Journal of Microbiology (2021) Vol. 59, No. 10, pp. 931–940
DOI 10.1007/s12275-021-1292-4
eISSN 1976-3794
pISSN 1225-8873

Discovery of novel glycoside hydrolases from *C*-glycoside-degrading bacteria using sequence similarity network analysis[§]

Bin Wei[1,2,3†], Ya-Kun Wang[1†], Jin-Biao Yu[1], Si-Jia Wang[1,4], Yan-Lei Yu[1], Xue-Wei Xu[3*], and Hong Wang[1,2*]

[1]*College of Pharmaceutical Science & Collaborative Innovation Center of Yangtze River Delta Region Green Pharmaceuticals, Zhejiang University of Technology, Hangzhou 310014, P. R. China*
[2]*Key Laboratory of Marine Fishery Resources Exploitment & Utilization of Zhejiang Province, Hangzhou 310014, P. R. China*
[3]*Key Laboratory of Marine Ecosystem and Biogeochemistry, State Oceanic Administration & Second Institute of Oceanography, Ministry of Natural Resources, Hangzhou 310012, P. R. China*
[4]*Center for Human Nutrition, David Geffen School of Medicine, University of California, Los Angeles, California 90024, USA*

(Received Jun 3, 2021 / Revised Jul 15, 2021 / Accepted Aug 4, 2021)

C-Glycosides are an important type of natural product with significant bioactivities, and the C-glycosidic bonds of C-glycosides can be cleaved by several intestinal bacteria, as exemplified by the human faeces-derived puerarin-degrading bacterium *Dorea* strain PUE. However, glycoside hydrolases in these bacteria, which may be involved in the C-glycosidic bond cleavage of C-glycosides, remain largely unknown. In this study, the genomes of the closest phylogenetic neighbours of five puerarin-degrading intestinal bacteria (including *Dorea* strain PUE) were retrieved, and the protein-coding genes in the genomes were subjected to sequence similarity network (SSN) analysis. Only four clusters of genes were annotated as glycoside hydrolases and observed in the genome of *D. longicatena* DSM 13814^T (the closest phylogenetic neighbour of *Dorea* strain PUE); therefore, genes from *D. longicatena* DSM 13814^T belonging to these clusters were selected to overexpress recombinant proteins (CG1, CG2, CG3, and CG4) in *Escherichia coli* BL21(DE3). *In vitro* assays indicated that CG4 efficiently cleaved the O-glycosidic bond of daidzin and showed moderate β-D-glucosidase and β-D-xylosidase activity. CG2 showed weak activity in hydrolyzing daidzin and *p*NP-β-D-fucopyranoside, while CG3 was identified as a highly selective and efficient α-glycosidase. Interestingly, CG3 and CG4 could be selectively inhibited by daidzein, explaining their different performance in kinetic studies. Molecular docking studies predicted the molecular determinants of CG2, CG3, and CG4 in substrate selectivity and inhibition propensity. The present study identified three novel and distinctive glycoside hydrolases, highlighting the potential of SSN in the discovery of novel enzymes from genomic data.

Keywords: *C*-glycosides, puerarin, daidzin, sequence similarity network, molecular docking, inhibition

Introduction

The human gut microbiota is a complex microbial community with an enormous number of microbes that benefit the host by fermentation of complex polysaccharides and natural glycosides into aglycones or short-chain fatty acids in the intestine (Ley *et al.*, 2006; Chassard and Lacroix, 2013; Kim *et al.*, 2015; Thursby and Juge, 2017; Valles-Colomer *et al.*, 2019; Dong *et al.*, 2020). Intestinal bacterial organisms have numerous carbohydrate-active enzyme (CAZyme)-coding genes in their genomes that encode many CAZymes targeting the glycosidic bonds within dietary polysaccharides and glycosides to produce nutrients and energy (Huang *et al.*, 2018).

Glycosides are compounds in which one or more sugars are linked with the aglycone through glycosidic bonds. They are predominantly presented in the form of *O*-glycosides in small amounts in the forms of *C*-glycosides, N-glycosides, and S-glycosides (Nakamura *et al.*, 2013). *O*-Glycosides are among the most abundant compounds found in plants; for example, daidzin (daidzein 7-*O*-glycoside) is the major isoflavone *O*-glycoside in soybean. Additionally, *C*-glycosides are an essential natural product whose glycosidic bond is formed by directly connecting the anomeric carbon of sugar to the carbon atom of aglycon. They are distributed in many medicinal plants and have various bioactivities, such as puerarin (daidzein 8-*C*-glycoside) (Nakamura *et al.*, 2013). Unlike the *O*-glycosidic bond, the *C*-glycosidic bond usually has good stability and cannot be easily degraded by acids, bases, and host enzymes (Ito *et al.*, 2014). Although CAZymes from intestinal bacteria have been widely studied, questions regarding the degradation of *O*-glycoside daidzin and *C*-glycoside puerarin have received little attention.

Hattori and coworkers first reported that human intestinal bacteria could break the *C*-glycosidic bond of homoorientin in 1988 (Hattori *et al.*, 1988a, 1988b). Jin *et al.* (2008) isolated the bacterium *Dorea* strain PUE from human faeces in 2008 that could break the *C*-glycosidic bond of puerarin. In 2011, Nakamura *et al.* (2011) reported that *Dorea* strain PUE broke the *C*-glycosidic bond through hydrolysis but not reduction. To date, only ten bacterial strains have been reported to cleave the *C*-glycosidic bond, and how the enzymes manipulate deglycosylation has been partially elucidated (Wei *et al.*, 2020).

†These authors contributed equally to this work.
*For correspondence. (H. Wang) E-mail: hongw@zjut.edu.cn; Tel.: +86-571-8832-0622 / (X.W. Xu) E-mail: xuxw@sio.org.cn
§Supplemental material for this article may be found at http://www.springerlink.com/content/120956.

For example, Braune *et al.* (2016) reported that the DfgA, DfgB, DfgC, DfgD, and DfgE proteins of *Eubacterium cellulosolvens* ATCC 43171 only metabolized the *C*-glycosidic bond of homoorientin and isovitexin in the presence of Mn^{2+} and NAD^+, but the findings could not explain why the strain manipulated the deglycosylation of puerarin. Recently, Nakamura and colleagues reported that the reaction mechanism of C-deglycosylation of puerarin by *Dorea* strain PUE comprises four reactions, oxidation, deglycosylation, H_2O addition, and reduction, cooperating with at least four enzymes (Nakamura *et al.*, 2019, 2020). However, glycoside hydrolases in these *C*-glycoside-degrading bacteria, which may be involved in the cleavage of *C*-glycosides, remain largely unknown.

In recent decades, with the rapid development of high-throughput sequencing technology, the amount of genomic sequence data has grown exponentially (Brown and Babbitt, 2012). Although many novel functional gene sequences have been identified, the structure and function of their encoding proteins remain unclear. Sequence similarity network (SSN) analysis provides a fast and simple calculation framework to observe the evolution of many related proteins (Atkinson *et al.*, 2009) and may be a strategy to solve this problem. SSNs can visualize relationships among protein sequences, in which the related proteins are grouped in clusters (Levin *et al.*, 2017). Recently, an SSN containing genes encoding GH29 and GH95 α-fucosidases from the CAZy database and α-fucosidase encoding genes from the genomes of the intestinal bacteria *Ruminococcus gnavus* E1 and *R. gnavus* ATCC 29149 was constructed to identify the representative α-fucosidases from *R. gnavus* E1 (Wu *et al.*, 2020).

Among ten intestinal bacterial strains capable of cleaving the *C*-glycosidic bond of *C*-glycosides, five of them, *Lactococcus* sp. MRG-IFC-1, *Enterococcus* sp. MRG-IFC-2, *Lachnospiraceae* strain CG19-1, *Dorea* strain PUE, and *E. cellulosolvens* ATCC 43171, could manipulate the deglycosylation of puerarin and have been taxonomically identified (Braune and Blaut, 2011, 2012; Nakamura *et al.*, 2011; Xu *et al.*, 2014; Kim *et al.*, 2015; Wei *et al.*, 2020). We hypothesize that all five *C*-glycoside-degrading bacteria may contain a specific group of genes responsible for degradation, and these genes probably show high sequence similarities to each other and encode glycoside hydrolases. Because many genes were already annotated as glycoside hydrolases in the genomes of these strains, the genomes of the closest relatives of the five strains were retrieved, and the protein-coding genes in the genomes were subjected to SSN analysis to quickly identify the potential genes involved in the *C*-glycosidic bond cleavage of puerarin. Only genes annotated as glycoside hydrolases and observed in the genome of the closest phylogenetic neighbour of *Dorea* strain PUE were selected to overexpress recombinant proteins in *Escherichia coli* BL21(DE3). Next, the activities of the recombinant proteins in cleaving puerarin, daidzin, and probe substrates of glycoside hydrolases were evaluated using *in vitro* cell-free assays, and the molecular determinants of the substrate selectivity and inhibition propensity were predicted by molecular docking studies.

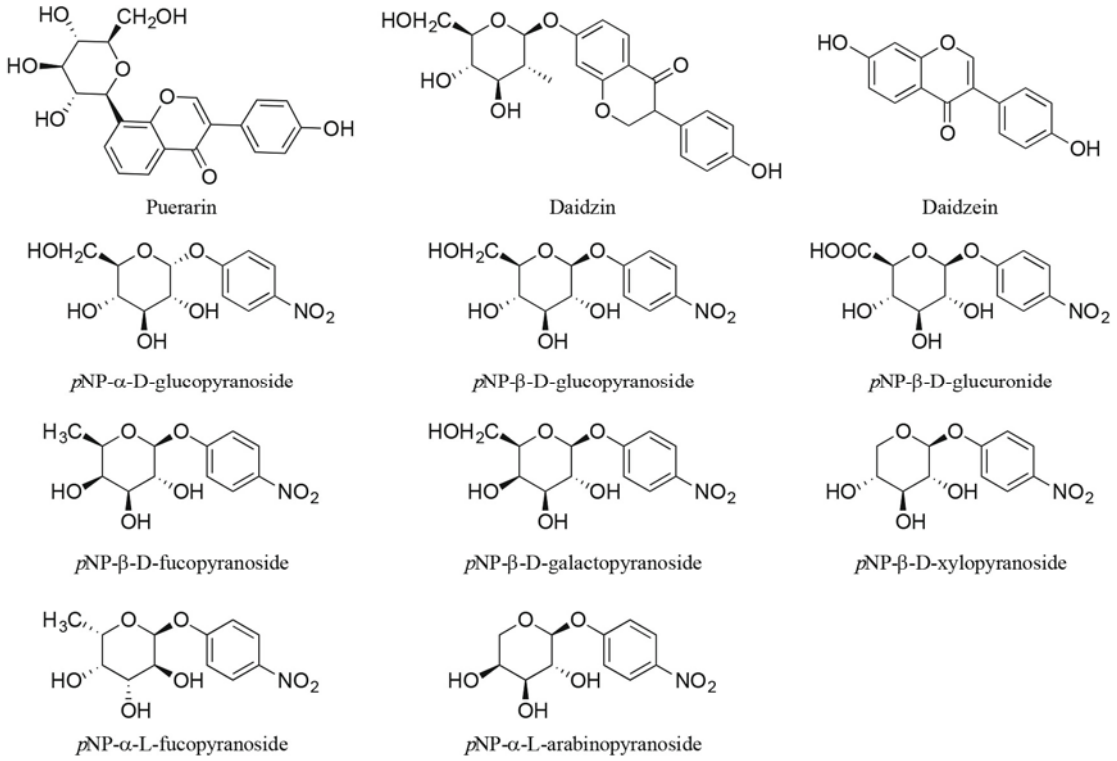

Fig. 1. **Chemical structures of substrates used in this study.**

Materials and Methods

Materials

Puerarin, daidzin, and daidzein (Fig. 1) with ≥ 98% purity by HPLC were purchased from Shanghai Yuanye Biotechnology Co., Ltd. Eight probe substrates for glycoside hydrolases, *p*NP-β-D-xylopranoside (*p*NP-β-Xyl), *p*NP-β-D-galactopyranoside (*p*NP-β-Gal), *p*NP-β-D-fucopyranoside (*p*NP-β-Fuc), *p*NP-α-L-fucopyranoside (*p*NP-α-Fuc), *p*NP-α-L-arabinopyranoside (*p*NP-α-Ara), *p*NP-α-D-glucopyranoside (*p*NP-α-Glc), *p*NP-β-D-glucopyranoside (*p*NP-β-Glc), and *p*NP-β-glucuronide (*p*NP-β-Glu) (Fig. 1) were obtained from Sigma. Dulbecco's phosphate-buffered saline (PBS) was provided by Life Technologies. Imidazole, kanamycin, and isopropyl β-D-1-thiogalactopyranoside (IPTG) were supplied by Biosharp. Other reagents and solvents were of analytical grade.

Construction of the SSN

The 16S rRNA gene sequences of *Lactococcus* sp. MRG-IFC-1 (GenBank accession No. KF803554.1), *Enterococcus* sp. MRG-IFC-2 (GenBank accession No. KF803555.1), *Lachnospiraceae* strain CG19-1 (GenBank accession No. FJ711049.1), *Dorea* strain PUE (GenBank accession No. EU377662.1), and *E. cellulosolvens* ATCC 43171 (GenBank accession No. X71860.1) were retrieved from the literature and identified using the EzBioCloud server against the type strain sequences. (https://www.ezbiocloud.net/identify) (Yoon *et al.*, 2017). The protein-coding gene sequences of the top hits were downloaded from the database, and the resulting amino acid sequences were used to generate the SSN using the Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST) (Gerlt *et al.*, 2015) (available at https://efi.igb.illinois.edu/efi-est/index.php). The network was generated from all-by-all BLAST comparisons of all 13,197 protein-coding gene sequences. The thresholds for sequence identity and alignment scores are mainly determined empirically. Each node represents a single sequence or a group of sequences with more than 95% sequence identity. Each edge in the network represents an alignment hit with an E-value of 10^{-150} or better. The SSN data were visualized using Cytoscape 3.6 (Zallot *et al.*, 2019).

Expression and purification of candidate glycoside hydrolases

Candidate glycoside hydrolase genes were codon-optimized and synthesized by GenScript. Next, gene fragments were cloned into the pET28a vector and transformed into *E. coli* BL21(DE3). The recombinant strains were grown in LB medium containing 50 mg/ml of kanamycin at 37°C until the OD_{600} reached 0.5. Gene expression was induced at 16°C for an additional 12 h by adding IPTG to a final concentration of 0.5 mM. Cells were harvested by centrifugation at 8,000 rpm at 4°C for 10 min. The pellet was suspended in PBS buffer and disrupted by sonication. The cell debris was removed by centrifugation at 10,000 rpm at 4°C for 10 min to obtain the crude extracts. The supernatant was applied onto Ni-NTA His binding resin to purify the C-terminal His-tagged recombinant proteins. The column was eluted using a linear gradient from 0 to 250 mM imidazole in NTA buffer (20 mM Tris-HCl at pH 7.9, 0.5 M NaCl, 10% glycerol). Purified recombinant proteins were concentrated using an ultrafiltration membrane (Amicon Ultra 10 kDa cut-off; Millipore) and analyzed by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) on an 8% separating gel. Protein concentrations were determined using the bicinchoninic acid assay (Beyotime Institute of Biotechnology).

Substrate selectivity of candidate glycoside hydrolases

To investigate the activity of the candidate glycoside hydrolases in hydrolyzing the probe substrates, such as *p*NP-β-Xyl, *p*NP-β-Gal, *p*NP-β-Fuc, *p*NP-α-Fuc, *p*NP-α-Ara, *p*NP-α-Glc, *p*NP-β-Glc, and *p*NP-β-Glu, enzymes were incubated with the corresponding *p*NP-glycosides (*p*NPG) at 37°C for 30 min, and the reaction mixture comprised 10 ml of *p*NPG (0.25 mM), 10 ml of enzymes, and 80 ml of PBS buffer (pH 7.2). The concentration of released *p*NP was determined by recording the absorbance at 405 nm using a microplate reader (Spectra Max 190; Molecular Devices). To compare the activity of the enzymes in cleaving the *C*-glycosidic and *O*-glycosidic bonds, the enzymes were incubated with puerarin and daidzin, respectively. The 100-ml reaction mixture comprised 88 ml of 50 mM PBS buffer, 10 ml of enzyme (final concentration: 0.2 mg/ml), and 2 ml of 12.5 mM substrate in DMSO and was incubated at 37°C for 30 min. The reaction was terminated by adding 100 ml of acetonitrile. Aliquots of the supernatants were injected onto an Agilent LC 1260 HPLC-DAD with a ZORBAX Extend-C_{18} column (4.6 × 250 mm; 5-mm particle size; Agilent Technologies).

Effects of pH, temperature, and metal ions on the activity of candidate glycoside hydrolases

The effects of pH on enzymatic activity were determined at 37°C at pH values ranging from 3.0 to 11.0 using different buffers. The buffer solutions used were as follows: 50 mM glycine-HCl buffer (pH 3.0), 50 mM sodium acetate buffer (pH 3.6, 4.0, and 5.0), PBS buffer (pH 5.8, 6.0, 7.0, and 8.0), and 50 mM glycine-NaOH buffer (pH 8.0, 8.6, 9.0, 10.0, and 11.0). The reaction mixtures were prepared by mixing 10 ml of *p*NPG (0.25 mM), 10 ml of enzyme solutions, and 80 ml of buffer solutions with different pH values. The effects of temperature on the activity of candidate glycoside hydrolases were characterized at the optimal pH at temperatures ranging from 20°C to 80°C. The impacts of Na^+, K^+, Mg^{2+}, Zn^{2+}, Mn^{2+}, Cu^{2+}, and Ca^{2+} (final concentration: 1 mM) on the activity of candidate glycoside hydrolases were evaluated at 37°C in buffers with optimal pH. The reactions lasted for 30 min, and the activity of the enzymes was quantified according to the methods described above.

Kinetics of candidate glycoside hydrolases in glycoside hydrolysis

The kinetic parameters of candidate glycoside hydrolases were determined using *p*NPG or daidzin as the substrates under the initial reaction rate. The concentration ranges of *p*NP-α-Glc, *p*NP-β-Glc, *p*NP-β-Fuc, *p*NP-β-Xyl, and daidzin were optimized as 0.25–5.00 mM, 50–1,000 mM, 50–1,000 mM, 50–800 mM, and 50–250 mM, respectively. The reaction mixtures comprised 2 ml of glycosides, 10 ml of enzymes, and 88 ml of PBS buffer. *p*NPG was incubated at 37°C for 30 min, and the incubation time for daidzin was 3 h. The ac-

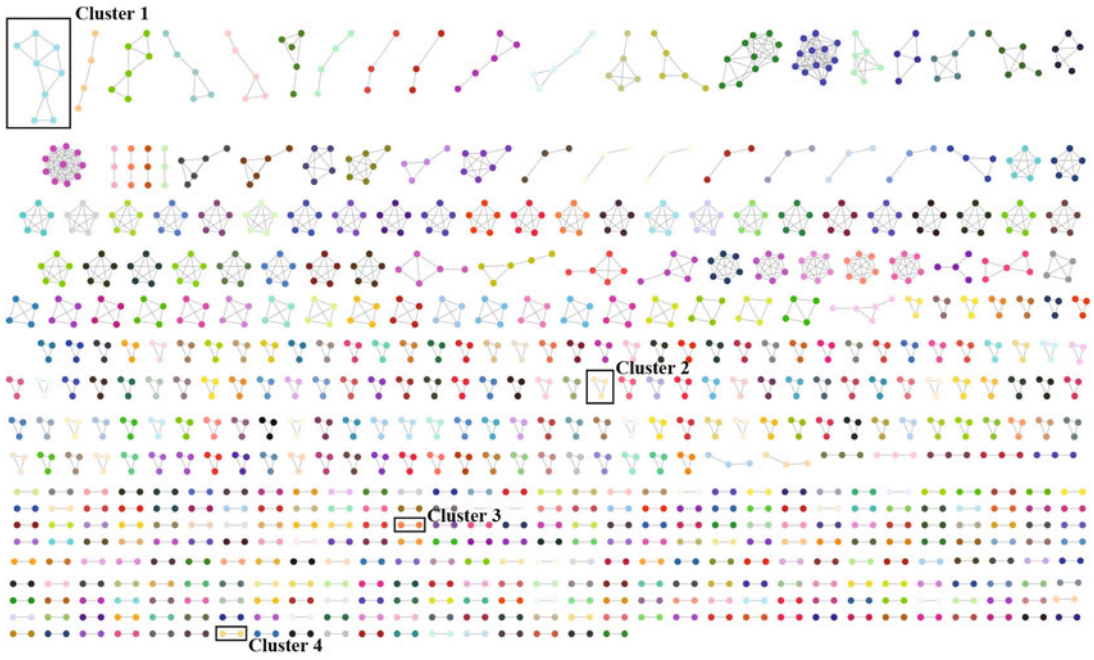

Fig. 2. **Sequence similarity network of protein-coding genes in the genomes of the closest relatives of *Lactococcus* sp. MRG-IFC-1, *Enterococcus* sp. MRG-IFC-2, *Lachnospiraceae* strain CG19-1, *Dorea* strain PUE, and *E. cellulosolvens* ATCC 43171.**

tivity of the enzymes was quantified according to the methods described above. The Michaelis constant (K_m), maximal velocity (V_{max}), and turnover number (k_{cat}) values were calculated using GraphPad Prism 5.0 (GraphPad Software).

Time-course curves of candidate glycoside hydrolases in daidzin hydrolysis and effects of daidzein on the activity of the enzymes

To explain the different behaviours of the candidate glycoside hydrolases in kinetic studies, the time course of candidate glycoside hydrolases in hydrolyzing daidzin and the effects of daidzein on the activity of the enzymes were assessed. For time course experiments, enzymes were incubated with 250 mM daidzin at 37°C for 2, 4, 8, 12, 16, and 24 h. The effects of daidzein on the activity of enzymes in hydrolyzing *p*NPG or daidzin were determined by preincubating the enzymes with 100 mM daidzein at 37°C for 5 min and then mixing them with the corresponding substrate solutions (final concentration: 250 mM). The formation of *p*NP and daidzein was also quantified as described above.

Homology modelling and molecular docking studies

Molecular docking studies were performed to predict the molecular determinants of the candidate glycoside hydrolases in substrate selectivity and inhibition propensity. The 3D structures of the enzymes were constructed by homology modelling and validated by Ramachandran plots using MOE (Version 2019.09; Chemical Computing Inc.) according to our previous report (Wei *et al.*, 2018). Substrates or inhibitors were docked into the active sites of the 3D models of the enzymes using the same parameters as described in our recent work (Zhou *et al.*, 2020).

Nucleotide sequence accession number

The codon-optimized gene sequences of candidate glycoside hydrolases CG1-CG4 have been deposited in GenBank under accession numbers MW770309-MW770312.

Results

SSN analysis reveals the gene sequences of candidate glycoside hydrolases

To date, only five bacterial strains, *Lactococcus* sp. MRG-IFC-1, *Enterococcus* sp. MRG-IFC-2, *Lachnospiraceae* strain CG19-1, *Dorea* strain PUE, and *Eubacterium cellulosolvens* ATCC 43171, could cleave the *C*-glycosidic bond of puerarin and have been taxonomically identified (Braune and Blaut, 2011, 2012; Nakamura *et al.*, 2011; Xu *et al.*, 2014; Kim *et al.*, 2015; Wei *et al.*, 2020). Therefore, the genomes of *L. lactis* subsp. *hordniae* NBRC 100931^T, *Enterococcus lactis* BT-

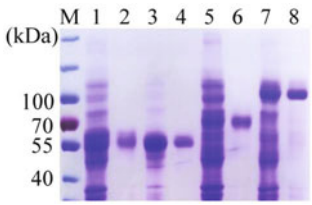

Fig. 3. **SDS-PAGE analysis of cell extracts and purified enzymes.** Lanes: M, protein marker; 1, cell extracts of CG1 after IPTG induction; 2, purified CG1; 3, cell extracts of CG2 after IPTG induction; 4, purified CG2; 5, cell extracts of CG3 after IPTG induction; 6, purified CG3; 7, cell extracts of CG4 after IPTG induction; 8, purified CG4.

Table 1. Substrate selectivity of CGs
The data are presented as the means ± standard error of the mean from three independent experiments.

Substrate	Activity (mmol/mg/h)			
	CG1	CG2	CG3	CG4
*p*NP-α-D-glucopyranoside	-[a]	-	381.6 ± 5.9	-
*p*NP-β-D-glucopyranoside	-	-	-	11.2 ± 0.3
*p*NP-β-D-glucuronide	-	-	-	1.5 ± 0.96
*p*NP-β-D-fucopyranoside	-	0.11 ± 0.02	-	-
*p*NP-β-D-galactopyranoside	-	-	1.9 ± 0.09	1.7 ± 0.02
*p*NP-β-D-xylopyranoside	-	-	-	14.6 ± 0.59
*p*NP-α-L-fucopyranoside	-	-	-	-
*p*NP-α-L-arabinopyranoside	-	-	-	-
Puerarin	-	-	-	-
Daidzin	-	0.08 ± 0.01	-	103.0 ± 5.3

[a] -, no activity was observed.

159[T], *Frisingicoccus* FJ711049_s, *D. longicatena* DSM 13814[T], and *Eubacterium cellulosolvens* ATCC 43171[T], whose 16S rRNA genes showed the highest sequence similarity (89.8–100%) to the corresponding strains (Supplementary data Table S1), were retrieved. All 13,197 protein-coding genes in these five genomes were subjected to SSN analysis. The protein-coding gene sequences were distributed in hundreds of clusters (Fig. 2). Only four clusters of genes (clusters 1–4) were annotated as glycoside hydrolases and contained protein-coding genes from *D. longicatena* DSM 13814[T]. Therefore, genes from *D. longicatena* DSM 13814[T] belonging to these clusters were selected to overexpress the candidate glycoside hydrolases (CG1, CG2, CG3, and CG4) (Supplementary data Table S2).

Expression and purification of CGs

The gene sequences of the four glycoside hydrolase candidates from *D. longicatena* DSM 13814[T] were code-optimized and directly synthesized by Genescript. The gene fragments were cloned into pET-28a and transformed into *E. coli* BL21 (DE3) to obtain the recombinant proteins. The molecular weights of CG1-CG4 calculated by ExPASy were 55, 56, 64, and 95 kDa. CG1-CG4 were abundantly overexpressed after IPTG induction, and the recombinant proteins were purified by affinity chromatography and ultrafiltration (Fig. 3).

CGs exhibit different substrate specificity profiles

The enzymatic activity of the four candidate glycoside hydrolases in hydrolyzing puerarin and daidzin, as well as the

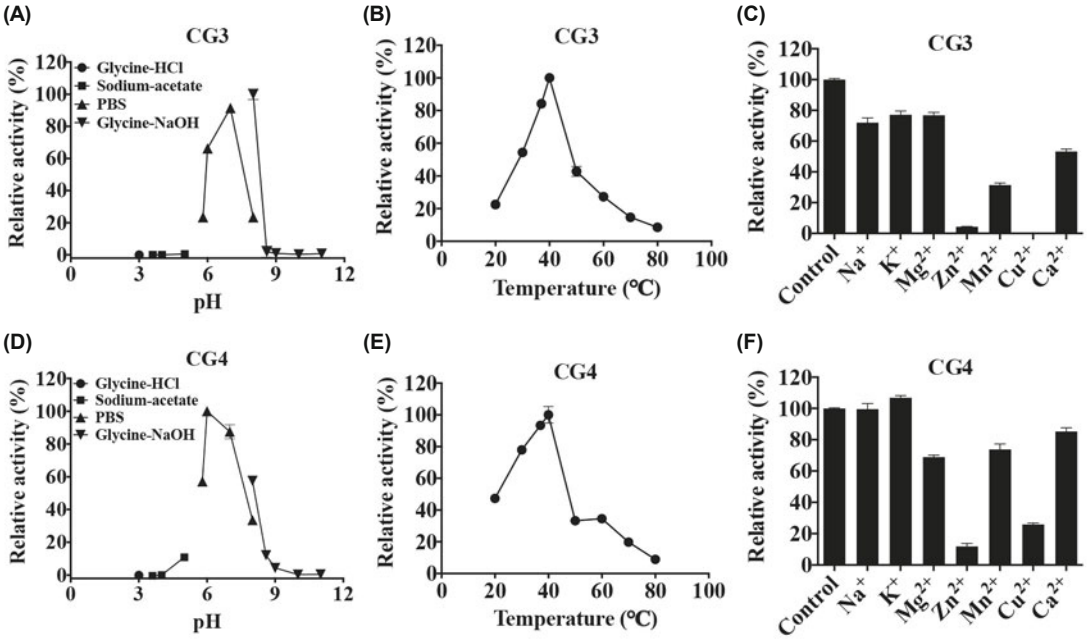

Fig. 4. Effects of pH, temperature, and metal on the *p*NPG hydrolysis activity of CG3 and CG4. Relative activity of CG3 at various (A) pH, (B) temperature, and (C) in the presence of various metal ions. Relative activity of CG4 at various (D) pH, (E) temperature, and (F) in the presence of various metal ions. All data were expressed as mean ± SD of triplicate reactions.

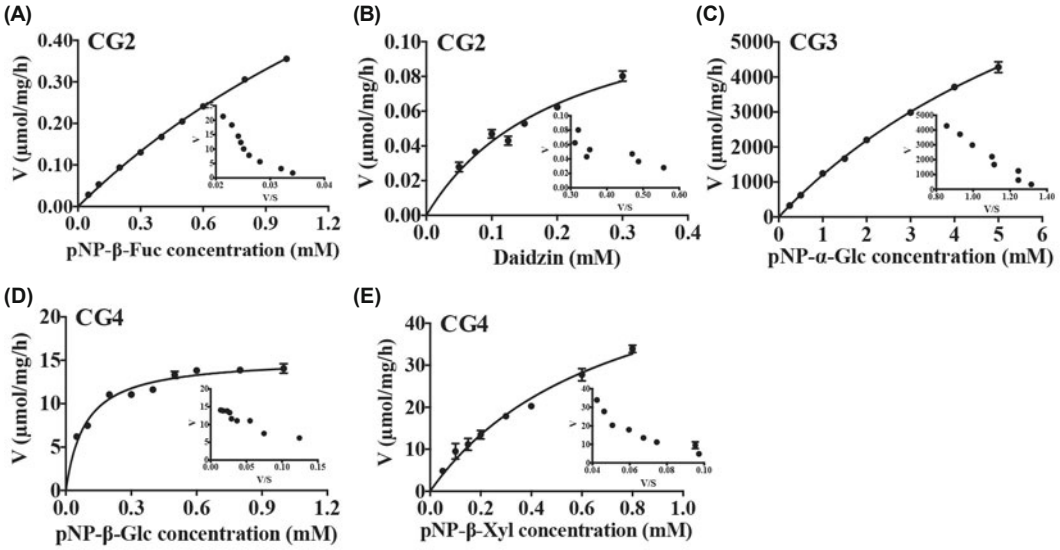

Fig. 5. Kinetic profiles of CG2 in hydrolyzing (A) *p*NP-β-fucopyranoside and (B) daidzin. Kinetic profiles of CG3 in hydrolyzing (C) *p*NP-α-L-glucopyranoside. Kinetic profiles of CG2 in hydrolyzing (D) *p*NP-β-D-glucopyranoside and (E) *p*NP-β-D-xylopyranoside. Insets are Eadie-Hofstee plots. All data were expressed as mean ± SD of triplicate reactions.

eight probe substrates (Fig. 1), were assessed at pH 7.2 and 37°C. Unfortunately, none of them could cleave the *C*-glycosidic bond of puerarin, but CG4 could efficiently hydrolyze the *O*-glycosidic bond of daidzin at a rate of 103.0 μmol/mg/h (Table 1). CG2 also showed slight activity towards daidzin (0.08 μmol/mg/h) and moderate activity towards *p*NP-β-Fuc (0.11 μmol/mg/h), and CG4 showed moderate enzymatic activity towards *p*NP-β-Glc (11.2 μmol/mg/h) and *p*NP-β-Xyl (14.6 μmol/mg/h), as well as weak activity against *p*NP-β-Glu (1.5 μmol/mg/h) and *p*NP-β-Gal (1.7 μmol/mg/h). CG3 was estimated to be a highly efficient α-glycosidase (381.6 μmol/mg/h) but also showed weak β-galactosidase activity (1.9 μmol/mg/h). CG1 showed no detectable activity towards the tested substrates. The findings revealed the different substrate specificity profiles of the three novel glycoside hydrolases.

Effects of pH, temperature, and metal ions on the enzymatic activity of CGs

Because CG3 and CG4 were more active in hydrolyzing glycosides than other CGs, the effects of pH, temperature, and metal ions on the enzymatic activity of CGs were assessed using the ability of CG3 to hydrolyze pNP-α-Glc and that of CG4 to hydrolyze pNP-β-Glc. Thus, the optimal conditions of representative CGs could be efficiently characterized. The optimal pH of CG3 in hydrolyzing *p*NP-α-Glc and that of CG4 in hydrolyzing *p*NP-β-Glc was determined at pH 3.0–11.0. CG3 showed maximum activity in glycine-NaOH buffer at pH 8.0 and retained more than 66% of activity in PBS buffer at pH 6.0 and 7.0, while CG4 displayed the optimal activity in PBS buffer at pH 6.0 and negligible activity at pH < 5 or pH > 9 (Fig. 4A and B). CG3 and CG4 were temperature-sensitive, with the optimal temperature estimated at approximately 40°C (Fig. 4B and E). The effect of metal ions on the activity of CG3 and CG4 is presented in Fig. 4C and F. The presence of Zn^{2+} and Cu^{2+} almost completely inhibited the activity of CG3 and CG4 (inhibition rate > 74%), and the addition of Na^{+} and K^{+} partially inhibited the activity of CG3 with an inhibition rate close to 30% but did not affect the activity of CG4. Mg^{2+}, Mn^{2+}, and Ca^{2+} also incompletely inhibited the activity of CG3 and CG4 with inhibition rates ranging from 14.7% to 68.5%.

Kinetics parameters of CGs in glycoside hydrolysis

The impact of the substrate concentration on the enzymatic activity of CGs is depicted by direct plots and Eadie-Hofstee plots (Fig. 5). According to the linear tendency in the Eadie-Hofstee plots, all the reactions followed typical Michaelis-Menten kinetics. The kinetic parameters are summarized in Table 2. The V_{max} and K_m values of CG2 in hydrolyzing *p*NP-β-Fuc were 1.3 ± 0.11 μmol/mg/h and 2.6 ± 0.27 mM, respectively, which were approximately ten and eight times as

Table 2. Kinetic parameters of CGs towards pNPG and daidzin
The data are reported as the means ± standard error of the mean from three independent experiments.

Substrate	Enzyme	K_m (mM)	V_{max} (μmol/mg/h)	k_{cat} (sec^{-1})	k_{cat}/K_m (sec^{-1} mM^{-1})
*p*NP-β-Fuc	CG2	2.6 ± 0.27	1.3 ± 0.11	0.02 ± 0.002	0.008 ± 0.0002
Daidzin	CG2	0.27 ± 0.078	0.16 ± 0.027	0.001 ± 0.0004	0.009 ± 0.001
*p*NP-α-Glc	CG3	9.1 ± 1.4	12,000 ± 1,400	215.6 ± 25.5	23.7 ± 0.92
*p*NP-β-Glc	CG4	0.089 ± 0.005	15.3 ± 0.32	0.4 ± 0.008	4.5 ± 0.16
*p*NP-β-Xyl	CG4	0.73 ± 0.029	62.6 ± 1.8	1.7 ± 0.05	2.3 ± 0.05

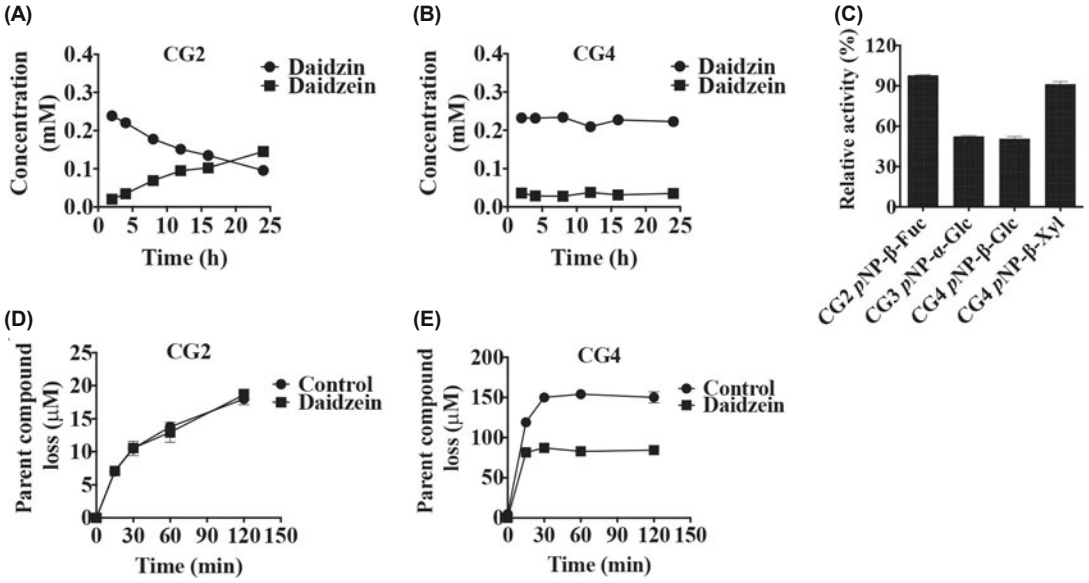

Fig. 6. The time-course curves of (A) CG2 and (B) CG4 in hydrolyzing daidzin. (C) The effects of daidzein on the pNPG-hydrolysis activity of CGs. The effect of daidzein on the activity of (D) CG2 and (E) CG4 in hydrolyzing daidzin.

much as those of CG2 in hydrolyzing daidzin, respectively, suggesting that CG2 had a high affinity but low capacity for daidzin. Thus, the catalytic efficiency (k_{cat}/K_m) of CG2 towards them was comparable (0.008 ± 0.0002 *vs.* 0.009 ± 0.0010 sec^{-1} mM^{-1}). The V_{max} and k_{cat}/K_m values of CG3 in hydrolyzing *p*NP-α-Glc were 12,000 ± 1,400 μmol/mg/h and 23.7 ± 0.92 sec^{-1} mM^{-1}, respectively, indicating that CG3 was a highly efficient α-glucosidase. CG4 exhibited a higher affinity and lower capacity towards *p*NP-β-Glc than *p*NP-β-Xyl. The results demonstrated the distinct behaviours of these three

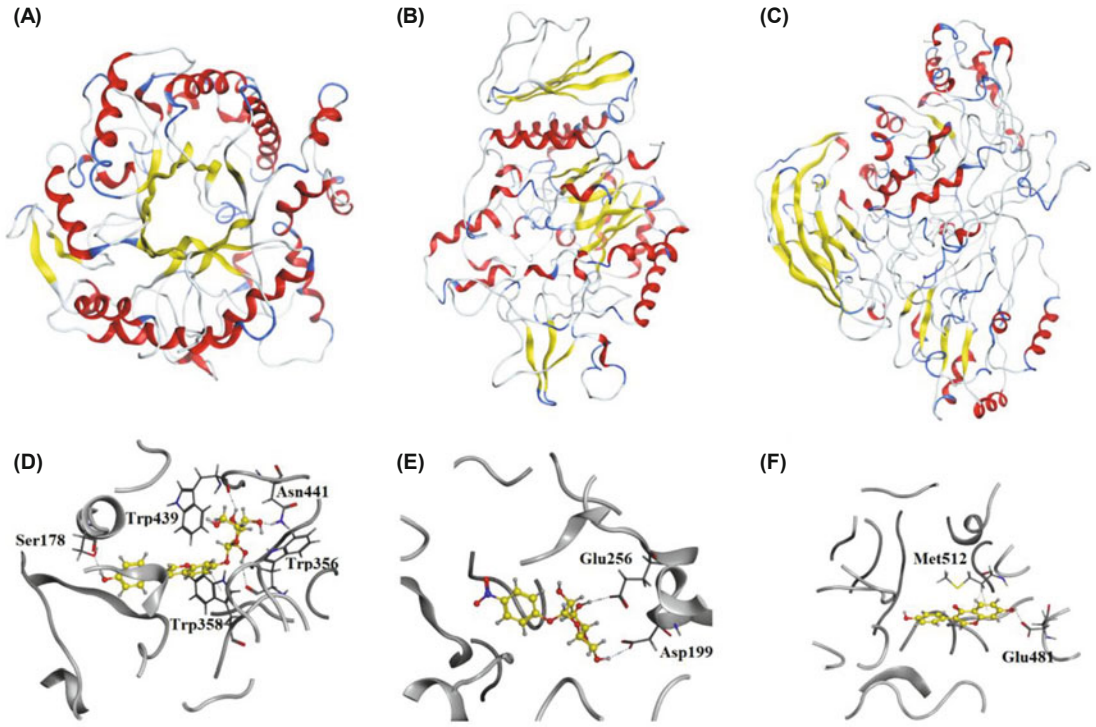

Fig. 7. Ribbon diagrams of the homology model of (A) CG2, (B) CG3, and (C) CG4. Binding modes of (D) daidzin in the active site of the model of CG2, (E) *p*NP-α-Glc in the active site of the model of CG3, and (F) daidzein in the active site of the model of CG4. The ligand is colored in yellow and the surrounding residues of the binding pockets are colored in gray. The backbone of the receptor is depicted as gray.

novel glycoside hydrolases.

Effect of daidzein on the enzymatic activity of CGs

Although CG4 showed excellent enzymatic activity in hydrolyzing the *O*-glycosidic bond of daidzin, unlike CG2, the time course of CG4 was quite unique, and the reaction reached equilibrium after 30 min of incubation (Fig. 6A and B), suggesting that the product daidzein may inhibit the enzymatic activity of CG4. Therefore, the effects of daidzein on the *p*NPG hydrolysis activity of CG2-CG4 were determined. Daidzein exhibited approximately 50% inhibition towards CG3 in hydrolyzing *p*NP-α-Glc and CG4 in hydrolyzing *p*NP-β-Glc and negligible inhibition towards CG2 in hydrolyzing *p*NP-β-Fuc and CG4 in hydrolyzing *p*NP-β-Xyl, demonstrating the varied inhibition propensity of CGs (Fig. 6C). Daidzein also did not affect the activity of CG2 in metabolizing daidzin (Fig. 6D) but could partially inhibit the activity of CG4 in metabolizing daidzin (Fig. 6E).

Homology modelling and molecular docking studies

After searching for the homologous proteins of CGs using MOE, a putative phospho-β-glucosidase from *Streptococcus mutans* UA159 (PDB 4F66), an oligo-1,6-glucosidase from *Bacillus cereus* (PDB 1UOK), and a β-glucosidase from *Thermotoga neapolitana* DSM 4359 (PDB 2X40) showed the highest similarity to CG2, CG3, and CG4, with similarities of 58.2%, 60.9%, and 36.3%, respectively. Top-ranked proteins were used as templates to construct the homology models of CG2-CG4. The ribbon diagrams of the homology model of CG2, CG3, and CG4 are presented in Fig. 7A–C. The Ramachandran plots showed that most of the residues of CGs were in the allowed regions, suggesting the good quality of the models (Supplementary data Fig. S1). Daidzin could properly dock into the active site of CG2 through hydrogen bond interactions with Ser178, Trp356, Trp358, Trp439, and Asn441, suggesting that these residues may be essential in daidzin hydrolysis by CG2 (Fig. 7D). *p*NP-α-Glc bound at the entrance of the active site of CG3 and developed hydrogen bond interactions with Asp199 and Glu256 (Fig. 7E). The binding mode of daidzein with CG4 is depicted in Fig. 7F. Daidzein formed hydrogen bond interactions with residues Glu481 and Met512 in the active site of CG4, implying the important roles of the two residues in the inhibition by daidzein.

Discussion

C-Glycoside-degrading bacteria have attracted increased attention because of the unclear degradation mechanism. Several intestinal bacterial isolates have been reported to break the *C*-glycosidic bond of some *C*-glycosides, such as puerarin, barbaloin, and mangiferin, but glycoside hydrolases in these bacteria, which are involved in the cleavage of *C*-glycosides, remain largely unknown, particularly for the puerarin-degrading bacterium *Dorea* strain PUE. Here, the protein-coding genes in the genomes of the closest relatives of five bacterial strains capable of breaking down the *C*-glycosidic bond were retrieved and submitted to SSN analysis. Only four gene sequences from *D. longicatena* DSM 13814^{T} exhibited the potential to be glycoside hydrolases. The recombinant proteins (CG1-CG4) were successfully overexpressed and purified from *E. coli* BL21(DE3). The *in vitro* assays demonstrated that none could break the *C*-glycosidic bond of puerarin, but CG2 and CG4 could cleave the *O*-glycosidic bond of daidzin. CG3 was identified as a highly efficient and selective α-glycosidase. Daidzein was characterized as a potent inhibitor of CG4, a selective inhibitor of CG3, and showed no inhibitory potency towards CG2. Molecular docking studies explored the molecular determinants of CGs in substrate selectivity and inhibition propensity.

In the present study, five strains were found to manipulate the deglycosylation of puerarin and were taxonomically identified. Because the five strains were not type strains and their genomic sequences were unavailable, we retrieved the protein-coding genes in the genomes of the closest phylogenetic neighbours of five strains from the EzBioCloud database (Supplementary data Table S2) and submitted them to SSN analysis with intentions to determine the potential glycoside hydrolases. Given that 199,907 qualified genomes are found in the EzBioCloud database, the gene sequences of the candidate glycoside hydrolases from *D. longicatena* DSM 13814^{T} (the closest phylogenetic neighbour of *Dorea* strain PUE) should exhibit high sequence identity to those from *Dorea* strain PUE, but the sequence similarity between them must be investigated further.

SSN is a powerful method to assign functions to uncharacterized genes and mine potential target gene sequences based on sequence similarity and functional gene annotation results (Atkinson *et al.*, 2009). Levin *et al.* (2017) identified a prominent glycyl radical enzyme in the human gut microbiome using SSN analysis. Recently, Wu *et al.* (2020) characterized an α-fucosidase from *R. gnavus* E1 by analyzing the sequence relationships between identified α-fucosidases and protein-coding genes in the genomes of *R. gnavus* E1 using SSN. Here, we constructed an SSN of protein-coding genes from five bacterial isolates and identified three novel glycoside hydrolases that showed distinct substrate selectivity and inhibition propensity. For example, CG2 could hydrolyze *p*NP-β-Fuc and daidzin and could not be inhibited by daidzein. CG3 was identified as a highly efficient and selective α-glycosidase and could be partially inhibited by daidzein. CG4 showed broad-spectrum hydrolase activity, and the inhibitory potency of daidzein varied depending on the substrates. Our findings highlight the potential of SSN in the discovery of novel enzymes from genomic data.

In the present study, three novel glycoside hydrolases (CG2-CG4) were identified from the human faeces-derived anaerobic bacterium *D. longicatena* DSM 13814^{T}. The enzymes show hydrolytic activity against various *p*NP-glycosides, suggesting the strong adaptability of intestinal microorganisms by utilizing various glycosides as energy resources. The catalytic efficiency of CG3 in hydrolyzing *p*NP-α-Glc is much higher (k_{cat}/K_m: 1.4 *vs.* 0.11 $\mu M^{-1}min^{-1}$) than that of α-glucosidase from *Gardnerella* spp. (Bhandari *et al.*, 2021). The application of CG3 in the preparation of oligosaccharides warrants further investigation. CG2 and CG3 could hydrolyze daidzin, demonstrating the important roles of intestinal bacterial enzymes in the *in vivo* efficacy of natural plant glycosides.

β-Glycosides are a unique group of bioactive metabolites that are abundantly found in Chinese medicine, fruits, and tea (Hollman, 2001; Jyotshna *et al.*, 2016; Wang *et al.*, 2020). Microbial β-glucosidases play important roles in the deglycosylation of natural β-glycosides, such as soybean isoflavone glycosides (Kim *et al.*, 2011; Pei *et al.*, 2016; Li *et al.*, 2018; Qu *et al.*, 2020), but no study has reported that the activity can be inhibited by daidzein. In the present study, CG2 and CG4 from the same strains could catalyze the cleavage of the *O*-glycosidic bond of daidzin, suggesting the high potential of intestinal bacteria in β-glycoside metabolism. Interestingly, CG2 could not hydrolyze the probe substrate for β-glucosidase but could hydrolyze the substrate for β-fucosidase and daidzin. According to the gene annotation results of cluster 2 (Supplementary data Table S1) and similarity of homologous proteins, it is likely to be a phospho-β-glucosidase.

Wei *et al*. (2017) reported that daidzein shows a more potent inhibitory effect on α-glucosidase than acarbose, and the amino acid residue Thr215 of α-glucosidase is necessary for the interaction. Seong *et al*. (2016) also confirmed that daidzein is a potent inhibitor of α-glucosidase with IC_{50} values of 8.58 ± 0.94 μM, but the amino acid residues of α-glucosidase that developed hydrogen bond interactions with daidzein are Arg270, Ala292, and His295. Here, we found that daidzein is an inhibitor of CG3 and CG4 but cannot inhibit the enzymatic activity of CG2. Notably, the inhibitory potency of daidzein on CG4 varied depending on the substrate. Residues Glu481 and Met512 in the active site of CG4 may play important roles in the inhibition by daidzein. The low sequence similarity (28.2%) between α-glucosidase and CG4 may explain the different key residues for inhibition. The molecular determinants of CGs in substrate selectivity were also predicted by molecular docking studies, but the precise mechanism warrants structural biology studies.

In conclusion, the present study using SSN analysis identified three novel glycoside hydrolases that show varied substrate selectivity and inhibition propensity and predicted the molecular determinants of the three glycoside hydrolases in substrate selectivity and inhibition propensity. The findings highlight the functional diversity of glycoside hydrolases and potential of SSN in the identifying of novel enzymes from genomic data.

Acknowledgements

This work was financially supported by the National Key Research and Development Program (2017YFE0103100), the programs of the National Natural Science Foundation of China (No. 81903534, No. 81773628, and No. 81741165), the High-Level Talent Special Support Plan of Zhejiang Province (No. 2019R52009) and the Scientific Research Fund of Second Institute of Oceanography, State Oceanic Administration (JB2002).

Conflict of Interest

All authors declare no conflicts of interest. This article does not contain any studies with human participants performed by any of the authors.

Author Contributions

HW and XX conceived and designed the study. BW and YW conducted experiments and wrote the manuscript. JY and SW contributed analytical tools and analyzed data. YY revised the manuscript. All the authors read and approved the manuscript.

References

Atkinson, H.J., Morris, J.H., Ferrin, T.E., and Babbitt, P.C. 2009. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS ONE* **4**, e4345.

Bhandari, P., Tingley, J.P., Palmer, D.R., Abbott, D.W., and Hill, J.E. 2021. Characterization of an α-glucosidase enzyme conserved in *Gardnerella* spp. isolated from the human vaginal microbiome. *J. Bacteriol.* **203**, e0021321.

Braune, A. and Blaut, M. 2011. Deglycosylation of puerarin and other aromatic *C*-glucosides by a newly isolated human intestinal bacterium. *Environ. Microbiol.* **13**, 482–494.

Braune, A. and Blaut, M. 2012. Intestinal bacterium *Eubacterium cellulosolvens* deglycosylates flavonoid *C*- and *O*-glucosides. *Appl. Environ. Microbiol.* **78**, 8151–8153.

Braune, A., Engst, W., and Blaut, M. 2016. Identification and functional expression of genes encoding flavonoid *O*- and *C*-glycosidases in intestinal bacteria. *Environ. Microbiol.* **18**, 2117–2129.

Brown, S. and Babbitt, P. 2012. Inference of functional properties from large-scale analysis of enzyme superfamilies. *J. Biol. Chem.* **287**, 35–42.

Chassard, C. and Lacroix, C. 2013. Carbohydrates and the human gut microbiota. *Curr. Opin. Clin. Nutr. Metab. Care* **16**, 453–460.

Dong, J., Liang, Q., Niu, Y., Jiang, S., Zhou, L., Wang, J., Ma, C., and Kang, W. 2020. Effects of *Nigella sativa* seed polysaccharides on type 2 diabetic mice and gut microbiota. *Int. J. Biol. Macromol.* **159**, 725–738.

Gerlt, J.A., Bouvier, J.T., Davidson, D.B., Imker, H.J., Sadkhin, B., Slater, D.R., and Whalen, K.L. 2015. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): a web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta* **1854**, 1019–1037.

Hattori, M., Kanda, T., Shu, Y., Akao, T., Kobashi, K., and Namba, T. 1988a. Metabolism of barbaloin by intestinal bacteria. *Chem. Pharm. Bull.* **36**, 4462–4466.

Hattori, M., Shu, Y.Z., el-Sedawy, A.I., Namba, T., Kobashi, K., and Tomimori, T. 1988b. Metabolism of homoorientin by human intestinal bacteria. *J. Nat. Prod.* **51**, 874–878.

Hollman, P.C.H. 2001. Evidence for health benefits of plant phenols: local or systemic effects? *J. Sci. Food Agric.* **81**, 842–852.

Huang, L., Zhang, H., Wu, P., Entwistle, S., Li, X., Yohe, T., Yi, H., Yang, Z., and Yin, Y. 2018. dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation. *Nucleic Acids Res.* **46**, D516–D521.

Ito, T., Fujimoto, S., Shimosaka, M., and Taguchi, G. 2014. Production of *C*-glucosides of flavonoids and related compounds by *Escherichia coli* expressing buckwheat *C*-glucosyltransferase. *Plant Biotechnol.* **31**, 519–524.

Jin, J.S., Nishihata, T., Kakiuchi, N., and Hattori, M. 2008. Biotransformation of *C*-glucosylisoflavone puerarin to estrogenic (3*S*)-equol in co-culture of two human intestinal bacteria. *Biol. Pharm. Bull.* **31**, 1621–1625.

Jyotshna, Khare, P., and Shanker, K. 2016. Mangiferin: a review of

sources and interventions for biological activities. *Biofactors* **42**, 504–514.

Kim, M., Lee, J., and Han, J. 2015. Deglycosylation of isoflavone *C*-glycosides by newly isolated human intestinal bacteria. *J. Sci. Food Agric.* **95**, 1925–1931.

Kim, Y.S., Yeom, S.J., and Oh, D.K. 2011. Characterization of a GH3 family β-glucosidase from *Dictyoglomus turgidum* and its application to the hydrolysis of isoflavone glycosides in spent coffee grounds. *J. Agric. Food Chem.* **59**, 11812–11818.

Levin, B.J., Huang, Y.Y., Peck, S.C., Wei, Y., Martínez-del Campo, A., Marks, J.A., Franzosa, E.A., Huttenhower, C., and Balskus, E.P. 2017. A prominent glycyl radical enzyme in human gut microbiomes metabolizes *trans*-4-hydroxy-L-proline. *Science* **355**, eaai8386.

Ley, R.E., Turnbaugh, P.J., Klein, S., and Gordon, J.I. 2006. Human gut microbes associated with obesity. *Nature* **444**, 1022–1023.

Li, X., Xia, W., Bai, Y., Ma, R., Yang, H., Luo, H., and Shi, P. 2018. A novel thermostable GH3 *β*-glucosidase from *Talaromyce leycettanus* with broad substrate specificity and significant soybean isoflavone glycosides-hydrolyzing capability. *BioMed Res. Int.* **2018**, 4794690.

Nakamura, K., Komatsu, K., Hattori, M., and Iwashima, M. 2013. Enzymatic cleavage of the *C*-glucosidic bond of puerarin by three proteins, Mn^{2+}, and oxidized form of nicotinamide adenine dinucleotide. *Biol. Pharm. Bull.* **36**, 635–640.

Nakamura, K., Nishihata, T., Jin, J.S., Ma, C.M., Komatsu, K., Iwashima, M., and Hattori, M. 2011. The *C*-glucosyl bond of puerarin was cleaved hydrolytically by a human intestinal bacterium strain PUE to yield its aglycone daidzein and an intact glucose. *Chem. Pharm. Bull.* **59**, 23–27.

Nakamura, K., Zhu, S., Komatsu, K., Hattori, M., and Iwashima, M. 2019. Expression and characterization of the human intestinal bacterial enzyme which cleaves the *C*-glycosidic bond in 3″-oxopuerarin. *Biol. Pharm. Bull.* **42**, 417–423.

Nakamura, K., Zhu, S., Komatsu, K., Hattori, M., and Iwashima, M. 2020. Deglycosylation of the isoflavone *C*-glucoside puerarin by a combination of two recombinant bacterial enzymes and 3-oxoglucose. *Appl. Environ. Microbiol.* **86**, e00607-20.

Pei, X., Zhao, J., Cai, P., Sun, W., Ren, J., Wu, Q., Zhang, S., and Tian, C. 2016. Heterologous expression of a GH3 β-glucosidase from *Neurospora crassa* in *Pichia pastoris* with high purity and its application in the hydrolysis of soybean isoflavone glycosides. *Protein Expr. Purif.* **119**, 75–84.

Qu, X., Ding, B., Li, J., Liang, M., Du, L., Wei, Y., Huang, R., and Pang, H. 2020. Characterization of a GH3 halophilic β-glucosidase from *Pseudoalteromonas* and its NaCl-induced activity toward isoflavones. *Int. J. Biol. Macromol.* **164**, 1392–1398.

Seong, S.H., Roy, A., Jung, H.A., Jung, H.J., and Choi, J.S. 2016. Protein tyrosine phosphatase 1B and α-glucosidase inhibitory activities of *Pueraria lobata* root and its constituents. *J. Ethnopharmacol.* **194**, 706–716.

Thursby, E. and Juge, N. 2017. Introduction to the human gut microbiota. *Biochem. J.* **474**, 1823–1836.

Valles-Colomer, M., Falony, G., Darzi, Y., Tigchelaar, E., Wang, J., Tito, R.Y., Schiweck, C., Kurilshikov, A., Joossens, M., Wijmenga, C., *et al.* 2019. The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.* **4**, 623–632.

Wang, S., Zhang, S., Wang, S., Gao, P., and Dai, L. 2020. A comprehensive review on *Pueraria*: Insights on its chemistry and medicinal value. *Biomed. Pharmacother.* **131**, 110734.

Wei, B., Wang, Y.K., Qiu, W.H., Wang, S.J., Wu, Y.H., Xu, X.W., and Wang, H. 2020. Discovery and mechanism of intestinal bacteria in enzymatic cleavage of C-C glycosidic bonds. *Appl. Microbiol. Biotechnol.* **104**, 1883–1890.

Wei, B., Yang, W., Yan, Z.X., Zhang, Q.W., and Yan, R. 2018. Prenylflavonoids sanggenon C and kuwanon G from mulberry (*Morus alba* L.) as potent broad-spectrum bacterial β-glucuronidase inhibitors: biological evaluation and molecular docking studies. *J. Funct. Foods* **48**, 210–219.

Wei, J., Zhang, X.Y., Deng, S., Cao, L., Xue, Q.H., and Gao, J.M. 2017. α-Glucosidase inhibitors and phytotoxins from *Streptomyces xanthophaeus*. *Nat. Prod. Res.* **31**, 2062–2066.

Wu, H., Rebello, O., Crost, E.H., Owen, C.D., Walpole, S., Bennati-Granier, C., Ndeh, D., Monaco, S., Hicks, T., Colvile, A., *et al.* 2020. Fucosidases from the human gut symbiont *Ruminococcus gnavus*. *Cell. Mol. Life Sci.* **78**, 675–693.

Xu, J., Qian, D., Jiang, S., Guo, J., Shang, E., Duan, J., and Yang, J. 2014. Application of ultra-performance liquid chromatography coupled with quadrupole time-of-flight mass spectrometry to determine the metabolites of orientin produced by human intestinal bacteria. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **944**, 123–127.

Yoon, S.H., Ha, S.M., Kwon, S., Lim, J., Kim, Y., Seo, H., and Chun, J. 2017. Introducing EzBioCloud: a taxonomically united database of 16S rRNA and whole genome assemblies. *Int. J. Syst. Evol. Microbiol.* **67**, 1613–1617.

Zallot, R., Oberg, N., and Gerlt, J.A. 2019. The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. *Biochemistry* **58**, 4169–4182.

Zhou, T.S., Wei, B., He, M., Li, Y.S., Wang, Y.K., Wang, S.J., Chen, J.W., Zhang, H.W., Cui, Z.N., and Wang, H. 2020. Thiazolidin-2-cyanamides derivatives as novel potent *Escherichia coli* β-glucuronidase inhibitors and their structure-inhibitory activity relationships. *J. Enzyme Inhib. Med. Chem.* **35**, 1736–1742.