolated from hyperalkaline soil, and demonstrating a wider pH tolerance range than previously described strain LIND7H and other closely related *Parabacteroides* sp.. The isolate is also the first obligate anaerobe described as a degrader of ISA, and here we present the first rates for its degradation. CAZy analysis revealed that the strain harbored a number of carbohydrate degrading enzymes, which merit further investigation to determine the metabolic pathways associated with ISA degradation.

References

B.S.I 2005. BS ISO 14853:2005 Plastics-Determination of the ultimate anaerobic biodegradation of plastic materials in an aqueous system- Method by measurement of biogas production. BS ISO 14853:2005, British Standards Institute, London, UK.

BASSIL, N. M., BRYAN, N. & LLOYD, J. R. 2014. Microbial degradation of isosaccharinic acid at high pH. *The ISME Journal,* 9**,** 310-320.

BOETZER, M., HENKEL, C. V., JANSEN, H. J., BUTLER, D. & PIROVANO, W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics,* 27**,** 578-579.

BOETZER, M. & PIROVANO, W. 2012. Toward almost closed genomes with GapFiller. *Genome Biology,* 13**,** R56.

BURKE, I. T., MORTIMER, R. J., PALANIYANDI, S., WHITTLESTON, R. A., LOCKWOOD, C. L., ASHLEY, D. J. & STEWART, D. I. 2012. Biogeochemical reduction processes in a hyper-alkaline leachate affected soil profile. *Geomicrobiology Journal,* 29**,** 769-779.

DUCKWORTH, A., GRANT, W., JONES, B. & VAN STEENBERGEN, R. 1996. Phylogenetic diversity of soda lake alkaliphiles. *FEMS Microbiology Ecology,* 19**,** 181-191.

GREENFIELD, B., HURDUS, M., SPINDLER, M. & THOMASON, H. 1997. The effects of the products from the anaerobic degradation of cellulose on the solubility and sorption of radioelements in the near field. *In:* PLC, A. T. (ed.). Oxford: AEA Technology.

GRISSA, I., VERGNAUD, G. & POURCEL, C. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research,* 35**,** W52-W57.

JABARI, L., GANNOUN, H., CAYOL, J.-L., HEDI, A., SAKAMOTO, M., FALSEN, E., OHKUMA, M., HAMDI, M., FAUQUE, G. & OLLIVIER, B. 2012. *Macellibacteroides fermentans* gen. nov., sp. nov., a member of the family Porphyromonadaceae isolated from an upflow anaerobic filter treating abattoir wastewaters. *International Journal of Systematic and Evolutionary Microbiology,* 62**,** 2522-2527.

KATAEVA, I. A., SEIDEL, R. D., LI, X.-L. & LJUNGDAHL, L. G. 2001. Properties and Mutation Analysis of the CelK Cellulose-Binding Domain from the *Clostridium thermocellum* Cellulosome. *Journal of Bacteriology,* 183**,** 1552-1559.

KIM, O.-S., CHO, Y.-J., LEE, K., YOON, S.-H., KIM, M., NA, H., PARK, S.-C., JEON, Y. S., LEE, J.-H., YI, H., WON, S. & CHUN, J. 2012. Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *International Journal of Systematic and Evolutionary Microbiology,* 62**,** 716-721.

KROGH, A., LARSSON, B., VON HEIJNE, G. & SONNHAMMER, E. L. L. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology,* 305**,** 567-580.

KUIPPERS, G., BASSIL, N. M., BOOTHMAN, C., BRYAN, N. & LLOYD, J. R. 2015. Microbial degradation of isosaccharinic acid under conditions representative for the far field of radioactive waste disposal facilities. *Mineralogical Magazine,* 79**,** 1443-1454.

KUMAR, S., STECHER, G. & TAMURA, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution***,** msw054.

MALBURG, S. R., MALBURG, L. M., LIU, T., IYO, A. H. & FORSBERG, C. W. 1997. Catalytic properties of the cellulose-binding endoglucanase F from *Fibrobacter succinogenes* S85. *Applied and Environmental Microbiology,* 63**,** 2449-2453.

MAMO, G. & MATTIASSON, B. 2016. Alkaliphilic Microorganisms in Biotechnology. *Biotechnology of Extremophiles.* Springer.

MARCHLER-BAUER, A., DERBYSHIRE, M. K., GONZALES, N. R., LU, S., CHITSAZ, F., GEER, L. Y., GEER, R. C., HE, J., GWADZ, M. & HURWITZ, D. I. 2014. CDD: NCBI's conserved domain database. *Nucleic Acids Research***,** gku1221.

MILODOWSKI, A. E., SHAW, R. P. & STEWART, D. I. 2015. The Harpur Hill Site: its geology, evolutionary history and a catalogue of materials present. . British Geological Survey, Keyworth, Nottingham, UK.

NDA 2010. Near-field Evolution Status Report Harwell, Didcot, Oxfordshire, UK Nuclear Decommissioning Authority.

OVERBEEK, R., OLSON, R., PUSCH, G. D., OLSEN, G. J., DAVIS, J. J., DISZ, T., EDWARDS, R. A., GERDES, S., PARRELLO, B. & SHUKLA, M. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research,* 42**,** D206-D214.

PARK, B. H., KARPINETS, T. V., SYED, M. H., LEUZE, M. R. & UBERBACHER, E. C. 2010. CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology,* 20**,** 1574-1584.

PEKAROVIČOVÁ, A. & MIKULÁŠOVÁ, M. 1991. Biodegradation of black liquor hydroxyacids by *Micrococcus lylae. Journal of Chemical Technology and Biotechnology,* 52**,** 539-543.

PETERSEN, T. N., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods,* 8**,** 785-786.

RAUT, M. P., KARUNAKARAN, E., MUKHERJEE, J., BIGGS, C. A. & WRIGHT, P. C. 2015. Influence of Substrates on the Surface Characteristics and Membrane Proteome of *Fibrobacter succinogenes* S85. *PLOS ONE,* 10**,** e0141197.

ROUT, S. P., CHARLES, C. J., DOULGERIS, C., MCCARTHY, A. J., ROOKS, D. J., LOUGHNANE, J. P., LAWS, A. P. & HUMPHREYS, P. N. 2015a. Anoxic Biodegradation of Isosaccharinic Acids at Alkaline pH by Natural Microbial Communities. *PLOS ONE,* 10**,** e0137682.

ROUT, S. P., CHARLES, C. J., GARRATT, E. J., LAWS, A. P., GUNN, J. & HUMPHREYS, P. N. 2015b. Evidence of the generation of isosaccharinic acids and their subsequent degradation by local microbial consortia within hyper-alkaline contaminated soils, with relevance to intermediate level radioactive waste disposal. *PLOS ONE,* 10.

ROUT, S. P., RADFORD, J., LAWS, A. P., SWEENEY, F., ELMEKAWY, A., GILLIE, L. J. & HUMPHREYS, P. N. 2014. Biodegradation of the Alkaline Cellulose Degradation Products Generated during Radioactive Waste Disposal. *PLOS ONE,* 9**,** e107433.

STRAND, S. E., DYKES, J. & CHIANG, V. 1984. Aerobic Microbial Degradation of Glucoisosaccharinic Acid. *Applied and Environmental Microbiology,* 47**,** 268-271.

TAMURA, K. & NEI, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution,* 10**,** 512-526.

TATUSOVA, T., DICUCCIO, M., BADRETDIN, A., CHETVERNIN, V., CIUFO, S. & LI, W. 2013. *Prokaryotic genome annotation pipeline,* NCBI, National Center for Biotechnology Information (US), Bethesda (MD).

WALKER, B. J., ABEEL, T., SHEA, T., PRIEST, M., ABOUELLIEL, A., SAKTHIKUMAR, S., CUOMO, C. A., ZENG, Q., WORTMAN, J. & YOUNG, S. K. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE,* 9**,** e112963.

WU, S., ZHU, Z., FU, L., NIU, B. & LI, W. 2011. WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC genomics,* 12**,** 1.

YE, J., COULOURIS, G., ZARETSKAYA, I., CUTCUTACHE, I., ROZEN, S. & MADDEN, T. L. 2012. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC bioinformatics,* 13**,** 134.

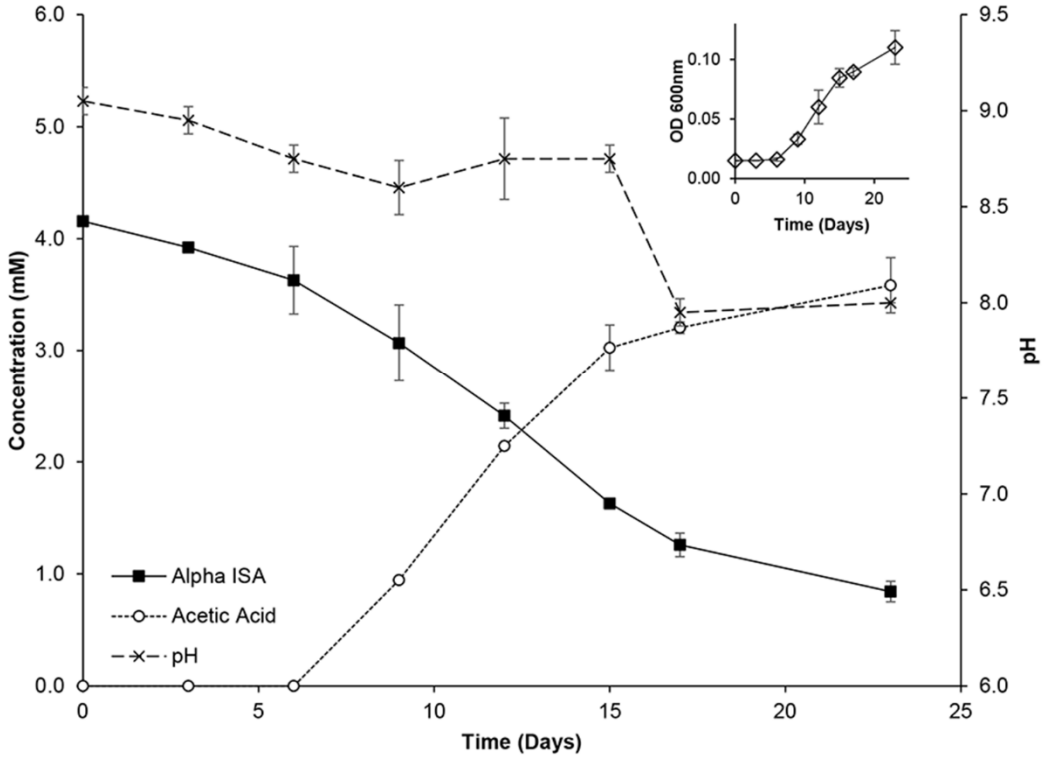

Fig 1. Isosaccharinic acid degradation profile of *Macellibacteroides fermentans* strain HH-ZS.

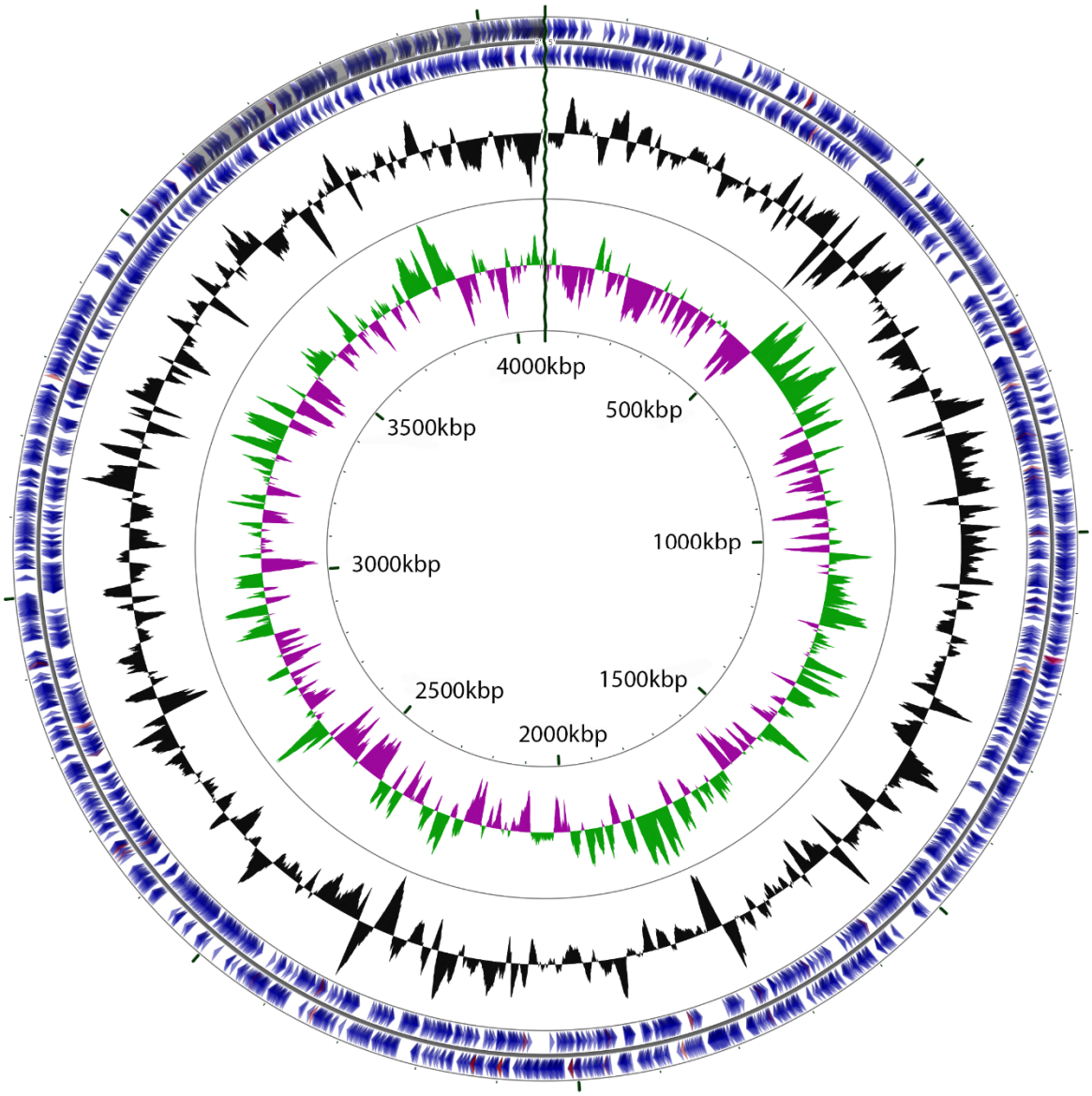

Fig 2. Circular representation of the *Macellibacteroides fermentans* strain HH-ZS complete genome. **Circles** (from inside to outside) **1 and 2** (protein coding sequences on the forward and reverse strand, blue indicates CDS, peach shows tRNAs, Pink rRNAs and grey other). **3** Shows GC content % and **4 and 5** show positive (green) and negative (magenta) GC skew.

Supplementary data

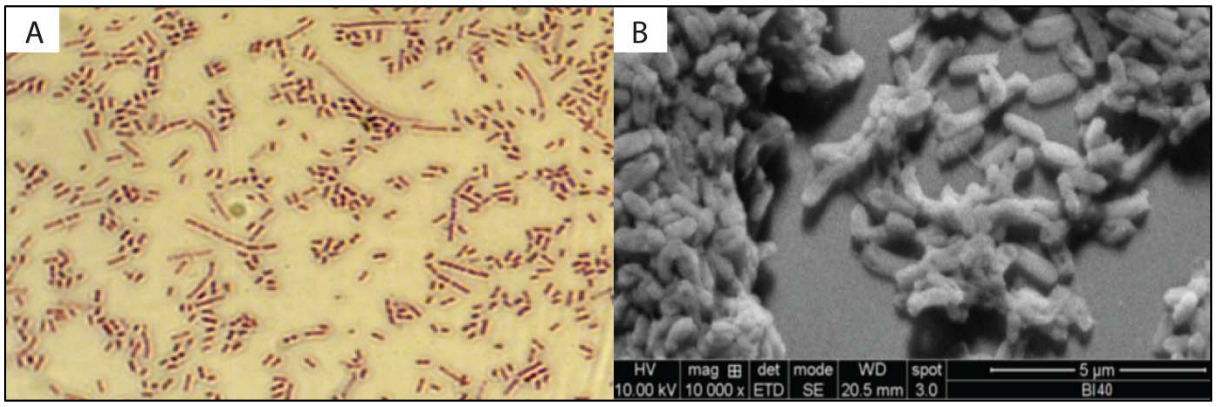

Fig S1: Gram stain (A) of *Macellibacteroides fermentans* strain HHZS viewed at 1000X magnification and SEM of the strain (B)

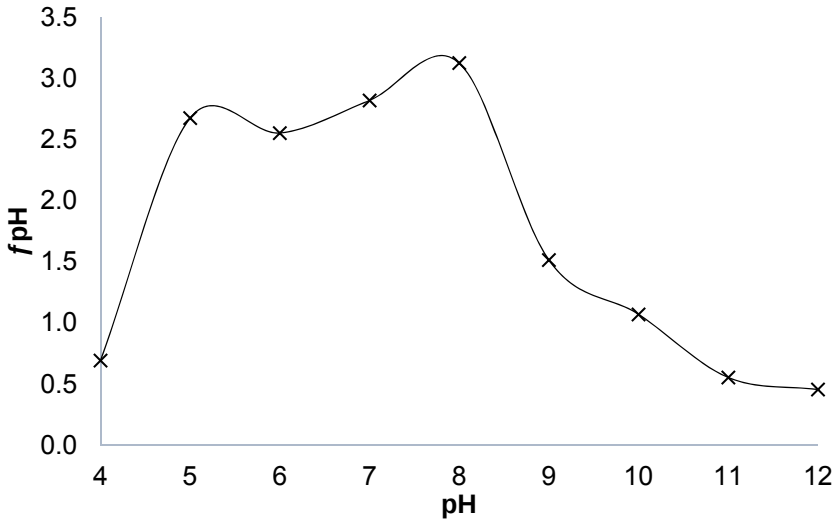

Fig S2. pH profile of *M.fermentans* HHZS^T determined via Bioscreen-C technology.

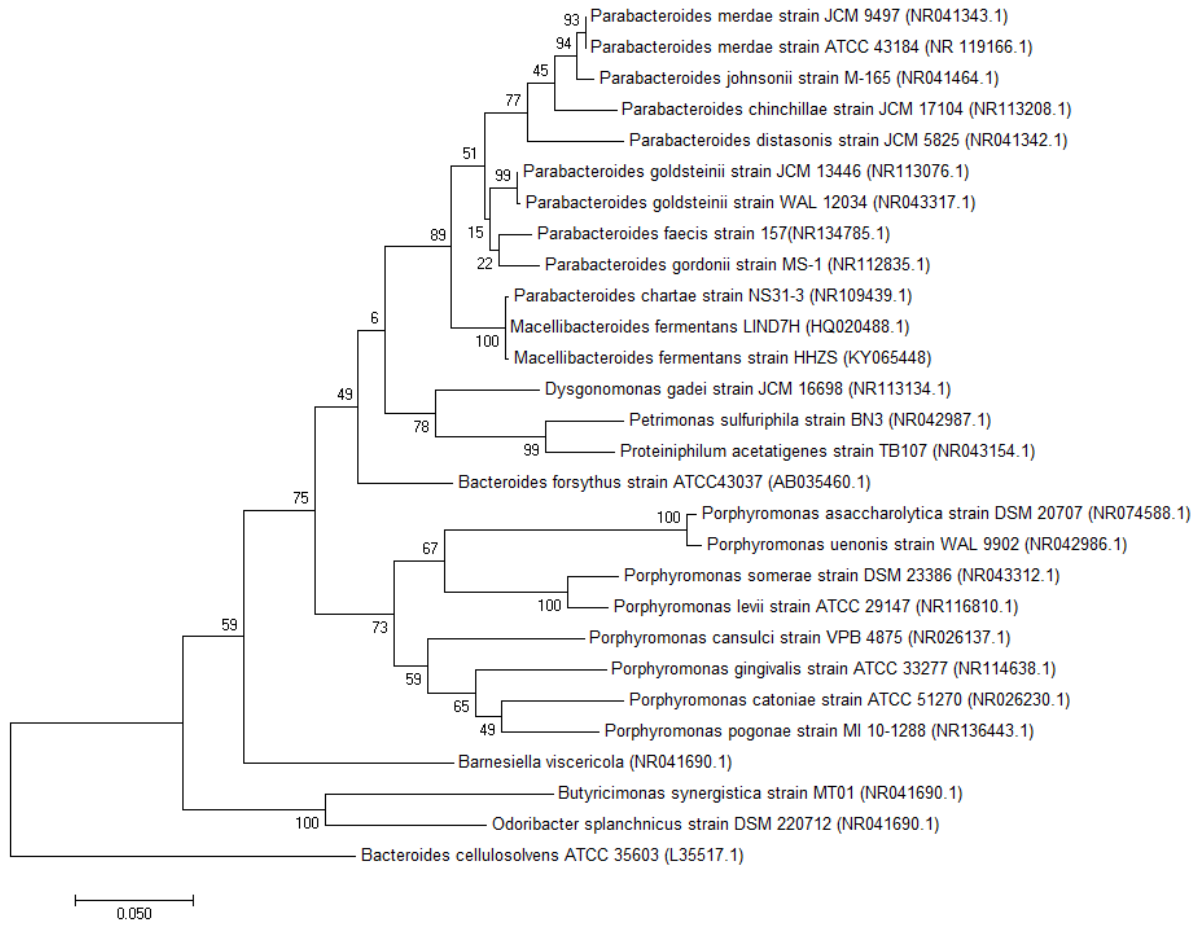

Fig S3. The evolutionary history of *M.fermentans* HHZS was inferred by using the Maximum Likelihood method based on the Tamura-Nei model (Tamura and Nei, 1993). The tree with the highest log likelihood (-6808.0249) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site (5 per 100 sites). The analysis involved 28 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 790 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016).

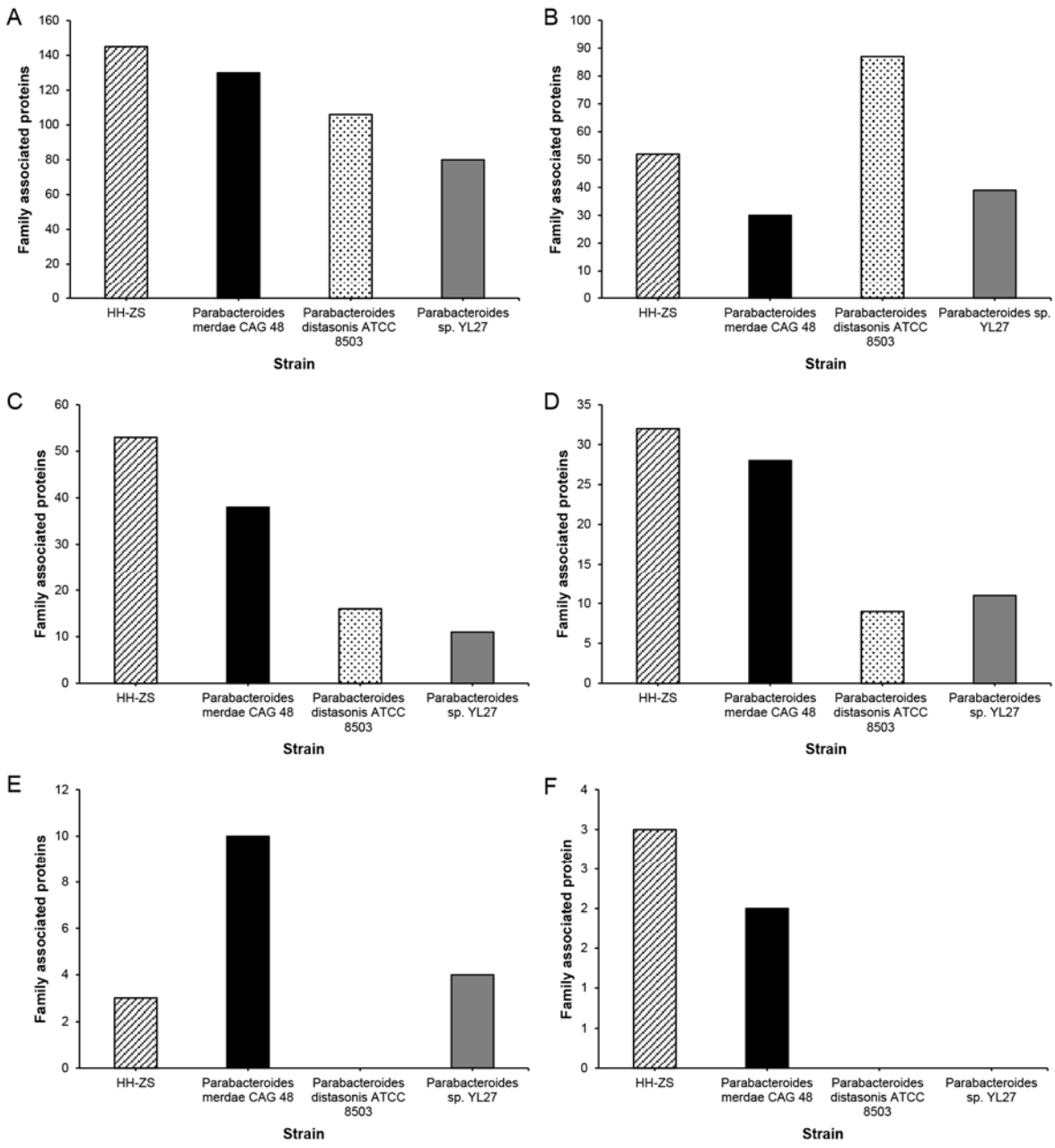

Fig S4: CAZy assignment of carbohydrate active enzymes of Strain HH-ZS and comparison with available whole genome sequences of Parabacteroides sp. Glycoside Hydrolase (A), Glycosyl Transferase (B), Carbohydrate Binding Module (C), Carbohydrate Esterase (D), Polysaccharide Lyase (E) and Auxiliary Activities (F) associated families are shown.

Gap containing region	Sequence (5'-3')	
1	F	AAC GCG GTA AGG CAG GTA AA
	R	TTG GCT CCG TCT GAT GAA CC
2	F	CCT GAC GGA AGA TCT GAG CC
	R	GCC TAC AGA GTT GCC GGA AT
3	F	AGC TGC CTG ATT TAT GTG TTG T
	R	CCA AAC AGT TGT GTT GAA TCT GC
4	F	GAT ACC GAG TTG ACG TGC CT
	R	AAC GGG TGT AAG TCC CCA AC
5	F	AGG AGC ACA TGC GGA AAG AT
	R	CGT CCA ATG TTT CCC ACC GA

Table S1. Primers used for completion of whole genome

Test	Result
Oxidase	-
Catalase	+
Indol	-
Urease	-
Protease (gelatine)	-
B-glucosidase (Esculin)	+
H_2S	+
D-Glucose	+
D-Mannitol	+
D-Lactose	+
D-Sucrose	+
D-Maltose	+
Salicin	+
D-xylose	+
L-arabinose	+
Glycerol	+
D-cellobiose	+
D-mannose	+
D-melezitose	+
D-raffinose	+
D-sorbitol	+
L-rhamnose	+
D-trehalose	+

Table S2. Results of biochemical testing performed as part of the API20A bacterial identification kit.

Code	Value	%age	Description
J	155	4.60	Translation, ribosomal structure and biogenesis
A	0	0.00	RNA processing and modification
K	185	5.49	Transcription
L	159	4.72	Replication, recombination and repair
B	0	0.00	Chromatin structure and dynamics
D	31	0.91	Cell cycle control, Cell division, chromosome partitioning
V	54	1.60	Defense mechanisms
T	140	4.15	Signal transduction mechanisms
M	245	7.27	Cell wall/membrane biogenesis
N	18	0.53	Cell motility
U	71	2.11	Intracellular trafficking and secretion
O	86	2.55	Posttranslational modification, protein turnover, chaperones
C	151	4.48	Energy production and conversion
G	219	6.50	Carbohydrate transport and metabolism
E	175	5.19	Amino acid transport and metabolism
F	74	2.20	Nucleotide transport and metabolism
H	141	4.18	Coenzyme transport and metabolism
I	73	2.17	Lipid transport and metabolism
P	187	5.55	Inorganic ion transport and metabolism
Q	21	0.62	Secondary metabolites biosynthesis, transport and catabolism
R	304	9.02	General function prediction only
S	179	5.31	Function unknown
-	702	20.83	Not in COGs

Table S3. Functional assignment of genes to COGs.