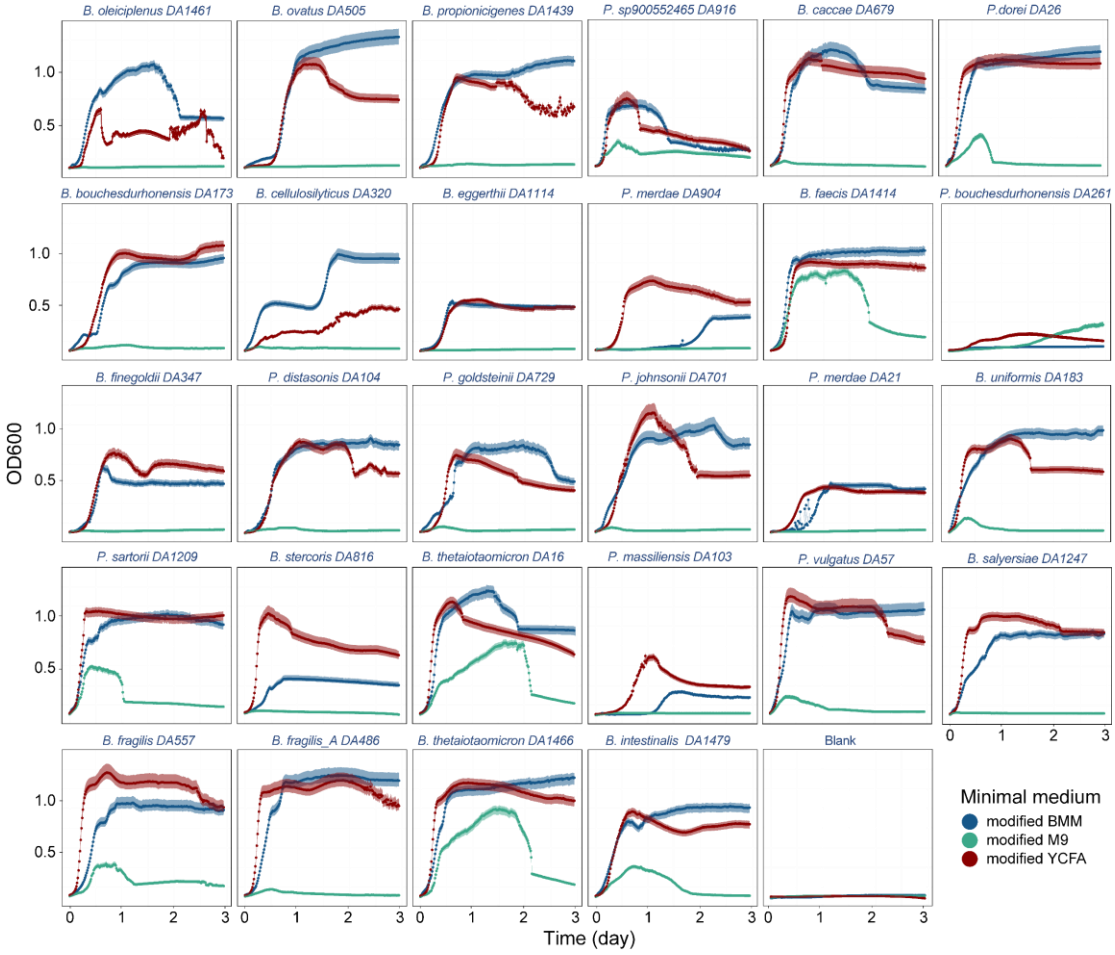
Selective utilization of medicinal polysaccharides by human gut *Bacteroides* species

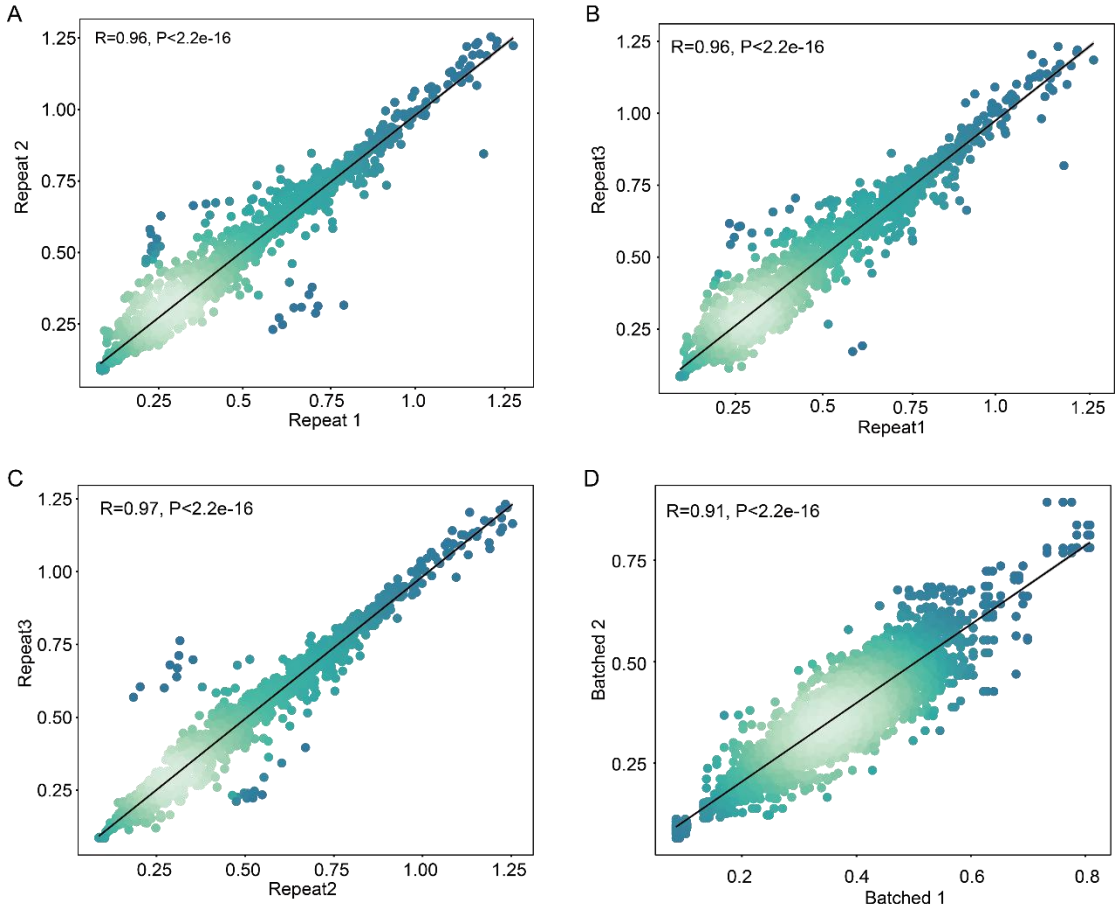
Zepeng Qu[1, 2, #], Hongbin Liu[2, #], Ji Yang[3, #], Linggang Zheng[1, 2], Jumin Huang[4], Ziming Wang[4], Chun Xie[4], Wenlong Zuo[2], Xiong, Xia[2], Lin Sun[5], Yifa Zhou[5], Ying Xie[6], Jingguang Lu[1], Yizhun Zhu[1], Lili Yu[1], Lihua Liu[7], Hua Zhou[6, *], Lei Dai[2, *], Elaine Lai-Han Leung[4, 8, *]

[1]School of Pharmacy, Faculty of Medicine & State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Taipa, Macau, China, 999078.
[2]CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, 518055.
[3]School of Chinese Materia Medica, Guangdong Pharmaceutical University, Guangzhou, China, 510006.
[4]Cancer Center, Faculty of Health Sciences, Ministry of Education (MOE) Frontiers Science Center for Precision Oncology, University of Macau, Macau, Macau SAR, China.
[5]Jilin Provincial Key Laboratory of Chemistry and Biology of Changbai Mountain Natural Drugs, Northeast Normal University, Changchun, China, 130024.
[6]State Key Laboratory of Traditional Chinese Medicine Syndrome, Guangdong Provincial Hospital of Chinese Medicine, Guangdong Provincial Academy of Chinese Medical Sciences, The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, Guangdong, China, 510006.
[7]School of Economics and Management, Yanbian University, Yanji, China,133000
[8]State Key Laboratory of Quality Research in Chinese Medicine, University of Macau, Macau.
[#]These authors contributed equally.
*Correspondence: lhleung@um.edu.mo (Elaine Leung); lei.dai@siat.ac.cn (Lei Dai); gutcmzhs@hotmail.com (Hua Zhou)

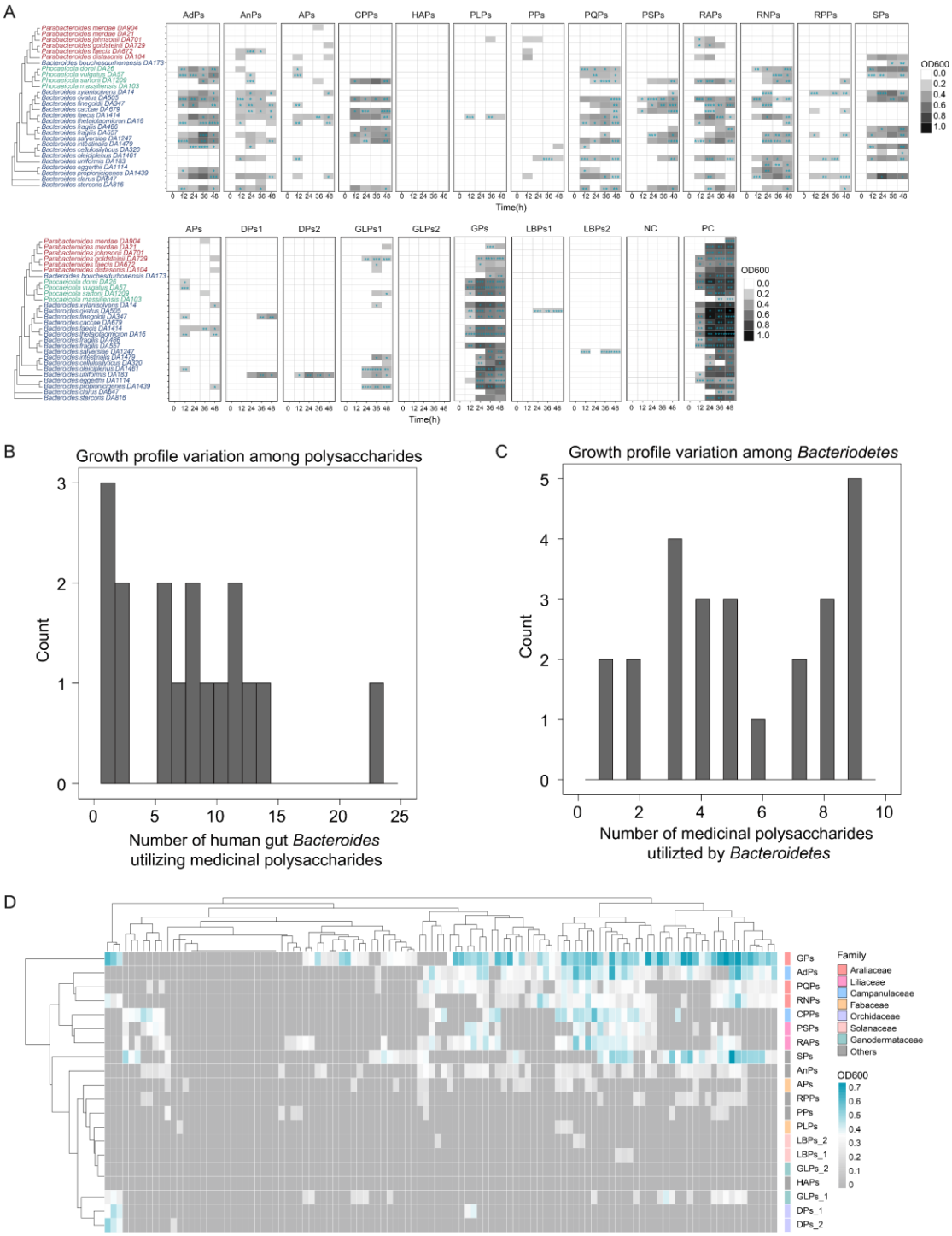
Supplementary Figure



Supplementary Figure 1 The mBMM medium with clear composition generally supported the growth of the *Bacteroidetes* spp. The growth curve of *Bacteroides* and *Parabacteroides* species in three basic medium, modified BMM, modified M9, and modified YCFA. Data are presented as mean values ± SD. Assays were performed in technical triplicates (n = 3).

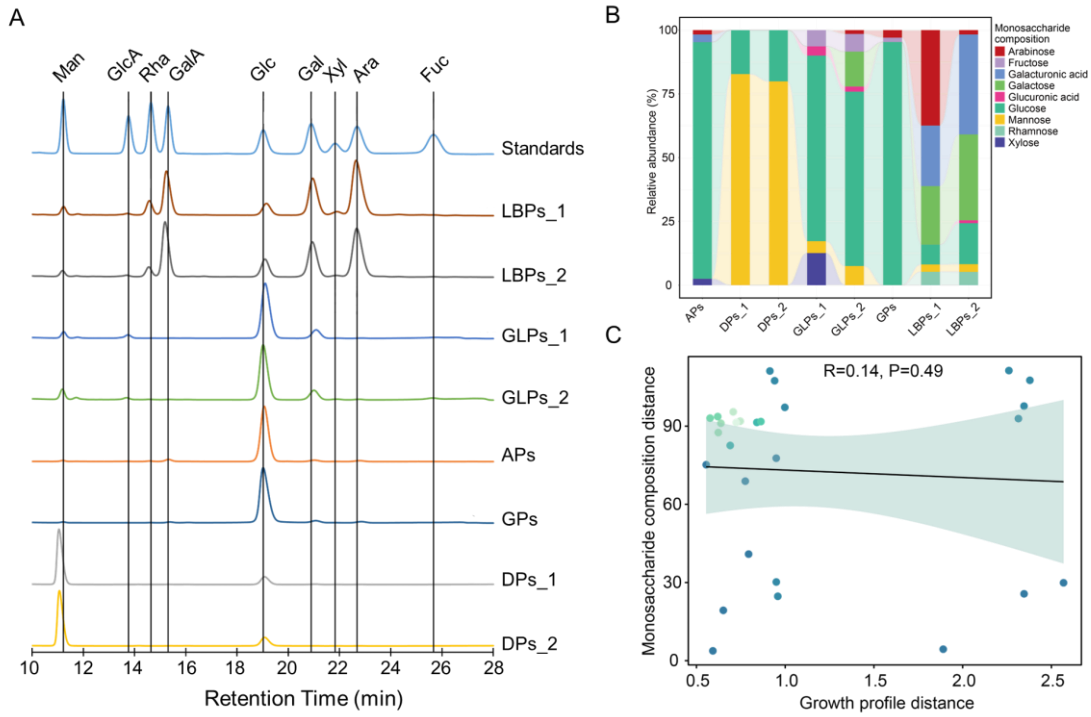


Supplementary Figure 2 The technical and batched repeatability of growth profiles. A-D) The point diagram displays the correlation between technical repeats 1 and 2 **(A)**, technical repeats 1 and 3 **(B)**, technical repeats 2 and 3 **(C)**, batched 1 and 2 **(D)**. The R and P values refer to the most parsimonious model, black line shows the mean regression based on all the data points. Light green dots represented high dot density and dark green dots represented low dot density.

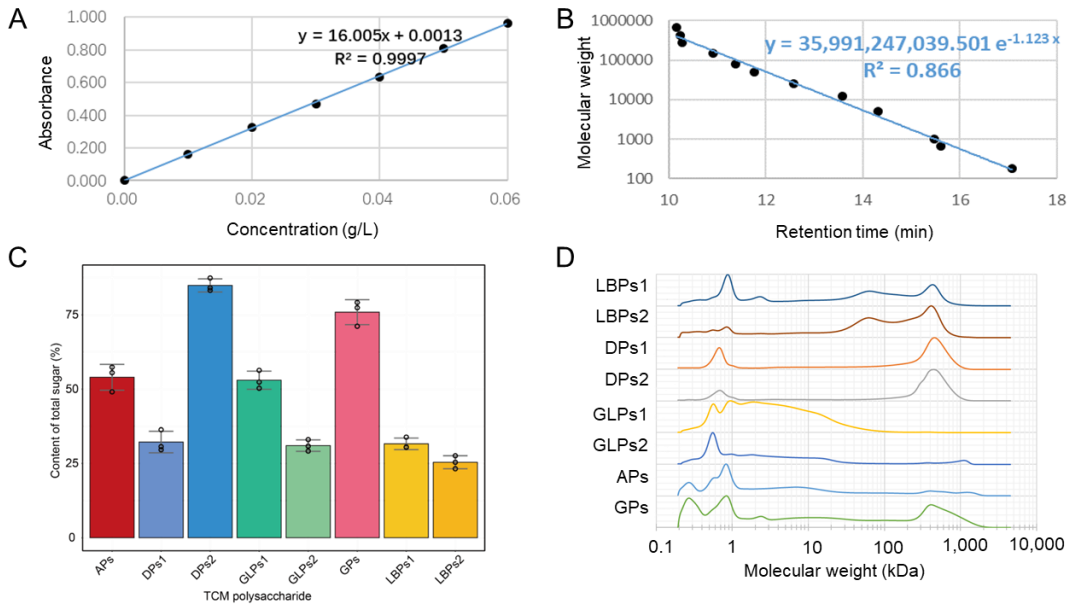


Supplementary Figure 3 The polysaccharides utilization spectrum of medicinal polysaccharides. A) The heat map showed the growth OD600 of 28 *Bacteroidetes* species at 0h, 12h, 24h, 36h, 48h under 20 medicinal polysaccharides. RNPs: Radix Notoginseng Polysaccharides; GPs: Ginseng polysaccharides; PQPs: Panax Quinquefolium Polysaccharides; PSPs: Polygonatum sibiricum Polysaccharides; RAPs: Rhizoma anemarrhenae Polysaccharides; CPPs: Codonopsis Pilosula Polysaccharides; AdPs: Adenophora Polysaccharides; PLPs: Pueraria lobata Polysaccharides; APs:

Astragalus polysaccharide; DPs_1 and DPs_2: Dendrobium polysaccharides; GLPs_1 and GLPs_2: Ganoderma lucidum polysaccharides; LBPs_1 and LBPs_2: Lycium barbarum polysaccharides; PPs: Poria Polysaccharides; RPPs: Radix Pseudoxtellariae Polysaccharides; AnPs: Angelica Polysaccharides; HAPs: Hovenia Acerba Polysaccharides; SPs: Scrophulariaceae Polysaccharide; PC: positive control; NC: negative control. Statistical significance was determined using a two-sided unpaired t-test between each experimental group and the negative control group, with adjustments for multiple comparisons made using the Bonferroni method. The P values are indicated as follows: * < 0.05; ** < 0.01; *** < 0.001; **** < 0.0001. Each experiment was conducted in triplicate to ensure reliability. **B)** Histogram showing distribution of the number of *Bacteroidetes* spp. utilizing medicinal polysaccharides. **C)** Histogram showing distribution of the number of medicinal polysaccharides utilized by *Bacteroidetes* spp. **D)** The heat map showed the growth phenotype, and the medicinal polysaccharides were clustered by Euclidean distance based on polysaccharide utilization spectrum. Each row is a high-dimensional vector composed of multiple *Bacteroidetes* species of a specific medicinal polysaccharides at multiple time points, and we calculate the Euclidean distance for each high-dimensional vector. medicinal polysaccharides were clustered based on the distance relations.

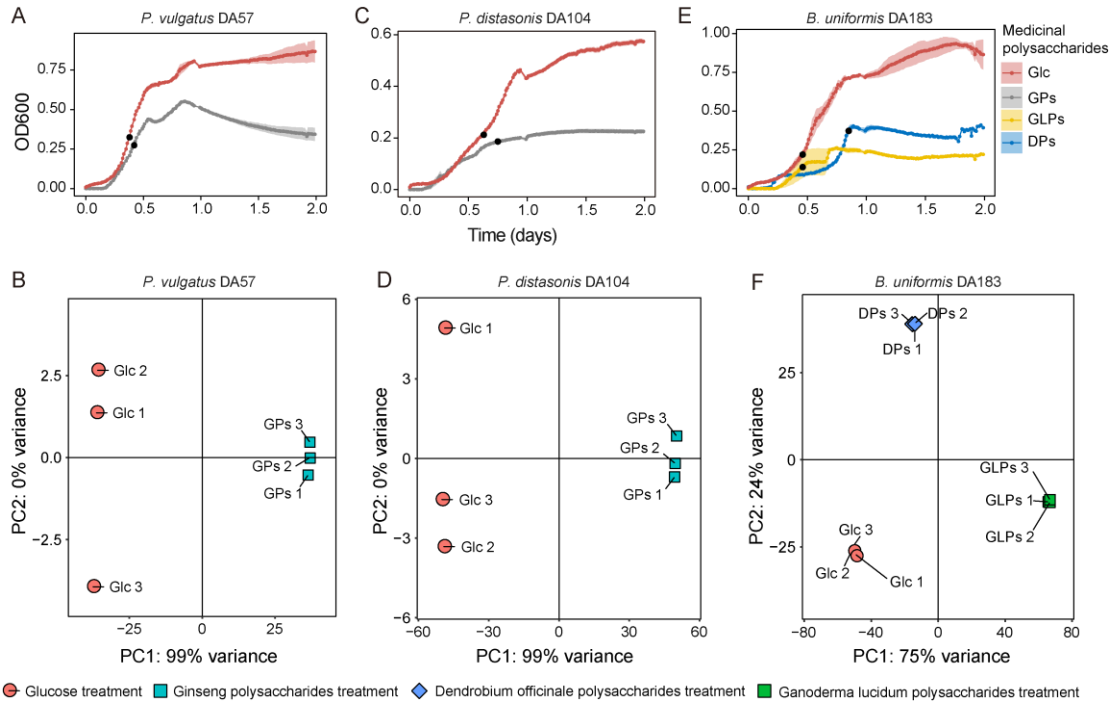


Supplementary Figure 4 Monosaccharide composition variation do not explain growth variation among medicinal polysaccharides. A) The HPLC peak diagram showed the monosaccharide composition. Ara: Arabinose; Fuc: Fructose; Gal: Galactose; GalA: Galacturonic acid; Glc: Glucose; GlcA: Glucuronic acid; Man: Mannose; Rha: Rhamnose; Xyl: Xylose. **B)** The stacked bar chart showed the composition and relative abundance of monosaccharide. **C)** The point diagram displays the correlation between growth profile distance and monosaccharide composition distance. The R and *P* values refer to the most parsimonious model. Statistical significance was determined using a two-sided Pearson correlation test (for R value) and an ordinary least squares (OLS) regression analysis (for the linear fit). The figure also includes a 95% confidence interval around the best linear fit. Light green dots represented high dot density and dark green dots represented low dot density.

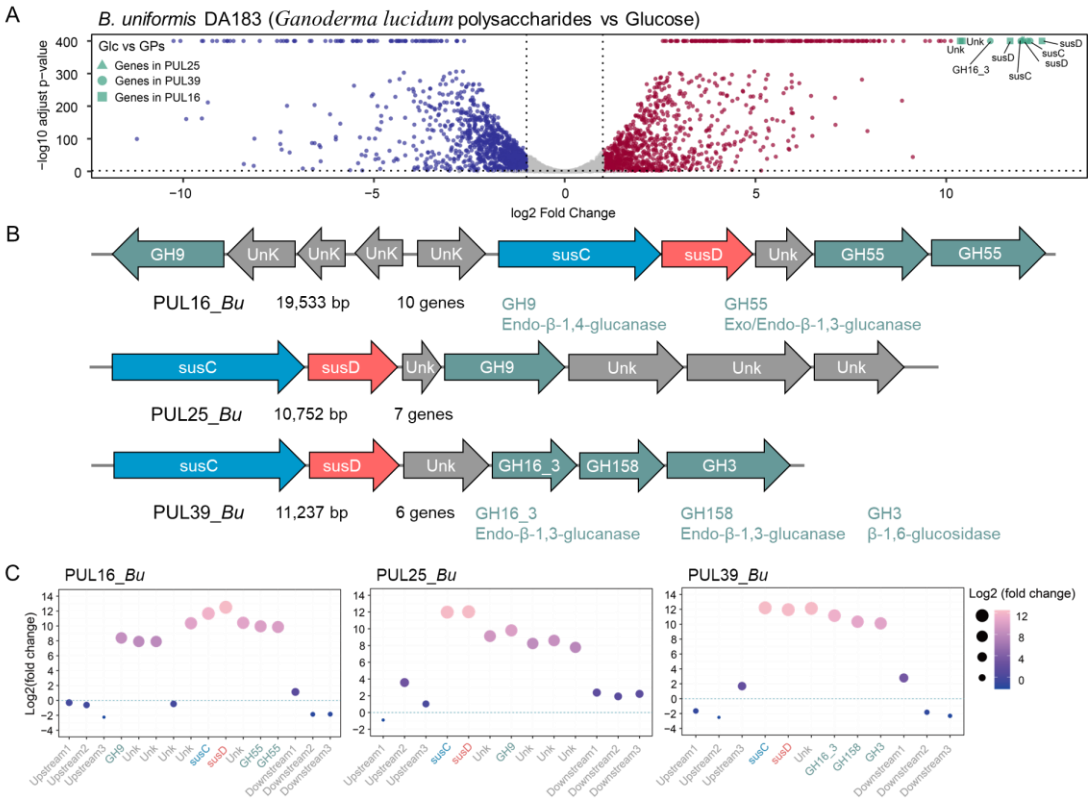


Supplementary Figure 5 The physical and chemical characterization of herbal polysaccharides. A) Standard curve for determination of total sugar content. Linear regression analysis is used to determine the slope, intercept, and equation. $R^2 = 0.9997$. **B)** Standard curve of molecular weight and chromatographic retention time of medicinal polysaccharides. $R^2 = 0.866$. **C)** The bar chart showed the content of total sugar. The content measurement of total sugar was repeated three times for each medicinal polysaccharide (black dots). Data are presented as mean values ± SD. Assays were performed in technical triplicates (n = 3). **D)** The stacked bar chart showed the composition and relative abundance of monosaccharide.

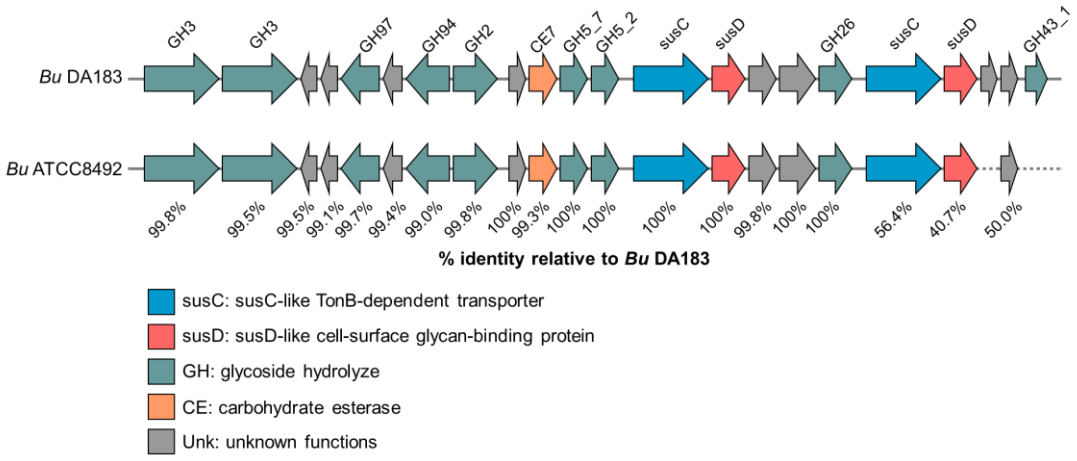
The total sugar content of APs was 54%, DPs_1 (32.2%), DPs_2 (84.9%), GLPs_1 (53%), GLPs_2 (31%), DPs_1 (31%), GPs (75.9%), LBPs_1 (31.6%), and LBPs_2 (25.4%) (Supplementary Fig. 4A, B, and Supplementary Table 5). The molecular weight analysis (Supplementary Fig. 4C, D, and Supplementary Table 6) showed that APs, GLPs-1, and GLPs-2 mainly consisted of polysaccharides with low molecular weight (0.4~1 kDa). GPs, DPs_1, DPs_2, LBPs_1, and LBPs_2 contained both low molecular weight (0.4~1 kDa) and high molecular weight (300~900 kDa) polysaccharides.



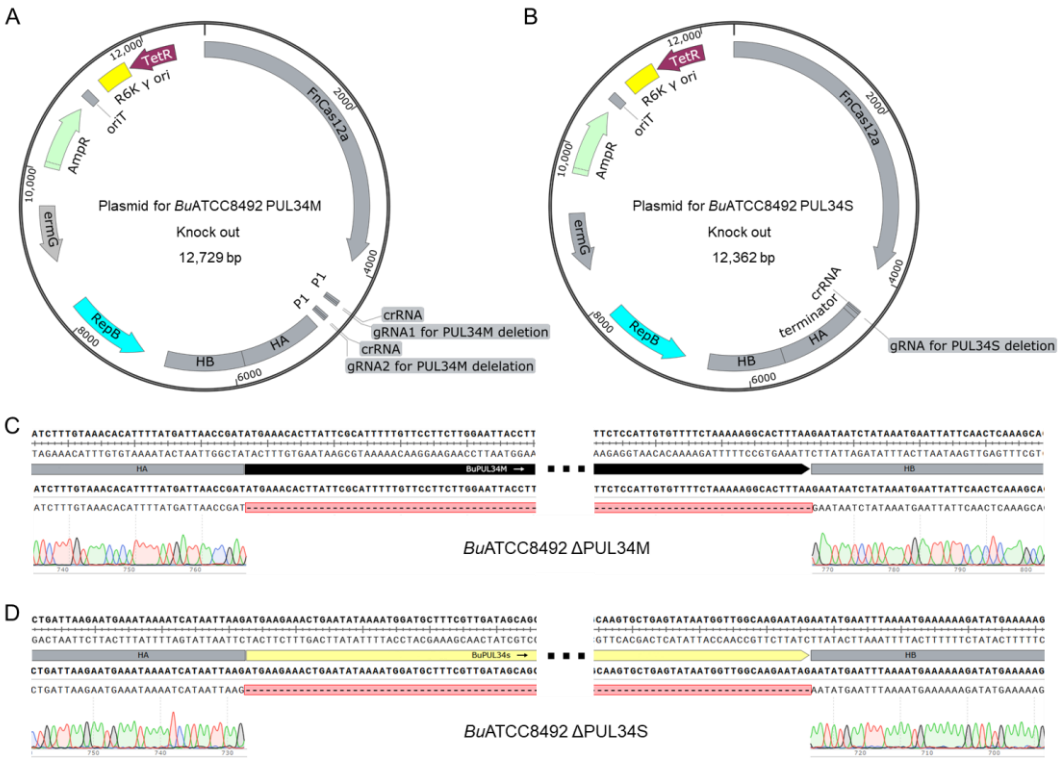
Supplementary Figure 6 Transcriptome sequencing and analysis of *Bacteroidetes* species. A, C, E) The scatter plot shows the sampling time point of transcriptome sequencing of *P. vulgatus* DA57 (A), *P. distasonis* DA104 (C), and *B. uniformis* DA183 (E). Data are presented as mean values ± SD. Assays were performed in technical triplicates (n = 3). **B, D, F)** The PCA diagram showed technical repeatability of transcriptome sequencing of *P. vulgatus* DA57 (B), *P. distasonis* DA104 (D), and *B. uniformis* DA183 (F). Each experiment was conducted in triplicate to ensure reliability. Glc: Glucose; GPs: Ginseng polysaccharides; DPs: *Dendrobium officinale* polysaccharides; GLPs: *Ganoderma lucidum* polysaccharides



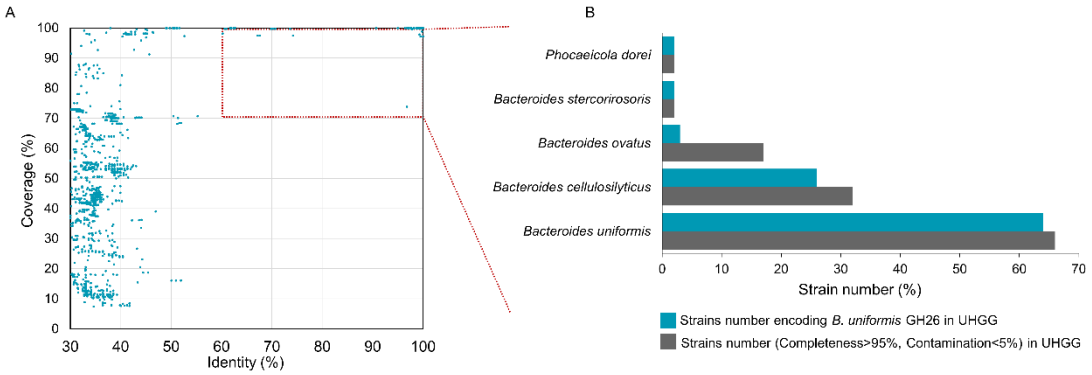
Supplementary Figure 7 Transcriptome sequencing and analysis of *B. uniformis* utilizing GLPs. A) Volcano plots of *B. uniformis* DA183 gene expression comparing minimal media supplemented with GLPs to glucose, showing the fold-change (log2, x-axis) vs. the differential significance (-log10 (adjust *p*-value), y-axis), with highlighted the Top 10 up-regulated genes colored in green. The adjust *p*-value was obtained by DESeq2 using the Wald test, after correcting for multiple hypothesis using the Benjamini and Hochberg method. **B)** The schematic diagram shows that PUL16 of *B. uniformis* DA183 (PUL16_*Bu*) contains 10 genes, with a cluster size of 19,533 base pairs. PUL25 of *B. uniformis* DA183 (PUL25_*Bu*) contains 7 genes, with a cluster size of 10,752 base pairs. PUL39 of *B. uniformis* DA183 (PUL39_*Bu*) contains 6 genes, with a cluster size of 11,237 base pairs. susC: susC-like TonB-dependent transporter; susD: susD-like cell-surface glycan-binding protein; GH: glycoside hydrolyze; Unk: unknown functions. **C)** Bubble plot displays the differential expression of the PUL16_*Bu*, PUL25_*Bu*, PUL39_*Bu* gene and its neighboring 3 genes, utilizing GLPs and glucose. The differential expression levels of genes were transformed into Log2 values. The size and color of the bubbles represent the magnitude of differential expression. The horizontal cyan dotted line represents a zero-fold change in gene expression level.



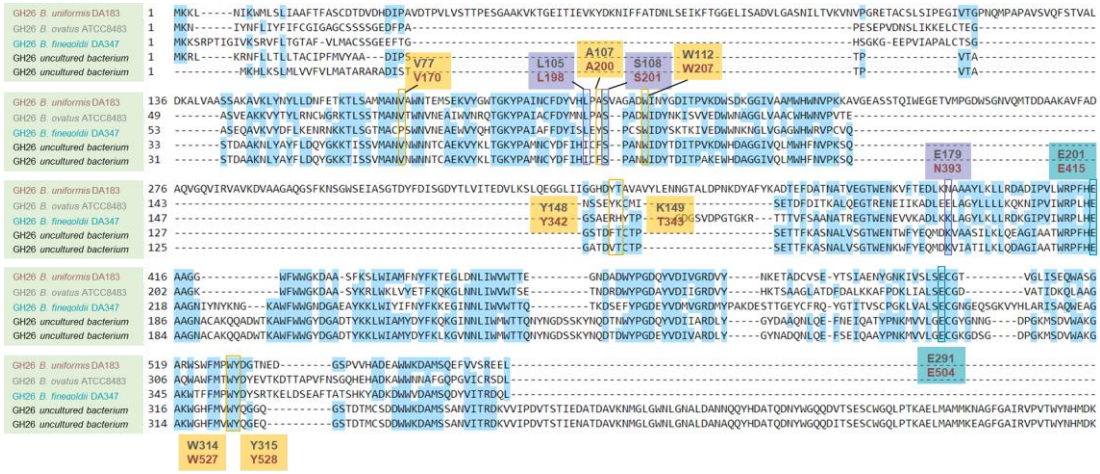
Supplementary Figure 8 The PUL34 gene clusters in *B. uniformis* strain DA183 and strain ATCC8492 are highly conserved. The gene distribution diagram shows the identity and structure of CAZymes genes in PUL34.



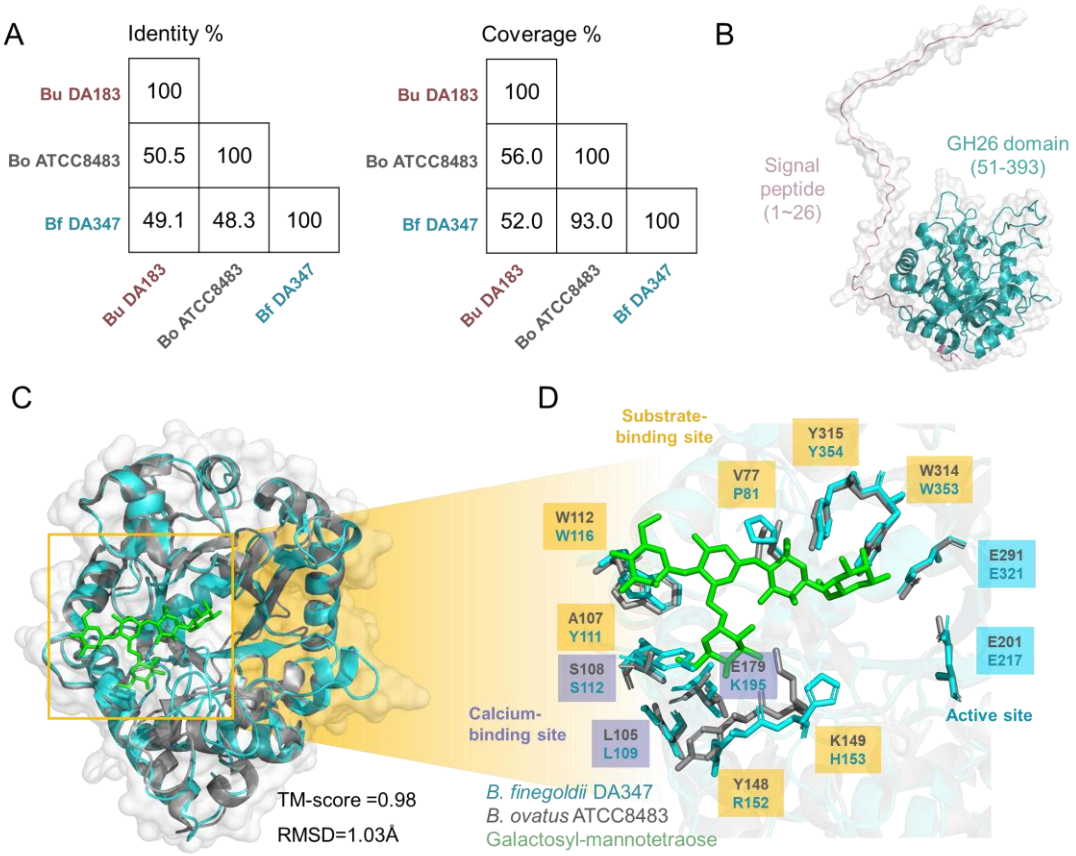
Supplementary Figure 9 Construction of *B. uniformis* knockout mutants. A-B) Plasmid map is used to knockout **A)** *Bu*ATCC8492 PUL34M and **B)** *Bu*ATCC8492 PUL34S. HA: homologous arm of the gene knockout on the left side, HB: homologous arm of the gene knockout on the right side, gRNA: guide RNA, P1: promoter, FnCas12a: CRISPR/Cas12a for genomic cutting, AmpR: ampicillin resistance gene, ErmG: erythromycin resistance gene, R6kγ: replication initiation site, ori: replication initiation site, oriT: conjugation initiation site. **C-D)** Gene knockout DNA sequencing result graph of **C)** *Bu*ATCC8492 ΔPUL34M and **D)** *Bu*ATCC8492 ΔPUL34S.



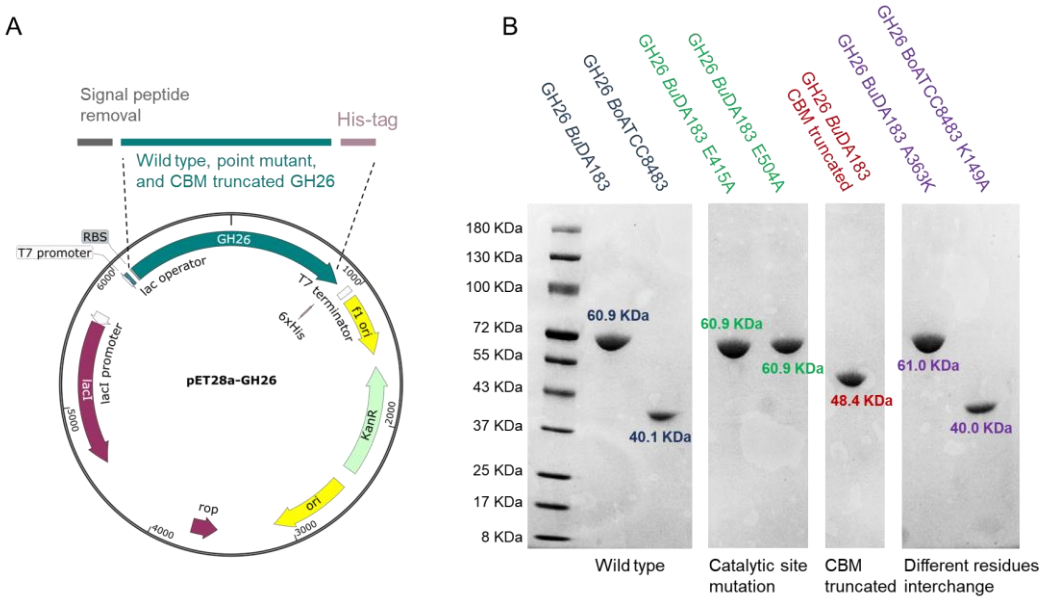
Supplementary Figure 10 Homologs of *B. uniformis* GH26 are found in several *Bacteroides* species. **A)** The sequence alignments of *B. uniformis* GH26 in UHGP[1] was performed using DIAMOND method[2]. The region (identity >60%, coverage>70%) was marked by a red dotted box. **B)** The bar plot showed the number of strains encoded *B. uniformis* GH26 (in cyan), the number of strains used for performing alignments in UHGG.



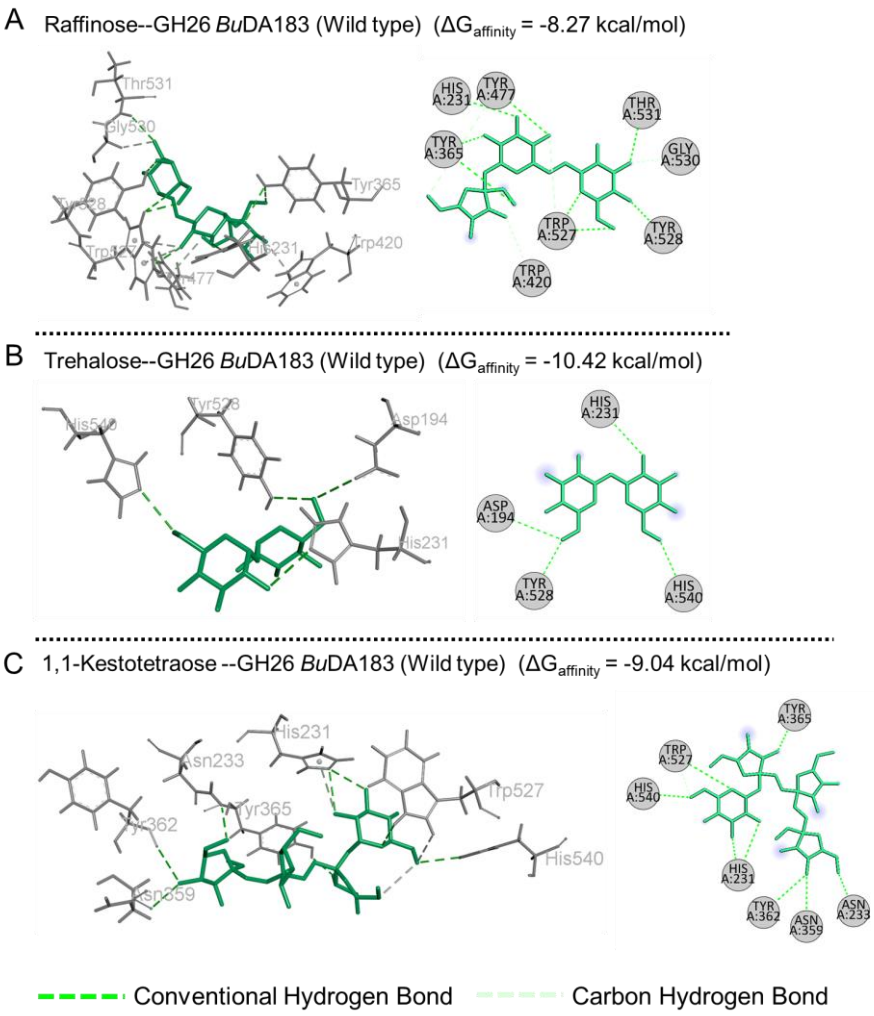
Supplementary Figure 11 The sequence alignment of *B. uniformis* GH26 neighboring protein. The alignment was performed by MUSCLE (version 5.1)[3]. The active sites (E415 and E504), the calcium-binding sites (L198 and S201), and the substrate-binding sites (V170, A200, W207, Y342, W527, and Y528) of *B. uniformis* GH26, and the active sites (E201 and E291), the calcium-binding sites (L105 and S108), and the substrate-binding sites (V77, A107, W112, Y148, W314, and Y315) of *B. ovatus* GH26 is conserved.



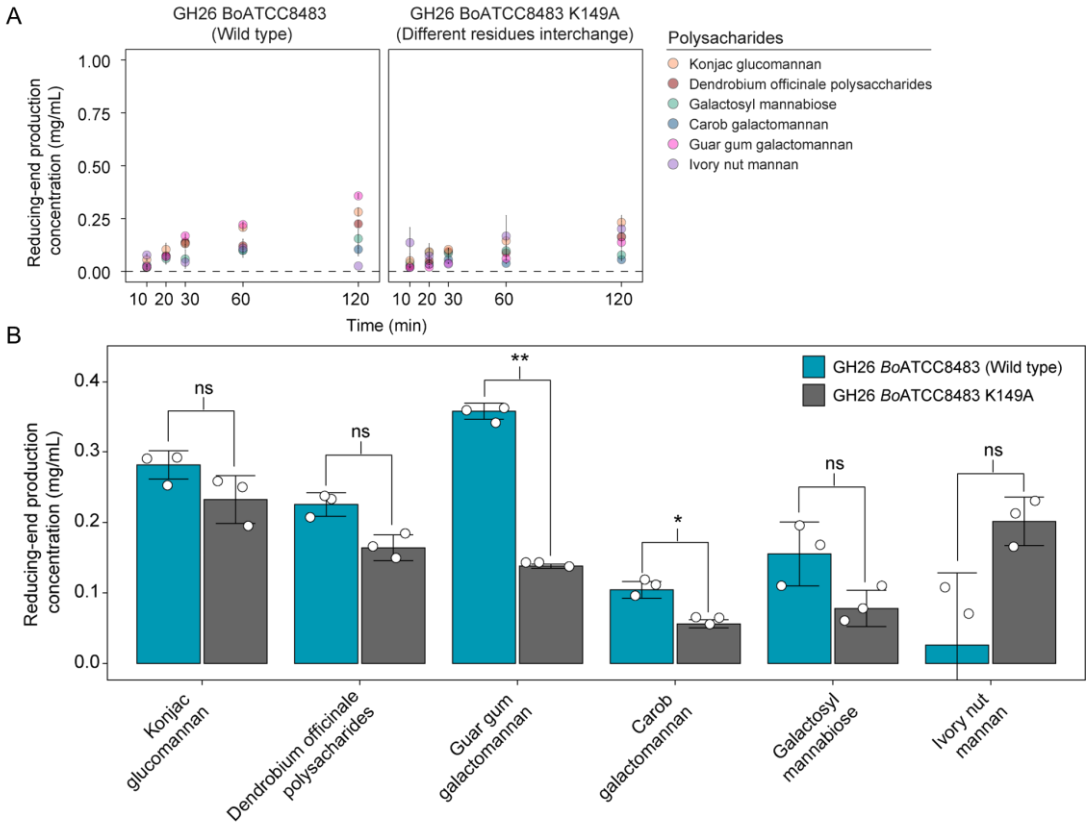
Supplementary Figure 12 The GH26 catalytic domains of *B. ovatus* ATCC8483 and *B. finegoldii* DA347 are highly similar. A) The sequence identity and coverage of GH26 from *B. uniformis* DA183, *B. ovatus* ATCC8483, and *B. finegoldii* DA347. Sequence identity and coverage were calculated using BLASTP (version 2.6.0)[4]. **B)** Overall structure of GH26 from *B. finegoldii* DA347 was predicted by AlphaFold2[5] based on the protein sequence. The signal peptide is colored in light pink, GH26 domain in cyan, respectively. **C)** The structure of GH26 from *B. finegoldii* DA347 (cyan) domain is superimposed on that of *B. ovatus* GH26 (grey) (PDB code 6HF4). **D)** The active sites (E217 and E321), the calcium-binding sites (L109 and S112), and the substrate-binding sites (W116, W353, and Y354) of GH26 from *B. finegoldii* DA347 (cyan stick models), and the active sites (E201 and E291), the calcium-binding sites (L105 and S108), and the substrate-binding sites (W112, W314, and Y315) of *B. ovatus* GH26 (grey stick models) is conserved (TM-score = 0.91, RMSD=1.37Å).



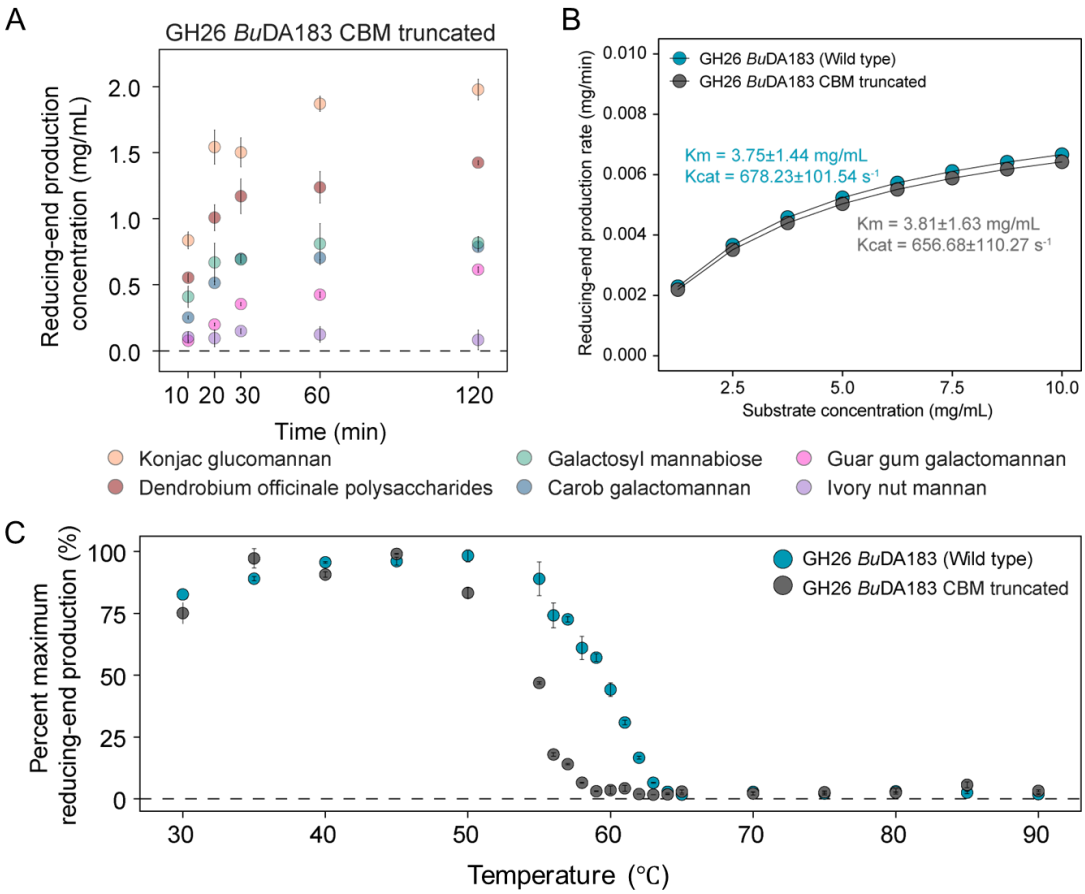
Supplementary Figure 13 The GH26 protein can be expressed solubly *in vitro*. A) The GH26 protein, with the N-terminal signal peptide deleted and a 6x His-tag added to the C-terminus, was constructed in the pET28a heterologous expression vector and transformed into *E. coli* BL21(DE3) for expression. **B)** SDS-PAGE gel image of wild type, point mutant, and CBM truncated GH26 enzyme.



Supplementary Figure 14 Molecular docking and visualizations for GH26 *Bu*DA183 and three non-reactive substrates. A-C) Interactions of GH26 *Bu*DA183 (Wild type) with raffinose **(A)**, 1,1-kestotetraose **(B)**, and trehalose **(C)** are visualized in 3D (left) and 2D (right). Substrates are shown as green sticks or circle, and interacting residues as grey sticks or circle. Interaction forces are displayed as dotted lines.



Supplementary Figure 15 Characterization of enzyme activity for wild type, point mutant, and CBM truncated GH26 enzyme. A) The dot plot shows the reducing-end production over time for six substrates catalyzed by wild-type GH26 *Bo*ATCC8483 and GH26 *Bo*ATCC8483 K149A. The substrates are: konjac glucomannan (yellow), *Dendrobium officinale* polysaccharides (red), galactosyl mannobiose (green), Carob galactomannan (blue), Guar gum galactomannan (pink), and Ivory nut mannan (purple). Two methods were used to predict the CBM domain boundaries (Methods). The enzyme concentration was 125 nM, substrate concentration was 5 mg/mL, reaction temperature was 37°C, and reaction time was 10, 20, 30, 60, 120 min. Data are presented as mean values ± SD. Assays were performed in technical triplicates (n = 3). **B)** The bar chart shows the reducing-end mannose produced by GH26 *Bo*ATCC8483 (Wild type) and GH26 *Bo*ATCC8483 K149A using six mannan substrates. The enzyme concentration was 125 nM, substrate concentration was 5 mg/mL, reaction temperature was 37°C, and reaction time was 2 hours. Data are presented as mean values ± SD. Assays were performed in technical triplicates (n = 3). Statistical significance was determined using a two-sided unpaired t-test between the wild-type group and the mutant group. The P values are indicated as follows: ns>0.05, * < 0.05; ** < 0.01.



190

191 **Supplementary Figure 16 CBM domain of GH26 is related to thermostability. A)** The dot plot
192 shows the reducing-end production over time for 6 reactive substrates catalyzed by CBM truncated
193 GH26 *Bu*DA183 enzyme. The substrates are: konjac glucomannan (yellow), *Dendrobium officinale*
194 polysaccharides (red), galactosyl mannabiose (green), carob galactomannan (blue), guar gum
195 galactomannan (pink), and ivory nut mannan (purple). The enzyme concentration was 125 nM,
196 substrate concentration was 5 mg/mL, reaction temperature was 37°C, and reaction time was 10, 20,
197 30, 60, 120 min. Data are presented as mean values ± SD. Assays were performed in technical triplicates
198 (n = 3). **B)** Michaelis-Menten plots of full-length wild type GH26 *Bu*DA183 (cyan) and CBM
199 truncated GH26 *Bu*DA183 (grey) acting on konjac glucomannan. Activities were measured as rates
200 of reducing-end production using the DNSA assay. The enzyme concentration was 12.5 nM,
201 substrate concentration was 1.25, 2.5, 3.75, 5, 6.25, 75, 8.75, 10 mg/mL, reaction temperature was
202 37°C, and reaction time was 10, 20, 30, 40 min. **C)** The dot plot shows the reducing-end production
203 over temperature for konjac glucomannan hydrolyzed by full-length wild type and CBM truncated
204 GH26 *Bu*DA183 enzyme. The enzyme concentration was 125 nM, substrate concentration was 5

mg/mL, reaction temperature ranged from 30°C to 90°C in 5°C increments and 55°C to 65°C in 1° C increments, and reaction time was 30 min. Data are presented as mean values ± SD. Assays were performed in technical duplicates (n = 2).

Reference

1 Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature biotechnology* **39**, 105-114 (2021).

2 Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature methods* **18**, 366-368 (2021).

3 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797 (2004).

4 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 1-9 (2009).

5 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).