Journal of Bioscience and Bioengineering
VOL. 128 No. 3, 274–282, 2019

www.elsevier.com/locate/jbiosc

Evaluation of synthetic gene encoding α-galactosidase through metagenomic sequencing of paddy soil

Yu-Pei Chen,[1,2,3] Li-Ling Liaw,[4] Jong-Tar Kuo,[3] Hong-Tan Wu,[1,2] Guey-Horng Wang,[1,2] Xiu-Qin Chen,[2] Chai-Fang Tsai,[5] and Chiu-Chung Young[5,6,*]

Department of Medical Technology, Xiamen Medical College, Xiamen, 361023 Fujian, China,[1] Application Technique Engineering Center of Natural Cosmeceuticals, College of Fujian Province, Xiamen Medical College, Xiamen, 361023 Fujian, China,[2] Department of Biological Science and Technology, China University of Science and Technology, Taipei 115, Taiwan,[3] Bioresource Collection and Research Center, Food Industry Research and Development Institute, HsinChu 300, Taiwan,[4] Department of Soil and Environmental Sciences, National Chung Hsing University, Taichung 40227, Taiwan,[5] and Innovation and Development Center of Sustainable Agriculture, National Chung Hsing University, Taichung 40227, Taiwan[6]

Received 9 December 2018; accepted 9 March 2019
Available online 5 April 2019

Many genes of industrial relevance can be found in soil. In this study, metagenome sequencing of paddy soil was performed with 55.68 Gb sequences and 1,787,113 putative open reading frames (ORFs). The functional profiles and metabolic pathway of soil metagenomes were examined using Gene Ontology, Metagenomics RAST, and Kyoto Encyclopedia of Genes and Genomes. To verify the protein function and assembly of ORFs, a putative gene encoding α-galactosidase, namely GalR, which shares 65% identity with an unpublished glycoside hydrolase (GH) 27 family protein, was synthesized using its optimal codon for overexpression in *Escherichia coli*. GalR was successfully obtained and characterized. The optimal temperature and pH for GalR activity were 30°C and pH 9, respectively. Enzymatic activity indicated that GalR was alkaliphilic and different from acidophilic α-galactosidase in the GH 27 family. Furthermore, 50% of the relative activity of GalR can be attained for 1.7 and 0.7 h preincubation at 40°C and 50°C, respectively. Significant inhibition of GalR was observed in the presence of ethylenediaminetetraacetic acid (EDTA), $MgCl_2$, sodium dodecyl sulfate (SDS), and H_2O_2; however, it was resistant to 0.1% methanol and ethanol and was slightly activated with NaCl and KCl. The specific activity of GalR was achieved at 11.6 and 0.59 μmol/min/mg of protein using *p*-nitrophenyl-α-D-galactopyranoside and raffinose as substrates, respectively. Consequently, the metagenomic sequencing–based strategy can provide information for mining novel genes.

[**Key words:** Paddy soil; Metagenome; Glycoside hydrolases; Codon optimization; α-Galactosidase]

The organic matter and nitrogen availability in soil are crucial for maintaining soil quality and agricultural productivity. Substantial research is being conducted to identify the link between soil organic matter and soil structural stability for improving soil strength and reducing soil erosion (1). Additionally, soil microbes are known to play crucial roles in the functioning of ecosystems by cycling nutrients, degrading organic matter and pollutants, and improving soil fertility and crop health (2). Numerous studies have been performed to determine how soil microbial diversity can be affected and changed with soil management (3–5). Therefore, considerable research is directed toward the soil microbial community and diversity assessment using culture-independent approaches. Nevertheless, few studies have been devoted to exploring the genetic diversity and gene function of the soil ecosystem through metagenomic analysis. Recently, forest soil–derived lignocellulolytic microbial consortia revealed that an overrepresentation of predicted carbohydrate transporters (ATP-binding cassette, TonB-dependent transporter, and phosphotransferases), two-component sensing systems, and β-glucosidases/galactosidases can be observed in soil in which wheat straw was cultivated (6). Moreover, in soil with crop succession or rotation, hydrolases including lipases, laccases, cellulases, proteases, amylases, and pectinases were abundantly identified (7).

In the Carbohydrate-Active Enzyme Database (CAZy), the GH family is involved in the hydrolysis of various carbohydrates (8). Several natural forms of biomass, including cellulose, hemicellulose, starch, and other polysaccharides, can be catalyzed with glycoside hydrolases (GHs), such as cellulase, amylase, and galactosidase (9). Among these, α-galactosidases (α-D-galactoside galactohydrolases; EC 3.2.1.22), or melibiases, are involved in the hydrolysis of α-1,6-linked galactoside moieties from different oligosaccharides such as raffinose, stachyose, and melibiose as well as the guar gum and locust bean gum of polysaccharides (10). They widely exist in prokaryotes, eukaryotes, and archaea and have potential applications in the medical, food, and environmental fields for treating Fabry disease, improving the nutritive value of animal feeds, and enhancing pulp bleaching, respectively (10). In the CAZy database, α-galactosidases are present in GH 4, 27, 36, 57, 97, and 110 families as distinguished by conserved amino acid sequences (8). Prokaryotic α-galactosidases are extensively found in all these families, whereas eukaryotic α-galactosidases are found

* Corresponding author at: Department of Soil and Environment Sciences, National Chung Hsing University, 145 Xingda Rd., South Dist., Taichung City 40227, Taiwan. Tel.: +886 4 22840373 ext 4302; fax: +886 4 22861495.
E-mail address: ccyoung@mail.nchu.edu.tw (C.-C. Young).

1389-1723/$ – see front matter
https://doi.org/10.1016/j.jbiosc.2019.03.006

in GH 27 and 36 families. In addition, several 3D structures of α-galactosidases from humans, rice, yeasts, filamentous fungi, and bacteria were examined based on the Protein Data Bank database.

For functional metagenomics-based gene screening, the DNA library construction approach is widely used (9). Fosmid and cosmid libraries are usually used to explore novel genes. Lee et al. (11) reported that a halotolerant and protease-resistant α-galactosidase was identified from the gut metagenome of *Hermetia illucens*, a polyphagous insect, using a fosmid library with 92,000 individual fosmid clones. However, the technique is time consuming and depends on the screening platform and probability. Metagenomic sequencing is a powerful alternative to functional gene isolation for comprehending complex microbial gene diversity (12). In this study, 55.68 Gb of sequence was obtained from paddy soil that had a long-term application of organic compost using the Illumina Genome Analyzer. After deep sequencing of soil DNA, several putative open reading frames (ORFs) were identified and annotated. Gene Ontology (GO), Metagenomics RAST (MG-RAST), and the Kyoto Encyclopedia of Genes and Genomes (KEGG) of soil metagenomes were further used for the analysis of functional profiles and metabolic pathways. To verify the identified ORFs whether functional genes can be obtained, expressed and characterized from the soil metagenome, database search and ORF prediction with over 250 amino acids in length provided information for the putative glycoside hydrolase genes. A putative gene encoding α-galactosidase with a low identical match and e-value, and high query coverage was synthesized and overexpressed in *Escherichia coli*. Then, the activity and characterization of α-galactosidase were determined and described.

MATERIALS AND METHODS

Field site and soil properties The experimental soil, which had undergone annual paddy upland rotation for 23 years, was obtained from the Taiwan Agricultural Research Institute, Council of Agriculture, Executive Yuan (24°01′N, 120°41′E). It was Chikuatitso clay loam soil (Fluvaquentic Dystrochrepts) in which rice and maize had been cultivated in rotation from August 1995. In this study, agricultural soils with compost application were harvested during paddy cultivation. The surface soil (0–10 cm depth) from five plots was sampled and pooled for metagenomic analyses. The applied compost comprised hog and green manures; the green manures obtained from *Trifolium alexandrinum* and *Sesbania roxburghii* had been used for rice and maize crops, respectively. The soil was sieved using a 2-mm mesh after which the soil properties were analyzed. The pH and electrical conductivity values of the soil were determined. The extractable elements of soil, such as P^{5+}, K^{+}, Ca^{2+}, Mg^{2+}, Cu^{2+}, and Zn^{2+}, were measured using inductively coupled plasma emission spectroscopy (13).

Metagenomic analysis of the paddy soil To examine comprehensively the metagenome of the paddy soil, metagenomic sequencing was conducted through next-generation sequencing (NGS). Soil DNA with OD 260/280 in a range of 1.8–2.0 and quantification using the Quant-iT dsDNA BR assay (Invitrogen, Carlsbad, CA, USA) were checked on 0.6% agarose gel for Illumina Solexa analysis (Illumina, San Diego, CA, USA). Library construction, metagenomic sequencing, and genome *de novo* analysis were performed by the Sequencing-Tech company (Taipei, Taiwan). A genomic DNA library was constructed using the Illumina TruSeq DNA LT Sample Prep Kit. DNA was sheared on a Bioruptor for 16 cycles of power on for 30 s and power off for 30 s. The end-repair and A-tailing reactions were used to treat sheared DNA, which was followed by adapter ligation. The adaptor-ligated DNA was analyzed on 1.5% agarose gel to obtain DNA with size between 250 and 530 bp. Size-selected DNA was amplified with ten cycles and purified with AMPure XP beads (Agencourt Bioscience Corporation, Beverly, MA, USA). The quality of the DNA library was checked using the Quant-iT dsDNA HS assay (Invitrogen) and real-time polymerase chain reaction and was validated using the Experion Automated Electrophoresis System (BioRad, Hercules, CA, USA). DNA sequencing reads were processed to remove the adapter, short sequences (<35 bp), and low-quality bases (Q < 20) using Trimmomatic software for acquisition of desired reads (14). The metagenomes of microbial populations based on the sequencing reads were analyzed using the MG-RAST server for the taxonomic distribution of information (15). MEGAHIT software was used for *de novo* assembly of cleaned and filtered reads (16). Genes and ORFs were annotated using MetaGeneMark (17) and searched against the National Center for Biotechnology Information (NCBI) nonredundant database using BLASTP. Genes encoding α-galactosidases were identified for phylogeny analysis through the neighbor-joining method in MEGA-X software with 1000 bootstrap replicates (18).

Codon optimization analysis and overexpression plasmid construction of α-galactosidase A putative ORF encoding α-galactosidase was designated as GalR. The GalR gene was analyzed using OptimumGene software (GenScript Biotech Corp., Nanjing, China) for α-galactosidase overexpression in *E. coli*. The GalR and its optimal codon (GalR-opt) including the predicted signal sequence at 5′ sequence, and *Eco*RV adaptors at 5′ and 3′ sequences were synthesized and provided by GeneDireX Inc. (Las Vegas, NV, USA). The synthesized DNA products were introduced into the pET-30a (+) plasmid with the digestion of the *Eco*RV site to obtain overexpression plasmids, namely pGalR and pGalR-opt.

Overexpression of α-galactosidase in *E. coli* and recombinant α-galactosidase protein purification To overexpress the α-galactosidase protein, pGalR and pGalR-opt plasmids were transformed into competent cells of *E. coli* C43 (DE3). The transformants were incubated on a Luria–Bertani (LB) agar plate with 0.5 mM isopropy-β-D-thiogalactoside (IPTG) and 30 μg/mL kanamycin. A single colony was selected and cultured overnight in LB broth. The cultured broth of *E. coli* was transferred into fresh LB broth and then grown at 37°C until OD_{600} reached 0.5. The culture was induced with IPTG (0.5 mM) and incubated at 25°C overnight. *E. coli* pellets were harvested and suspended in buffer A (25 mM Tris–HCl and 150 mM NaCl, pH 7). The cells were disrupted using a JY92-IIN sonicator (Ningbo Scientz Biotechnology, Ningbo, China), and the debris was removed through centrifugation at 10,000 ×*g* for 15 min at 4°C. The supernatant containing α-galactosidase protein was purified using a Ni-NTA purification system (Invitrogen). Furthermore, a 3-kDa Amicon Ultra-15 Centrifugal Filter (Millipore, Bedford, MA, USA) was used for concentrating and replacing elution buffer with buffer A. Protein was identified through sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and Coomassie blue staining.

Western blotting analysis of α-galactosidase The α-galactosidase proteins from the transformants harboring pGalR and pGalR-opt plasmids were obtained and separated through SDS-PAGE. The manipulation of protein transfer and hybridization for western blotting analysis were performed as described by Sambrook and Russell (19). The SDS-PAGE gel was transferred onto a polyvinylidene difluoride membrane, and immunoblot analysis was conducted using a primary antibody against His-tag at a dilution of 1:1000 (BBI Life Sciences, Shanghai, China). Horseradish peroxidase-conjugated antibody to mouse immunoglobulin G (1:5000 dilution; West Grove, PA, USA) was then used. The protein bands were identified using a SuperSignal West Pico Kit (Thermo Fisher Scientific Pierce, IL, USA) according to the manufacturer's instructions.

Enzymatic activity assay of α-galactosidase Enzymatic activity was determined through the detection of *p*-nitrophenol released from *p*-nitrophenyl-α-D-galactopyranoside (pNPGal). The mixture used in the enzyme activity assay contained purified α-galactosidase and 10 mM pNPGal (dissolved in 25 mM Tris–HCl and 150 mM NaCl, pH 8), which were incubated at a constant temperature for 30 min. The reaction was terminated with 100°C incubation for 5 min. A spectrophotometer at OD_{420} was used for detecting the released *p*-nitrophenol. The optimal enzyme activity was determined to be 25°C–60°C. The pH profile of α-galactosidase activity was assayed between 4 and 10. For the assays at pH 4.0–5.0, pH 6.0–8.0, and pH 9.0–10, 50 mM sodium acetate/acetic acid buffer, 100 mM sodium phosphate, and glycine/NaOH buffers, respectively, were used. To determine the effects of different ions and reagents on α-galactosidase, various chemical compounds were used: EDTA (0–200 mM), NaCl (50 mM), KCl (50 mM), $CaCl_2$ (50 mM), $MgCl_2$ (50 mM), SDS (0.1%), H_2O_2 (0.1%), methanol (0.1%), and ethanol (0.1%). The thermostability of α-galactosidase was detected through incubation of enzymes at 40°C and 50°C from 0 to 8 h. The enzyme was immediately chilled, and the residual activity was assayed through the pNPGal method. To explore the α-galactosidase activity on raffinose, the reduced sugar obtained from the digestion of raffinose with α-galactosidase was estimated through the 3,5-dinitrosalicylic acid (DNS) method. The purified α-galactosidase and 1% raffinose (dissolved in 100 mM glycine/NaOH buffer, pH 9) were mixed at a constant temperature for 1 h and incubated. The reaction was interrupted with 100°C incubation for 5 min. The DNS solution (1% 3,5-dinitrosalicylic acid, 30% potassium sodium tartrate, and 1.5% NaOH) was added to the mixture for detecting the reducing sugars. The resulting mixture was heated in boiling water for 5 min, and reducing sugars were detected through a spectrophotometer at OD_{540}.

Nucleotide sequence accession number This metagenomic project was deposited at DDBJ/ENA/GenBank under accession numbers BioProject PRJNA498754 and BioSample SAMN10334111. The accession numbers of cloning vectors including pGalR and pGalR-opt have been submitted to the NCBI. The accession numbers are MK249730 and MK249731.

RESULTS

Soil properties The experimental soil was obtained from land that had undergone annual paddy upland rotation for 23 years

using compost as a fertilizer. The paddy soil was collected after the stem elongation stage of rice crops. The results of soil analyses were as follows. The pH and EC values of soil were 5.23 and 757 μscm^{-1}, respectively. Extractable elements of P^{5+}, K^{+}, Ca^{2+}, Mg^{2+}, Cu^{2+}, and Zn^{2+} were found to be 125.06, 136.80, 1327.67, 161.60, 1.19, and 8.63 mg kg^{-1} through inductively coupled plasma emission spectroscopy.

Soil metagenomic analysis Microbial structure and gene diversity were examined using high-throughput sequencing. After raw reads were trimmed and chimeric reads were removed, the average reads were 119 bp in length; furthermore,467,184,214 reads were obtained from the paddy soil (Table S1). The total bases after quality trim were 55.68 Gb. Because high sequence diversity was expected, the reads were assembled, and 1,856,411 contigs of length >100 bp could be perceived (Table S2). The taxonomic hit distribution of sequences was deduced using MG-RAST pipeline (15). Most of the domains were occupied with sequences of bacteria (96.5%), followed by archaea (2.0%), eukaryotes (1.3%), and viruses (0.05%) (Fig. 1). Among bacteria, Proteobacteria, Actinobacteria, and Firmicutes successively were the three dominant phyla, accounting for 57% of the total sequences. Proteobacteria was the most dominant phylum (38%). According to gene prediction and annotation by MetaGeneMark (17), 1,787,113 putative ORFs could be recognized (Table S3). Of the identified ORFs, 381,819 were >100 amino acids in length, whereas only nine ORFs were >1000 amino acids in length. The amino acids of the nine ORFs were deduced using the BLAST program from the NCBI. Six ORFs were 31%–100% identical to hypothetical proteins, and the other ORFs encoded carboxypeptidase regulatory-like domain protein, sugar-binding protein, and von Willebrand factor type A domain protein with 36%, 49%, and 59% similarities, respectively.

GO provides a representation of gene product properties, covering three domains, namely the biological process, cellar components, and molecular function (20). The metabolic process, membrane, and catalytic activity had the most abundant unigenes of the metagenome with 74,187, 22,125, and 106,588 genes, respectively, in the three domains (Fig. 2). Furthermore, the MG-RAST server was used for functional annotation of the sequenced genome (15). The results indicated that 689,590 genes of the paddy soil metagenome were functionally assigned in the MG-RAST server. The genes involved in carbohydrates were predominant (15.9%), followed by clustering-based subsystems (14.0%) and amino acids and derivatives (9.8%). Pathway-based analysis of the KEGG was used to explore the biological functions of different proteins (21). The main biochemical metabolism in which proteins participate was identified through pathway analysis. In this study,

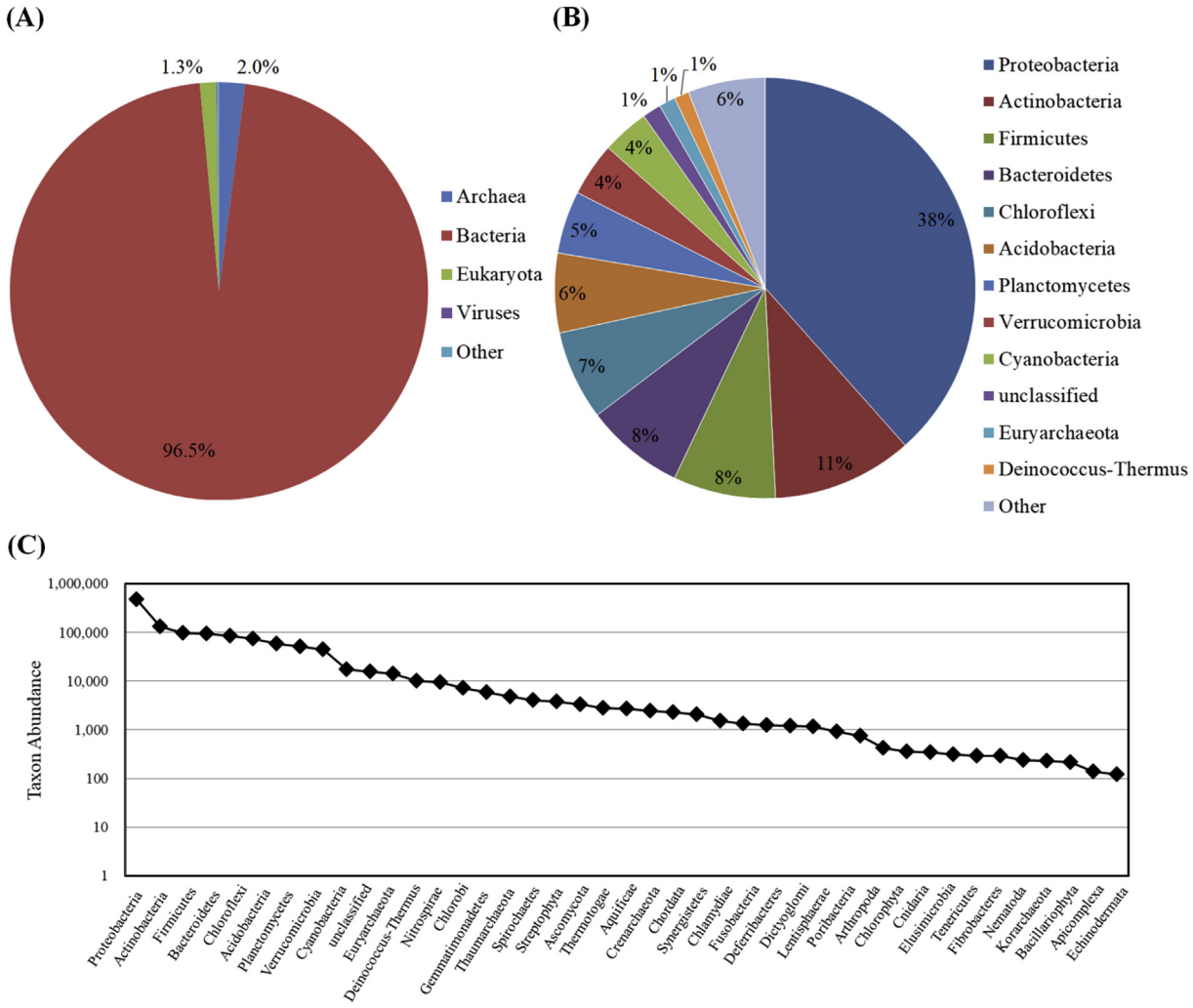

FIG. 1. Distribution of microbial community structures in paddy soil interpreted using MG-RAST server for domain and phylum taxonomy. The percentage of microbial community was classified according to the domain (A) and bacterial phylum (B). (C) The abundance of bacterial phylum was present in detail.

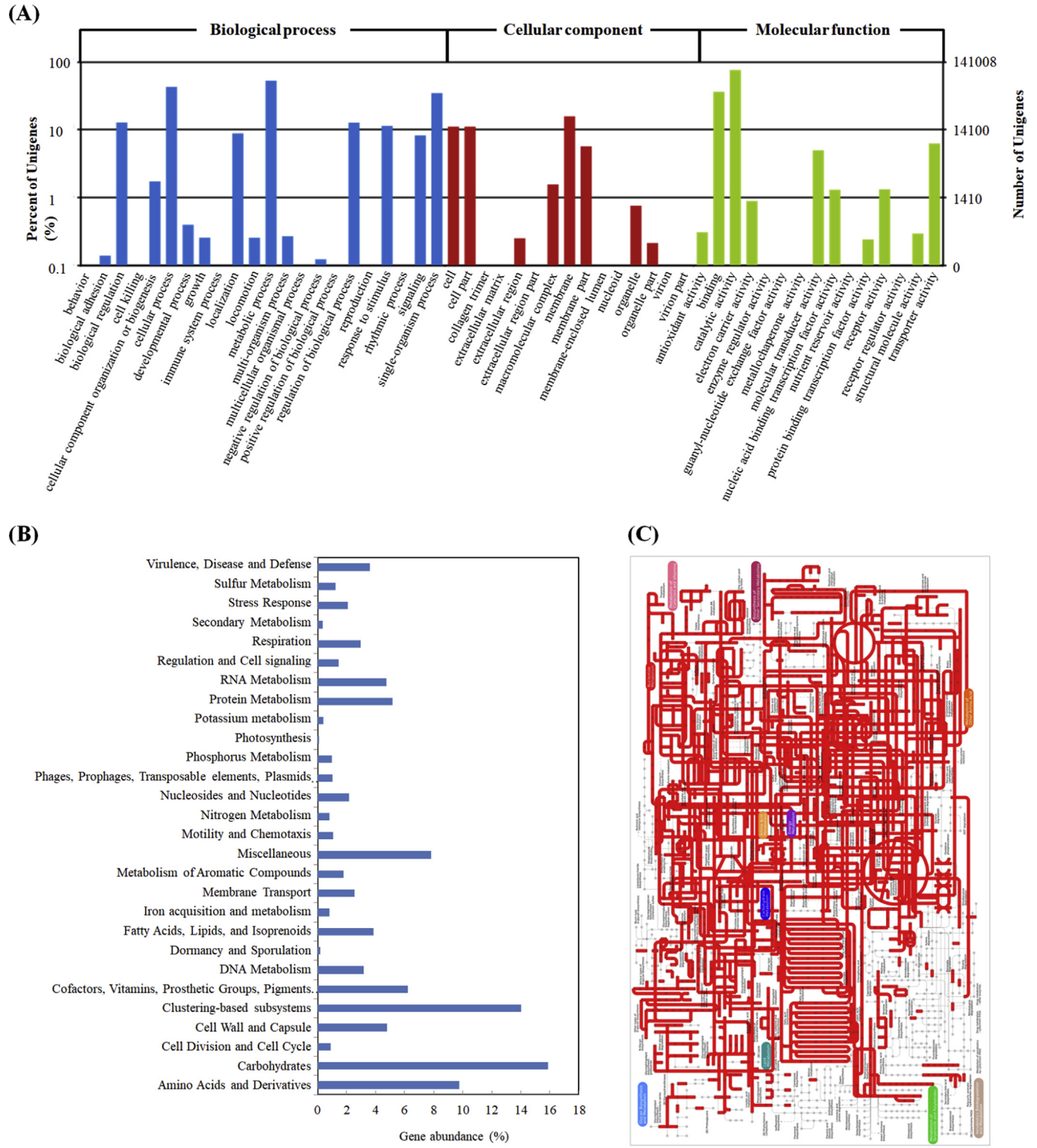

FIG. 2. Gene classifications of paddy soil metagenome for protein functional categories and metabolic pathway using (A) GO, (B) MG-RAST, and (C) KEGG analysis. KEGG pathway was implemented in Ipath (48). The detail information of KEGG pathway is provided in supplementary data (Table S4).

49.64% genes were assigned to metabolic pathways, followed by biosynthesis of secondary metabolites (26.41%; Table S4). In addition, several pathways involved in cellular processes, genetic information, environmental information, human diseases, and organismal systems were found in the paddy metagenome.

Construction and overexpression of recombinant α-galactosidase in *E. coli* GH can hydrolyze the glycosidic bond between carbohydrates such as cellulose, hemicellulose, and starch through the catalysis of cellulase and amylase (9). They are widely used in food, environmental science, and medicine. Therefore, in

this study, database search and ORF prediction provided information for the putative glycoside hydrolase genes (Table S5). Of the hypothetical GH, a putative ORF encoding α-galactosidase with 1122 bp, a low identical match and e-value and high query coverage was identified through the conserved domain search. It was designated as GalR with the 374 deduced amino acids, and the predicted molecular weight was 42.4 kDa. The amino acid of GalR genes was identified using the BLAST program from the NCBI; it had 65% identity with the GH 27 protein of *Maribacter polysiphoniae*, which was unpublished and uncharacterized. This was consistent with the phylogenetic analysis in which GalR was classified into the GH 27 family. Moreover, according to the conserved domain comparison of deduced amino acid sequences and the prediction of three-dimensional structure based on the SWISS-MODEL, the catalytic sites were hypothetically recognized as the 206th and 258th Asp (Figs. S1 and S2) (22).

To verify the enzymatic activity of GalR and codon usage bias in *E. coli*, the codons of GalR were optimized using OptimumGene for *E. coli* overexpression. The 204 codons of GalR were modified, and 20% difference of bases was observed between the original and optimal GalR (Fig. S3). The original and optimal GalR sequences were synthesized and introduced into the *Eco*RV site of a pET-30a (+) expression vector with a His-tag fusion of the N-terminus triggered by the T7 promoter to obtain the plasmids of pGalR and pGalR-opt. The pGalR and pGalR-opt plasmids were used for DNA transformation in *E. coli* C43 (DE3). The recombinant GalR proteins could be overexpressed with *E. coli* C43 (DE3) harboring pGalR and pGalR-opt on SDS-PAGE gel (Fig. 3). In addition, GalR fused with His-tag was demonstrated through western blotting analysis. *E. coli* C43 (DE3) harboring pGalR-opt expressed significantly more proteins than that harboring pGalR. The expected band of protein showed a 1.7-fold greater intensity in pGalR-opt expression than in pGalR expression. Thus, *E. coli* C43 (DE3) harboring pGalR-opt was used to achieve the purified protein.

Characterization of recombinant α-galactosidase The purified protein of GalR from the pGalR-opt expression was used to determine the enzymatic activities through pNPGal analysis. The optimal temperature and pH profiles of GalR activities are shown in Fig. 4. The optimal temperature of GalR was 30°C. However, the residual activity of GalR remained >23% at 50°C and 60°C. The optimal pH of GalR was 9 when glycine/NaOH buffer was used. The GalR activity drastically dropped when pH was <7 or >10. The residual activity of GalR was 2.3%–58.0% at pH 6–8 with sodium phosphate buffer. With the acetic buffer, GalR activity was almost lost at pH 4–5. To calculate the specific activity of α-galactosidase, *p*-nitrophenol was used as the standard. α-Galactosidase activity represents the production of 1 μmole of *p*-nitrophenol per min. According to the optimal condition, the specific activity of GalR was 11.6 μmol/min/mg of protein.

For wide industrial application, numerous studies have been performed to identify the thermal resistance of enzymes. Thermostability analysis of GalR was conducted for 8 h at 40°C and 50°C (Fig. 4). The GalR maintained 61% activity at 40°C after 60 min but only 25% at 50°C. The half-life, the time at which 50% activity is lost, of GalR was achieved at approximately 1.7 and 0.7 h at 40°C and 50°C, respectively, according to the regression calculation. However, the residual activity of GalR was attained for 21.8% at 40°C after 8 h of incubation. Furthermore, GalR was almost completely lost after 2 h of incubation at 50°C.

Effects of various reagents on α-galactosidase The effects of different metal ions, chemical reagents, and raffinose on GalR activity were determined. EDTA from 0 to 200 mM was used to explore the α-galactosidase activity of GalR (Fig. 5). The GalR activity gradually decreased with increasing EDTA, whereas enzymatic activity was rarely detectable at 200 mM EDTA. Moreover, GalR was significantly inhibited with 50 mM $MgCl_2$ and 0.1% SDS with 52.6% and 81.1% residual activities, respectively, whereas it was completely arrested with 0.1% H_2O_2. By contrast, the GalR activity was slightly activated with NaCl and KCl. Although the concentration of alcohol used was 0.1%, GalR seems to be highly resistant to alcohol.

Raffinose composed of α-1,6-linked galactoside moieties can be catalyzed to galactose and sucrose using α-galactosidase. In this study, raffinose was used as the substrate to examine the GalR activity (Fig. 6). According to the specific activity of α-galactosidase, D-galactose was used as the standard of reducing sugar. α-Galactosidase activity represents the production of 1 μmole of reducing sugars per min. GalR revealed low specific activity with 0.59 μmol/min/mg of protein. Additionally, the temperature profile of GalR treated with raffinose was similar to that revealed by pNPGal analysis. Addition of 50 mM EDTA and 0.1% SDS also can inhibit the enzymatic activity of GalR.

DISCUSSION

High-throughput sequencing through NGS, which can provide a large amount of information at the sequence level, has been applied in various metagenomes, such as agricultural soil, insects, and the human gut, to interpret the microbial diversity directly (6,7,11,23).

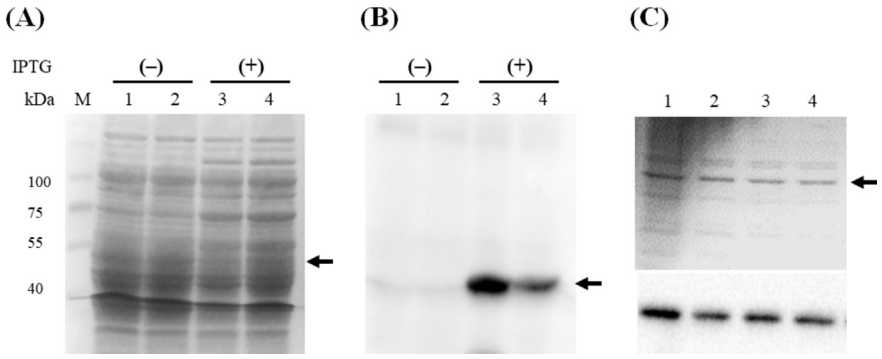

FIG. 3. SDS-PAGE and western blotting analysis of GalR proteins from *E. coli* C43 (DE3) harboring pGalR and pGalR-opt plasmids. (A) SDS-PAGE gel was stained with Coomassie blue. Lanes 1 and 3 indicate GalR protein obtained from soluble cell-free extract with optimal codon in absence (–) and presence (+) of IPTG at 0.5 mM. Lanes 2 and 4 indicate original GalR protein obtained from soluble cell-free extract in absence (–) and presence (+) of IPTG at 0.5 mM, respectively. Arrow indicates the predicted molecular weight of recombinant GalR with His-tag. (B) Western blotting analysis of original and optimal GalR protein obtained from soluble cell-free extract with and without 0.5 mM IPTG. Lanes 1–4 are the same with the SDS-PAGE gel stained with Coomassie blue. (C) Major band on SDS-PAGE gel was observed (upper), suggesting α-galactosidase purification through nickel column. Arrow indicates purified proteins. Western blotting analysis further confirmed purified proteins (lower).

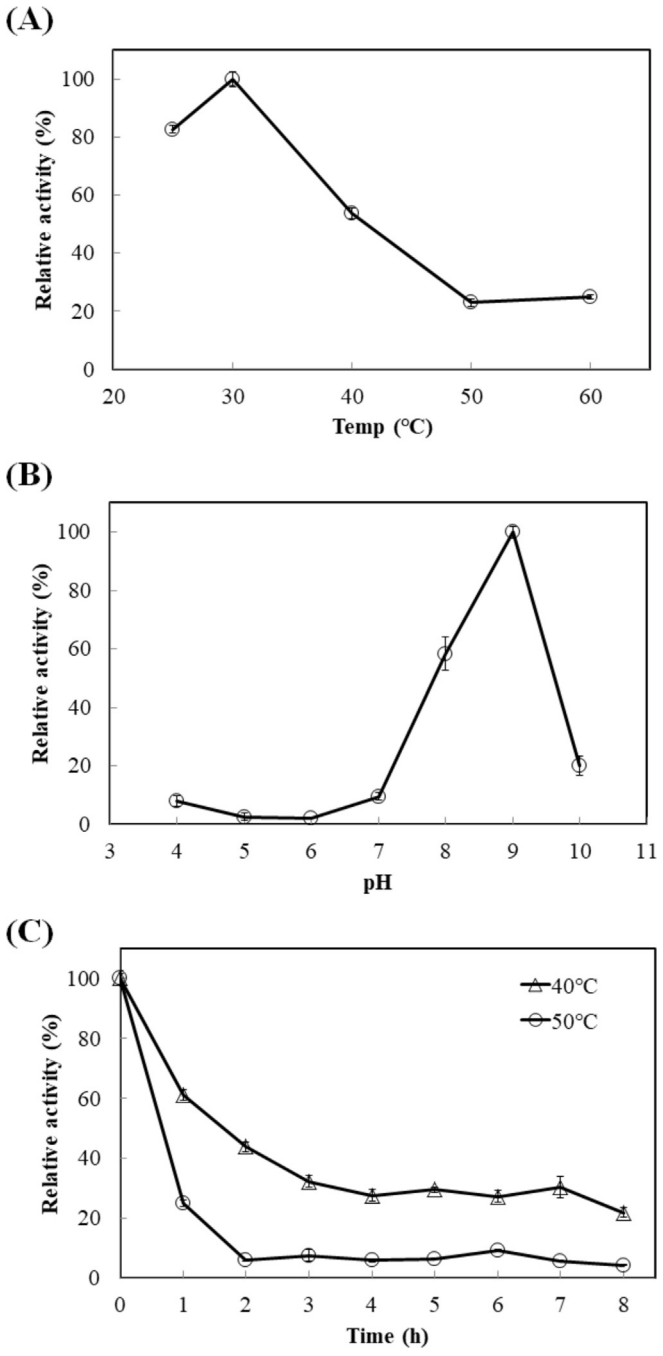

FIG. 4. Optimal temperature, pH, and thermostability analyses of GalR. (A) Temperature profiling of GalR was performed at 25°C-60°C and pH 8 with Tris–HCl buffer for 30 min. (B) The pH profiling of GalR was performed at 30°C with 50 mM sodium acetate/acetic acid buffer (pH 4–5), 100 mM sodium phosphate buffer (pH 6–8), and glycine/NaOH buffer (pH 9–10) for 30 min. (C) Thermostability analysis of GalR. GalR was incubated at 40°C and 50°C for 8 h. Remaining enzyme activity was measured at 30°C with glycine/NaOH buffer (pH 9). All enzymatic reactions were terminated using 100°C incubation for 5 min. A spectrophotometer at OD_{420} was used for detection of released *p*-nitrophenol.

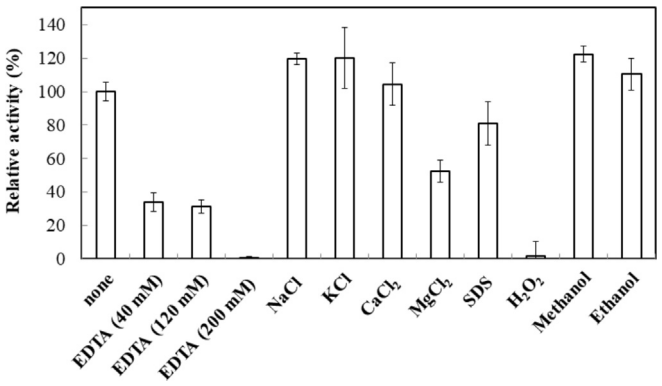

FIG. 5. Effect of various reagents on GalR activity. EDTA (40–200 mM) and various reagents (50 mM NaCl, KCl, $CaCl_2$, and $MgCl_2$; 0.1% SDS, H_2O_2, methanol, and ethanol) were added into glycine/NaOH buffer (pH 9) for GalR activity analysis at 30°C. Enzymatic activity was determined through pNPGal method at OD_{420}.

In this study, an in-depth analysis of paddy soil microbial sequencing was conducted with 55.68 Gb sequences using the Illumina Genome Analyzer, and the metagenome sequences were submitted to the NCBI database. The total bases and predicted genes of the soil metagenome were 12,000- and 397-fold greater than that of *E. coli* K12 with 4.64 Mb and 4498 genes (24). The sequences of the soil metagenome relating to Proteobacteria were determined to be the most abundant Eubacteria (Fig. 1), corresponding to previous characterizations of the different microbial consortia through 16S rRNA analysis (4,5). At the genus level, a large number of reads (2.4%) associated with *Candidatus Solibacter*, a member of the phylum Acidobacteria, were recognized, followed by *Candidatus Koribacter*, *Geobacter*, *Streptomyces*, *Bradyrhizobium*, and *Bacillus*. *Candidatus Solibacter* and *Candidatus Koribacter* are usually predominant in contaminated sites (25). In paddy soil with long-term application of organic compost, they could play a crucial role in decomposing organic matter for the carbon cycle. The relative abundance of the main nitrogen-fixing bacterial genera including *Bradyrhizobium*, *Bacillus*, *Burkholderia*, *Pseudomonas*, *Rhizobium*, *Acinetobacter*, *Mesorhizobium*, *Paenibacillus*, *Achromobacter*, *Agrobacterium*, and *Stenotrophomonas* was also observed in the metagenome (26). In addition, all sequences of viruses were derived from unclassified species with 0.05%. In Eukaryota, fungi are the most dominant category, the largest phylum of which is Ascomycota (0.3%) (27). However, 0.15% of fungi belonged to unclassified species in the metagenomic analysis.

With respect to the functional profiles, GO, MG-RAST, and KEGG were performed (Fig. 2). The significantly enriched GO terms were cellular, metabolic, and single-organism of the biological process as well as cell, cell part, and membrane of the cellular component together with binding and catalytic activity of the molecular function. These remarkable terms were the same as the soil rhizosphere microbial communities of the transgenic soybean line ZUTS31 (26). This indicated that these GO terms were essential and abundant in the microbial structure of agricultural soil. According to functional annotation using MG-RAST, sequences corresponding to carbohydrates were dominant. Of the identified sequences, genes involved in one-carbon, di-carbon, and oligosaccharide metabolisms such as decarboxylase, UDP-glucose 4-epimerase, and galactosidase, which provide energy for supporting life, were relatively abundant (28). Moreover, several CO_2 fixation genes associated with the Calvin–Benson cycle and photorespiration could be identified and may be derived from Cyanobacteria. Nitrogen cycling, including nitrogen fixation, nitrification, and denitrification, is essential for limiting the nutrient supply for microbiological processes and rice production (29). In this study, FixA and FixB required for free-living microaerophilic nitrogen fixation in *Bradyrhizobium* could be found based on the KEGG pathway (30). On the other hand, the functional genes of ammonia mono-oxygenase, nitrite oxidoreductase, nitrate reductase, and nitrite reductase were used as molecular markers to examine the dynamics of nitrifier and denitrifier communities (29). However,

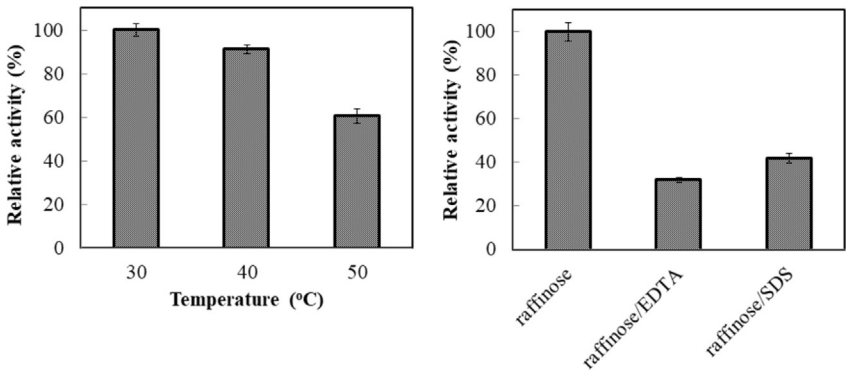

FIG. 6. Effect of raffinose on GalR activity. GalR and 1% raffinose were mixed at different temperatures and with different reagents (0.1% EDTA and SDS) at pH 9 (glycine/NaOH buffer) for 1 h. Enzymatic activity was detected through DNS method. Reaction was heated in boiling water for 5 min, and reducing sugars released were measured using a spectrophotometer at OD_{540}.

only nitrate and nitrite reductases were exhibited in the paddy soil metagenome. This suggested that spatial and geochemical factors could result in the uneven distribution of ammonium oxidizers (31). Furthermore, a few sequences related to infectious diseases corresponding to amoebiasis, toxoplasmosis, African trypanosomiasis, Chagas disease, leishmaniasis, and malaria were detected. Nevertheless, a complete pathway of infectious diseases was absent.

The GH family has potential applications in various industries. A putative ORF encoding α-galactosidase, namely GalR, with a low identical match and e-value and high query coverage in the metagenome was selected because it demonstrated considerable activity. GalR shared 65% identity with the GH 27 protein of *M. polysiphoniae*, which was not experimentally characterized. Much research has characterized GH 27 α-galactosidase in Eukaryota, such as *Hericium erinaceus* and *Fusarium oxysporum* fungi, which usually have a small molecular weight and are monomeric proteins (32,33). According to the Naumoff's study (34), GH 27 can be classified into three major subfamilies (GH 27a, GH 27b, and GH 27c). GH 27a enzymes containing four cysteine residues in the N-terminus are highly conserved to form two disulfide bridges for the maintenance of catalytic structure (33). However, only three cysteine residues were observed in GalR, and no disulfide bridge could be determined according to the DISULFIND prediction (35). In addition, on the basis of the *in silico* prediction of SignalP 4.1 Server, the first 25 amino acids of GalR with the high hydrophobic amino acids were assumed to be the signal secretion sequence (36). The C-terminal domain of GH 27 containing a β-sandwich structure was less conserved and its function was still unknown (37). In this study, based on the prediction of protein structure, GalR did not have a C-terminal β-sandwich domain (Fig. S2). Nevertheless, the hydrolysis activity of α-1,6-linked galactoside moieties from galacto-oligosaccharides can be identified in GalR. This suggested that a β-sandwich domain did not involve in the enzymatic function of GalR.

Many studies have been performed regarding the heterologous overexpression of α-galactosidase in *E. coli*. However, incompatibility of codon usage bias can cause a decrease in recombinant protein productivity or protein misfolding, resulting in the formation of an inclusion body (38). In this study, to improve the protein productivity, a gene optimization program using synonymous codon substitution by OptimumGene was performed for the GalR design. Our result revealed that the optimal GalR accounted for >1.7 times the original GalR protein content with high translational efficiency, which might result from the codon usage bias in *E. coli* (Fig. 3) (38). According to our review of relevant literature, five reports have been published on the characterization and overexpression of GH 27 α-galactosidases in *E. coli* (33,39—42). The optimal temperature and pH of the six reported α-galactosidases were detected at 35°C—50°C and pH 4—8.2. Four of these α-galactosidases from *Bacteroides fragilis*, *F. oxysporum*, and rice were acidophilic between pH 4 and 5. By contrast, our study indicated that GalR was alkaliphilic at the optimal pH 9 (Fig. 4). The decrease of lysine-aspartate ion pairs replaced by the increase of arginine-glutamate ion pairs and high hydrophobic contacts were important for stabilizing proteins in alkaline media (43). Comparing the amino acid composition between these α-galactosidases in GH 27, the lysine frequency of GalR was lower than that of *B. fragilis*, *F. oxysporum* and rice (α-Gal III). Furthermore, GalR showed high arginine and glutamate residues. The 36.4% frequency of hydrophobic residues, including Ala, Ile, Leu, Phe, Trp, and Val in GalR, was higher than that of the other reported α-galactosidases. These factors may reinforce the interactions of protein residues and contribute to enzymatic stabilization in alkaline conditions (43).

The 50% relative activity of GalR was attained for 1.7 and 0.7 h preincubation at 40°C and 50°C. The result showed higher thermostability than those of previous studies in GH 27 α-galactosidases (39,42). This may result from the thermostability characterizations containing nonpolar residues (Val, Ala, Leu, and Ile) and aromatic residues (Phe and Trp) for enhancing hydrophobicity (44). In addition, determination of the enzymatic sensitivity to different reagents, and metal ions indicated that GalR was gradually inhibited with the increase of EDTA, implying the involvement of metal ions in the enzymatic activity (Fig. 5). This was consistent with a GH 27 α-galactosidase from *B. fragilis*, whereas the enzymatic activity of several GH 36 α-galactosidases from *Bacillus coagulans*, *Bacillus megaterium*, and *Sulfolobus solfataricus* was not altered in the presence of EDTA (40,45—47). It suggested that the different GH families indeed display different enzymatic characteristics. Moreover, Ca^{2+} had no significant effect on GalR, whereas Na^{+} and K^{+} slightly promoted enzyme activity. This was similar to the property of α-galactosidase from *F. oxysporum* (33). GalR had higher tolerance to organic solvents of methanol and ethanol than to Mg^{2+}, H_2O_2, and SDS. Furthermore, GalR was also active against raffinose (Fig. 6), but its activity was lower than that of pNPGal utilization and was similar to that of other GH 27 enzymes (33,39,41).

In summary, metagenome sequencing of paddy soil was performed with 55.68 Gb sequences, and subsequent data mining was performed for functional profiles and metabolic pathways. Several GH family proteins, such as amylases, cellulases, xylanases, and galactosidases, were explored in the paddy soil metagenome, and they have the potential to be applied in various industrial fields. To demonstrate the functional property of the assembly ORFs, a

putative gene encoding α-galactosidase, namely GalR, was successfully synthesized using its optimal codon and overexpressed in *E. coli*. The alkaliphilic GalR was characterized in detail. The strategy of the metagenomic sequencing–based gene identification can not only discover various novel genes but also save time compared with screening DNA libraries.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbiosc.2019.03.006.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Science and Technology, Taiwan (Grant 102-2313-B-157-003-MY3), the Doctor Startup Fund Project of Xiamen Medical College (K2016-31), Fujian, China, and the Innovation and Development Center of Sustainable Agriculture from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

References

1. **Bronick, C. J. and Lal, R.:** Soil structure and management: a review, Geoderma, **124**, 3–22 (2005).
2. **Jacobsen, C. S. and Hjelmso, M. H.:** Agricultural soils, pesticides and microbial diversity, Curr. Opin. Biotechnol., **27**, 15–20 (2014).
3. **Maron, P. A., Mougel, C., and Ranjard, L.:** Soil microbial diversity: methodological strategy, spatial overview and functional interest, C. R. Biol., **334**, 403–411 (2011).
4. **He, J.-Z., Zheng, Y., Chen, C.-R., He, Y.-Q., and Zhang, L.-M.:** Microbial composition and diversity of an upland red soil under long-term fertilization treatments as revealed by culture-dependent and culture-independent approaches, J. Soils Sediments, **8**, 349–358 (2008).
5. **Souza, R. C., Cantão, M. E., Vasconcelos, A. T. R., Nogueira, M. A., and Hungria, M.:** Soil metagenomics reveals differences under conventional and no-tillage with crop rotation or succession, Appl. Soil Ecol., **72**, 49–61 (2013).
6. **Jimenez, D. J., Chaves-Moreno, D., and van Elsas, J. D.:** Unveiling the metabolic potential of two soil-derived microbial consortia selected on wheat straw, Sci. Rep., **5**, 13845 (2015).
7. **Souza, R. C., Cantao, M. E., Nogueira, M. A., Vasconcelos, A. T. R., and Hungria, M.:** Outstanding impact of soil tillage on the abundance of soil hydrolases revealed by a metagenomic approach, Braz. J. Microbiol., **49**, 723–730 (2018).
8. **Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B.:** The carbohydrate-active enzymes database (CAZy) in 2013, Nucleic Acids Res., **42**, D490–D495 (2014).
9. **Sathya, T. A. and Khan, M.:** Diversity of glycosyl hydrolase enzymes from metagenome and their application in food industry, J. Food Sci., **79**, R2149–R2156 (2014).
10. **Katrolia, P., Rajashekhara, E., Yan, Q., and Jiang, Z.:** Biotechnological potential of microbial α-galactosidases, Crit. Rev. Biotechnol., **34**, 307–317 (2014).
11. **Lee, C. M., Kim, S. Y., Song, J., Lee, Y. S., Sim, J. S., and Hahn, B. S.:** Isolation and characterization of a halotolerant and protease-resistant α-galactosidase from the gut metagenome of *Hermetia illucens*, J. Biotechnol., **279**, 47–54 (2018).
12. **Pozzo, T., Higdon, S. M., Pattathil, S., Hahn, M. G., and Bennett, A. B.:** Characterization of novel glycosyl hydrolases discovered by cell wall glycan directed monoclonal antibody screening and metagenome analysis of maize aerial root mucilage, PLoS One, **13**, e0204525 (2018).
13. **Hendershot, W. H., Lalande, H., Reyes, D., and MacDonald, D.:** Trace element assessment, pp. 109–120, in: Carter, M. R. and Gregorich, E. G. (Eds.), Soil sampling and methods of analysis, 2nd ed. CRC Press Taylor & Francis, Boca Raton, FL (2008).
14. **Bolger, A. M., Lohse, M., and Usadel, B.:** Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics, **30**, 2114–2120 (2014).
15. **Keegan, K. P., Glass, E. M., and Meyer, F.:** MG-RAST, a metagenomics service for analysis of microbial community structure and function, Methods Mol. Biol., **1399**, 207–233 (2016).
16. **Li, D., Luo, R., Liu, C. M., Leung, C. M., Ting, H. F., Sadakane, K., Yamashita, H., and Lam, T. W.:** MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices, Methods, **102**, 3–11 (2016).
17. **Zhu, W., Lomsadze, A., and Borodovsky, M.:** Ab initio gene identification in metagenomic sequences, Nucleic Acids Res., **38**, e132 (2010).
18. **Kumar, S., Stecher, G., and Tamura, K.:** MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets, Mol. Biol. Evol., **33**, 1870–1874 (2016).
19. **Sambrook, J. and Russell, D. W.:**, in: Molecular cloning: a laboratory manual, 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. (2001).
20. **Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., and other 10 authors:** Gene ontology: tool for the unification of biology. The gene ontology consortium, Nat. Genet., **25**, 25–29 (2000).
21. **Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M.:** From genomics to chemical genomics: new developments in KEGG, Nucleic Acids Res., **34**, D354–D357 (2006).
22. **Lansky, S., Salama, R., Solomon, H. V., Feinberg, H., Belrhali, H., Shoham, Y., and Shoham, G.:** Structure-specificity relationships in Abp, a GH27 β-L-arabinopyranosidase from *Geobacillus stearothermophilus* T6, Acta Crystallogr. D Biol. Crystallogr., **70**, 2994–3012 (2014).
23. **Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., and other 44 authors:** A human gut microbial gene catalogue established by metagenomic sequencing, Nature, **464**, 59–65 (2010).
24. **Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., and other 7 authors:** The complete genome sequence of *Escherichia coli* K-12, Science, **277**, 1453–1462 (1997).
25. **Wang, H., Zeng, Y., Guo, C., Bao, Y., Lu, G., Reinfelder, J. R., and Dang, Z.:** Bacterial, archaeal, and fungal community responses to acid mine drainage-laden pollution in a rice paddy soil ecosystem, Sci. Total Environ., **616–617**, 107–116 (2018).
26. **Lu, G. H., Hua, X. M., Cheng, J., Zhu, Y. L., Wang, G. H., Pang, Y. J., Yang, R. W., Zhang, L., Shou, H., Wang, X. M., Qi, J., and Yang, Y. H.:** Impact of glyphosate on the rhizosphere microbial communities of an EPSPS-transgenic soybean line ZUTS31 by metagenome sequencing, Curr. Genom., **19**, 36–49 (2018).
27. **Moll, J., Hoppe, B., Konig, S., Wubet, T., Buscot, F., and Kruger, D.:** Spatial distribution of fungal communities in an arable soil, PLoS One, **11**, e0148130 (2016).
28. **Diniz, R. H., Silveira, W. B., Fietto, L. G., and Passos, F. M.:** The high fermentative metabolism of *Kluyveromyces marxianus* UFV-3 relies on the increased expression of key lactose metabolic enzymes, Antonie Van Leeuwenhoek, **101**, 541–550 (2012).
29. **Wang, H., Li, X., Wang, J., and Zhang, H.:** Changes of microbial population and N-cycling function genes with depth in three Chinese paddy soils, PLoS One, **12**, e0189506 (2017).
30. **Weidenhaupt, M., Rossi, P., Beck, C., Fischer, H. M., and Hennecke, H.:** *Bradyrhizobium japonicum* possesses two discrete sets of electron transfer flavoprotein genes: *fixA*, *fixB* and *etfS*, *etfL*, Arch. Microbiol., **165**, 169–178 (1996).
31. **Stempfhuber, B., Engel, M., Fischer, D., Neskovic-Prit, G., Wubet, T., Schoning, I., Gubry-Rangin, C., Kublik, S., Schloter-Hai, B., Rattei, T., and other 6 authors:** pH as a driver for ammonia-oxidizing archaea in forest soils, Microb. Ecol., **69**, 879–883 (2015).
32. **Ye, F., Geng, X. R., Xu, L. J., Chang, M. C., Feng, C. P., and Meng, J. L.:** Purification and characterization of a novel protease-resistant GH27 α-galactosidase from *Hericium erinaceus*, Int. J. Biol. Macromol., **120**, 2165–2174 (2018).
33. **Maruta, A., Yamane, M., Matsubara, M., Suzuki, S., Nakazawa, M., Ueda, M., and Sakamoto, T.:** A novel α-galactosidase from *Fusarium oxysporum* and its application in determining the structure of the gum Arabic side chain, Enzyme Microb. Technol., **103**, 25–33 (2017).
34. **Naumoff, D. G.:** Phylogenetic analysis of α-galactosidases of the GH27 family, Mol. Biol., **38**, 388–399 (2004).
35. **Ceroni, A., Passerini, A., Vullo, A., and Frasconi, P.:** DISULFIND: a disulfide bonding state and cysteine connectivity prediction server, Nucleic Acids Res., **34**, W177–W181 (2006).
36. **Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H.:** SignalP 4.0: discriminating signal peptides from transmembrane regions, Nat. Methods, **8**, 785–786 (2011).
37. **Fernandez-Leiro, R., Pereira-Rodriguez, A., Cerdan, M. E., Becerra, M., and Sanz-Aparicio, J.:** Structural analysis of *Saccharomyces cerevisiae* α-galactosidase and its complexes with natural substrates reveals new insights into substrate specificity of GH27 glycosidases, J. Biol. Chem., **285**, 28020–28033 (2010).
38. **Wang, Y., Jia, P., Sharif, R., Li, Z., Li, Y., and Chen, P.:** High-level production of DNA-specific endonuclease AsEndI with synonymous codon and its potential utilization for removing DNA contamination, Appl. Biochem. Biotechnol., **185**, 641–654 (2018).
39. **Zhou, J., Liu, Y., Lu, Q., Zhang, R., Wu, Q., Li, C., Li, J., Tang, X., Xu, B., Ding, J., Han, N., and Huang, Z.:** Characterization of a glycoside hydrolase family 27 α-galactosidase from *Pontibacter* reveals its novel salt-protease tolerance and transglycosylation activity, J. Agric. Food Chem., **64**, 2315–2324 (2016).
40. **Gong, W., Xu, L., Gu, G., Lu, L., and Xiao, M.:** Efficient and regioselective synthesis of globotriose by a novel α-galactosidase from *Bacteroides fragilis*, Appl. Microbiol. Biotechnol., **100**, 6693–6702 (2016).

41. **Li, S., Kim, W. D., Kaneko, S., Prema, P. A., Nakajima, M., and Kobayashi, H.:** Expression of rice (*Oryza sativa* L. var. Nipponbare) α-galactosidase genes in *Escherichia coli* and characterization, Biosci. Biotechnol. Biochem., **71**, 520–526 (2007).
42. **Halstead, J. R., Fransen, M. P., Eberhart, R. Y., Park, A. J., Gilbert, H. J., and Hazlewood, G. P.:** α-Galactosidase A from *Pseudomonas fluorescens* subsp. *cellulosa*: cloning, high level expression and its role in galactomannan hydrolysis, FEMS Microbiol. Lett., **192**, 197–203 (2000).
43. **Popinako, A., Antonov, M., Tikhonov, A., Tikhonova, T., and Popov, V.:** Structural adaptations of octaheme nitrite reductases from haloalkaliphilic *Thioalkalivibrio* bacteria to alkaline pH and high salinity, PLoS One, **12**, e0177392 (2017).
44. **Zhou, X. X., Wang, Y. B., Pan, Y. J., and Li, W. F.:** Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins, Amino Acids, **34**, 25–33 (2008).
45. **Zhao, R., Tu, Y., Zhang, X., Deng, L., and Chen, X.:** A novel α-galactosidase from the thermophilic probiotic *Bacillus coagulans* with remarkable protease-resistance and high hydrolytic activity, PLoS One, **13**, e0197067 (2018).
46. **Huang, Y., Zhang, H., Ben, P., Duan, Y., Lu, M., Li, Z., and Cui, Z.:** Characterization of a novel GH36 α-galactosidase from *Bacillus megaterium* and its application in degradation of raffinose family oligosaccharides, Int. J. Biol. Macromol., **108**, 98–104 (2018).
47. **Brouns, S. J., Smits, N., Wu, H., Snijders, A. P., Wright, P. C., de Vos, W. M., and van der Oost, J.:** Identification of a novel α-galactosidase from the hyperthermophilic archaeon *Sulfolobus solfataricus*, J. Bacteriol., **188**, 2392–2399 (2006).
48. **Letunic, I., Yamada, T., Kanehisa, M., and Bork, P.:** iPath: interactive exploration of biochemical pathways and networks, Trends Biochem. Sci., **33**, 101–103 (2008).