BMC Bioinformatics

RESEARCH

Open Access

Functional glyco-metagenomics elucidates the role of glycan-related genes in environments

Hayato Takihara[1], Nobuaki Miura[1], Kiyoko F. Aoki-Kinoshita[2] and Shujiro Okuda[1*]

*Correspondence:
okd@med.niigata-u.ac.jp
[1] Division of Bioinformatics, Niigata University Graduate School of Medical and Dental Sciences, 1-757 Asahimachi-dori, Chuo-ku, Niigata 951-8510, Japan
Full list of author information is available at the end of the article

Abstract

Background: Glycan-related genes play a fundamental role in various processes for energy acquisition and homeostasis maintenance while adapting to the environment in which the organism exists; however, their role in the microbiome in the environment is unclear.

Methods: Sequence alignment was performed between known glycan-related genes and complete genomes of microorganisms, and optimal parameters for identifying glycan-related genes were determined based on the alignments. Using the constructed scheme (> 90% of identity and > 25 aa of alignment length), glycan-related genes in various environments were identified from 198 different metagenome data.

Results: As a result, we identified 86.73 million glycan-related genes from the metagenome data. Among the 12 environments classified in this study, the percentage of glycan-related genes was high in the human-associated environment, suggesting that these environments utilize glycan metabolism better than other environments. On the other hand, the relative abundances of both glycoside hydrolases and glycosyltransferases surprisingly had a coverage of over 80% in all the environments. These glycoside hydrolases and glycosyltransferases were classified into two groups of (1) general enzyme families identified in various environments and (2) specific enzymes found only in certain environments. The general enzyme families were mostly from genes involved in monosaccharide metabolism, and most of the specific enzymes were polysaccharide degrading enzymes.

Conclusion: These findings suggest that environmental microorganisms could change the composition of their glycan-related genes to adapt the processes involved in acquiring energy from glycans in their environments. Our functional glyco-metagenomics approach has made it possible to clarify the relationship between the environment and genes from the perspective of carbohydrates, and the existence of glycan-related genes that exist specifically in the environment.

Keywords: Functional glyco-metagenomics, Metagenome, Microbiome, Environment, Glycan-related genes, Glycan

Background

Genome sequencing has been carried out for the past two decades with respect to unicellular and multicellular microorganisms, as well as microbial communities isolated from a variety of environments, including the ocean [1, 2], soil [3], animal digestive tracts [4], and humans [5, 6]. Metagenomics elucidates the nature of life through the exploration of the genetic content of various bacteria from the reads of samples taken from different environments [7–11]. The issue at present is not to obtain more sequence data, but to infer the functions of the myriad of proteins already identified [12]. Although next-generation sequencing has led to an explosion of sequence data, our functional understanding of the data is still lacking [11, 13]. This bias is caused by the lack of diverse lineages and genetic compositions of the microbiome, basically due to the fact that genome sequences registered in databases tend to be biased toward culturable species.

Functional metagenomics is known as a method to search for functional genes in microbial communities based on the results of metagenomic sequencing and experimental screenings [14, 15]. Functional metagenomics has been mostly conducted on intestinal bacteria, it is used to infer what kind of phenomenon in the environment by associating metagenome information related to disease with known metabolic information [7, 8]. Databases that organize information on various functional genes in a broad range of species [16–20] have been constructed, which are more enhancing functional metagenomics. Such functional metagenomics have revealed that an abundance of glycan-related genes reside in the intestinal environment, identifying 95 of 124 (77%) carbohydrate hydrolase families [16, 21, 22]. Since glycan-related genes play a key role in energy acquisition, cell–cell interactions, molecular recognition, signaling transduction, etc. [23], of various organisms, it is considered to be an essential target in terms of its ecological significance. Information on genes involved in polysaccharides and glycans in individual species have drastically increased [22, 24], but functional information about such genes found in environments is still unknown.

Enzymes that assemble and degrade glycans have been classified into sequence-based families by Henrissat et al. [25–30] The functional diversity or specificity of these enzymes is enormous and reflects the wide variety of glycan structures found in nature. CAZy is a database that was launched in 1991[25] which collects and organizes information based on these data. Carbohydrate active enzymes registered in CAZy are called carbohydrate-active enzymes (CAZymes). The classification system in the CAZy database is based on the results of biochemical experiments, and genes are classified into specific categories, including glycoside hydrolase (GH), polysaccharide lyase (PL), glycosyltransferase (GT), carbohydrate-binding module (CBM), carbohydrate esterase (CE), as well as several other categories of genes that act on carbohydrates (AA)[22].

The purpose of this study is to elucidate carbohydrate metabolism such as energy acquisition by taking advantage of functional metagenomics and observing an overview of CAZymes (glycan-related genes) in various environments. Using information of metagenome reads in various environment, we found that the environment influences the abundance and types of glycan-related genes, suggesting that their functional roles could be adapted to their environmental glycan characteristics. From these findings, we propose a model for the use of glycan-related genes for energy acquisition of

microorganisms in the environments. The functional glyco-metagenomics established in this study can make it possible to infer the role of glycan-related genes and microorganisms in the environment.

Results

Evaluation of the accuracy of the identified glycan-related genes based on complete genomes

To obtain information on the distribution of glycan-related genes in the environment, we developed a method to identify glycan-related genes from metagenomic data that are comprehensively sequenced by next-generation sequencers to determine the DNA sequences in the environment (Fig. 1). Metagenomic sequences often consist of short reads of 100–200 bp, and we evaluated the conditions for identifying glycan-related genes based on the sequence homologies of these short reads when aligned against a database of known glycan-related gene sequences. First, we generated virtual shotgun metagenomic data by randomly fragmenting the genomic sequence of 39 completely sequenced genomes that differ at Genus level including 17 Gram-positive bacteria and 22 Gram-negative bacteria (see Additional file 2: Table S1). These virtual metagenomic reads were aligned with 820,000 protein sequences registered in dbCAN [24] as a reference database using GhostX [31]. The alignment results were divided into positive and negative groups based on certain cutoff values using two variables, alignment identity (60–90%) and length (5–25 aa). For the evaluation of prediction accuracy, genes registered in CAZy were used as the correct set, and if a candidate gene was above a certain cutoff value in terms of alignment identity (60–90%) and length (5–25 aa), it was considered true, and otherwise false. The accuracy of glycan-related genes extracted in

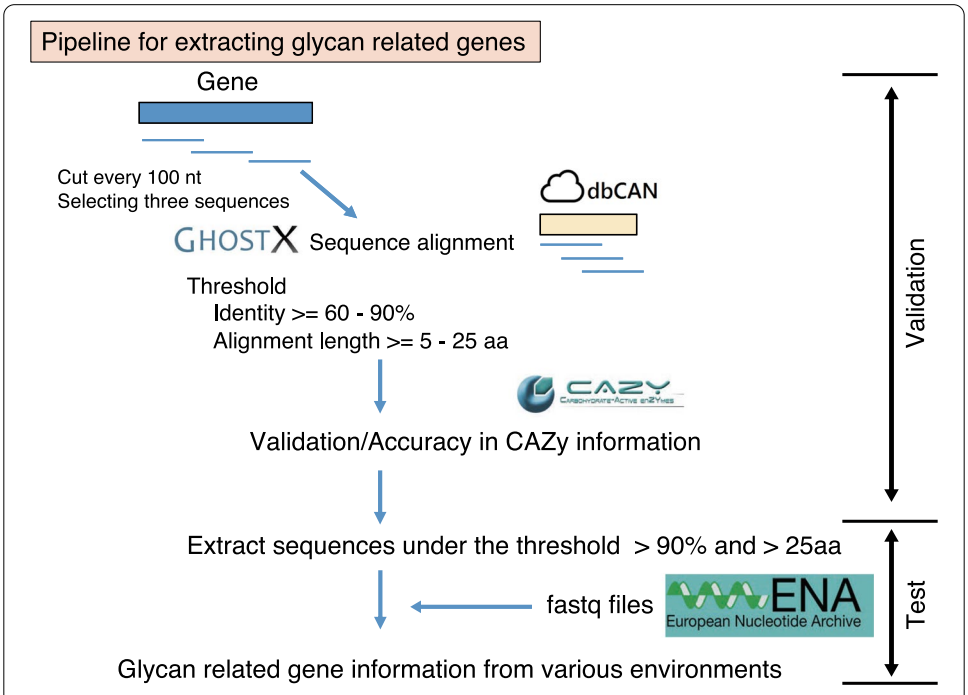

Fig. 1 A strategy of identification of glycan-related genes in environments

each bacterial genome was evaluated using Precision, Recall, and FDR (Fig. 2). These results showed that the highest accuracy was obtained when the identity was > 90% and the alignment length was > 25 aa, where Precision was > 90% (Fig. 2a), Recall was < 10% (Fig. 2c) and FDR was < 10% (Fig. 2b). Therefore, this condition was used for the subsequent identification of glycan-related genes.

Glycan-related genes identified from metagenome data

To investigate the distribution of glycan-related genes in microbial communities in various environments, we used the above method to detect glycan-related genes in the actual environmental metagenome (Fig. 1). We identified 86.73 million glycan-related genes from 198 metagenomic data obtained from ENA [32] (see Additional file 2: Table S3). The ratio of the number of identified glycan-related genes to the total number of reads included in each metagenomic data was high in the human-related environment such as human intestine and oral cavity, but 0.01% or less in the remaining 19 metagenome data (eleven from the deep sea, two from a hot spring, five from oil contamination, and one from rhizosphere).

In addition, the identified glycan-related genes differed greatly in their proportions in the environment compared to individual metagenomes. Thus, we calculated the distribution of the coverage of glycan-related genes in each environment (Fig. 3a). As a result, glycan-related genes were identified more in the human-associated metagenomes (average 2.64%) consisting of Gut (0.8–8.0%), Oral (0.9–2.8%), tGut (1.5–8.5%) and Skin (0.01–2.1%) than those in other environments (average 0.27%).

Although the number and content of glycan-related genes in the environment suggest that they are affected by the environment, their roles in each sample is unknown because glycan-related genes have a high variety of functions, from hydrolysis to transfer reactions of glycosidic bonds between sugars. Therefore, the identified glycan-related genes were annotated with the six functional categories defined by CAZy. The functions of the identified glycan-related genes were classified using the sequence annotations of the top hits. In each metagenome data set, we calculated the ratio of the six categories in which all of the identified glycan-related genes were annotated (Fig. 3b, Additional file 2: Table S4). We found that the proportion of the six classifications showed little difference among the samples. However, the ratio of GH (glycoside hydrolase) and GT (glycosyltransferase) in each environment was very high at 40–60%, and, surprisingly,

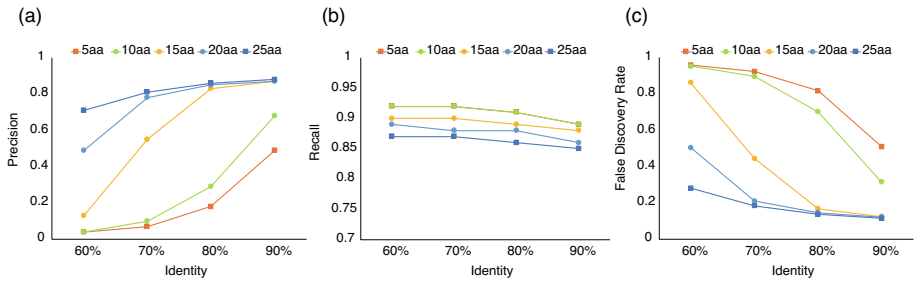

Fig. 2 Accuracy assessment of glycan-related gene identification using short read alignments. Accuracy indices used in this study are precision (**a**), recall (**b**) and false discovery rate (**c**)

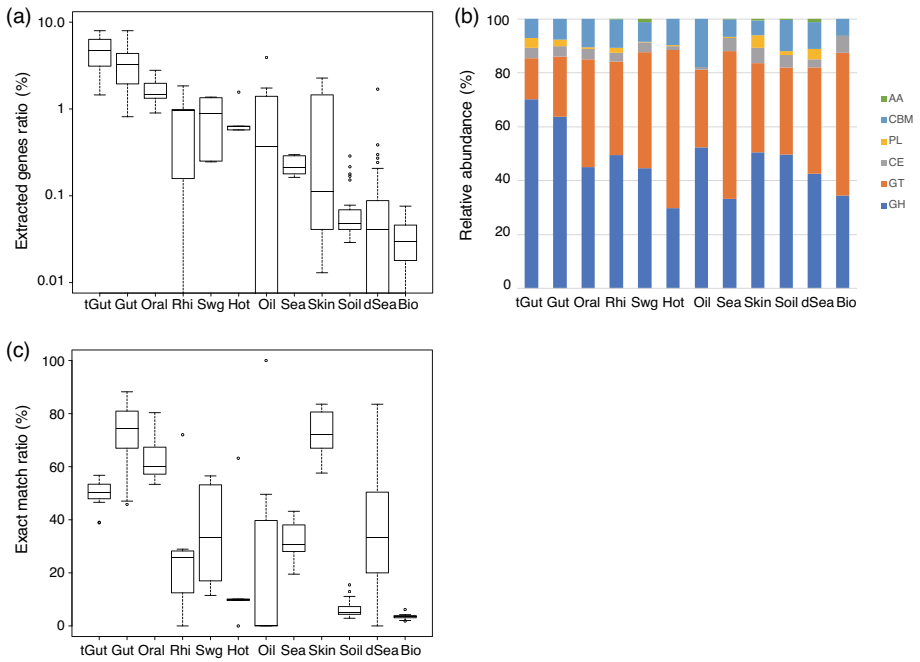

Fig. 3 Glycan-related genes identified from environmental metagenome sequences. **a** Ratio of glycan-related genes in each environment. To calculate the ratio, the number of reads mapped to glycan-related genes was divided by the total number of reads. **b** Relative abundance of enzyme classes of CAZy in the identified glycan-related genes. *GH* glycoside hydrolase, *GT* glycosyltransferase, *CE* carbohydrate esterase, *PL* polysaccharide lyase, *CBM* carbohydrate binding module, *AA* auxiliary activities. **c** Ratio of glycan-related genes perfectly matched with the sequence stored in the dbCAN database. The body of box plots in (**a**) and (**c**) goes from the first to third quartiles of the distribution and the center line is at the median. The abbreviation of the environmental names are *Bio* biofilm, *dSea* deep sea, *Gut* gut, *Hot* hot spring, *Oil* oil contaminated, *Oral* oral, *Rhi* rhizosphere, *Sea* sea, *Swg* sewage, *Skin* skin, *Soil* soil, *tGut* tumor gut

the combined ratio of both was constantly over 80% in all environments. There are six environments with the highest number of GH (Gut, tGut, Rhizosphere, Oil contamination, Skin, Soil) and six environments with the highest number of GTs (Oral, Sewage, Hot spring, Sea, Deep sea, Biofilm). In particular, the relative amount of GT was highest in Hot spring, Sea, and Biofilm. Thus, GTs are expected to be important in such aquatic feature and biome. From these results, it is possible that the ratio of the functions of glycan-related genes differs depending on the environment, and that the functions of the assembly of these genes also differ depending on the environment.

In this study, > 90% identity was used as a parameter to identify novel glycan-related genes. To examine the extent to which known glycan-related genes in the environment were identified, we calculated the percentage of amino acid sequences that completely matched the amino acid sequences in the reference database we used (Fig. 3c, Additional file 2: Table S5). Sequences that exactly matched the sequences registered in dbCAN varied among the metagenomic samples. However, when compared by environment, Gut, Skin, and Oral had a large number of exact matches, exceeding 50%, whereas Soil had a low number of less than 10%. This suggests that human-related environments, such as Gut and Skin, are often well-studied as research subjects, indicating that a large number of CAZymes are registered from these environments.

Enrichment of GH and GT family in various environments

Since carbohydrates are a source of energy for many microorganisms, their glycan-related genes may be adapted to different types and compositions of carbohydrates in their environment. Here, we defined I_{ms} as the proportion of metagenomic samples in the environment in which glycan-related genes for a family have been identified (see Methods). First, for the GH families, we examined the distribution of each family in the environment. We performed clustering analysis in order to classify the distribution of GH families in each environment (Fig. 4, Additional file 2: Table S6). As a result, we found that there are GH families specific to an environment, and therefore, GH families can be classified into several groups depending on the pattern of GH distribution. The family of GHs commonly detected in most of the environments, those in a few environments, and those in between were designated general-GH, specific-GH, and moderate general-GH, respectively (Fig. 4a). The general-GH group contained enzymes such as α-amylase found in many species. However, rare sugar hydrolytic enzymes such as xylanase and fucosidase were found in moderate general-GH, whereas polysaccharide-degrading enzymes such as glucanase and chitinase were found in specific-GH. Specific and moderate general families showed high I_{ms} in the rhizosphere (> 0.6 of I_{ms}), whereas most of the others showed lower values in oil contamination samples (< 0.3 of I_{ms}).

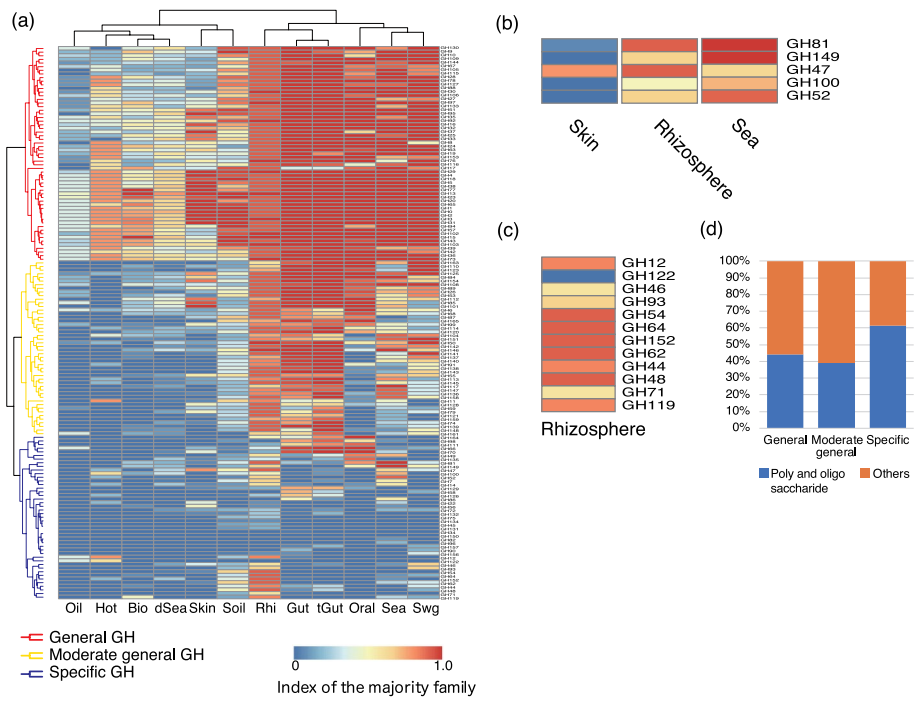

Fig. 4 Clustering of the families and distribution of the substrates in GH. **a** Cells in the heatmap show the I_{ms} of GH families (see Methods). **b** Remarkable family of specific-GH in skin, rhizosphere and sea. This figure is an enlargement of (**a**). **c** Remarkable family of specific-GH in rhizosphere. This figure is an enlargement of (**a**). **d** Ratio of the number of families that avail polysaccharide and oligo-saccharide as substrates among all GH families. The abbreviation of the environmental names are *Bio* biofilm, *dSea* deep sea, *Gut* gut, *Hot* hot spring, *Oil* oil contaminated, *Oral* oral, *Rhi* rhizosphere, *Sea* sea, *Swg* sewage, *Skin* skin, *Soil* soil, *tGut* tumor gut

As a distinctive family of specific-GH, α-mannosidase (GH47) showed high I_{ms} (>0.7 of I_{ms}) in skin, sea, and rhizosphere (Fig. 4b). This is an enzyme that trims high- (oligo-) mannose type *N*-glycans, and it is unique to eukaryotes. Therefore, it was considered to be a gene derived from yeast and fungi in the environment [33]. In Rhizosphere, a group of GHs showing high I_{ms}, endoglucanase (GH44, GH64, GH152), chitinase (GH48), xylosidase (GH54), α-L-arabinofuranosidase (GH62), nigeran digestion enzyme (GH71), and α-Amylase (GH119) were found (Fig. 4c). These are polysaccharide-degraders, which are thought to originate from species living in the plant-derived polysaccharide-rich rhizosphere [34, 35].

To investigate role of specific-GHs in each environment, based on the GH family information in CAZy, the substrates of each family were classified into polysaccharides, oligosaccharides, disaccharides and monosaccharides. The fraction of polysaccharide and oligosaccharide families in specific-GH was 61%, whereas that of general-GH was 44%, and moderate-GH was 39% (Fig. 4d, Additional file 2: Table S8). This fraction in the specific-GH group was significantly higher than in general-GH ($p=0.01$, Chi-squared test) and moderate-GH ($p=0.03$, Chi-squared test). GH is a hydrolytic enzyme, an enzyme that builds the metabolic system for energy acquisition mainly through the degradation of sugars. This suggests that these enzymes break down polysaccharides specific to each environment into small sugars to obtain energy.

Next, we performed clustering analysis in order to classify the distribution of GT families in each environment (Fig. 5, Additional file 2: Table S7). Two groups of GT types that are specific to each environment were detected based on their pattern of GT distribution (see Additional file 2: Table S9). The family of GTs commonly detected in 11 environments including 170 metagenomes was designated general-GT, and those in specific environments were designated specific-GT. A number of GTs such as trehalose phosphatase were found in many species in the general-GT group and specific enzymes such as mannosyl-transferase in the specific-GT group.

As a characteristic example of specific-GTs, there was a group of inverted I_{ms} pattern between gut and oral (Fig. 5b). The species harboring families showing high I_{ms} in Oral (>0.7 of I_{ms}), such as α-L-arabinosyltransferase (GT53), α-D-arabinofuranosyltransferase (GT85), α-1,2-mannosyltransferase (GT87), and β-1,2-arabinofuranosyltransferase (GT89), were investigated. As a result, it was found that α-L-arabinosyltransferase (GT53) is an enzyme widely identified in prokaryotes, while galactane α-D-arabinofuranosyltransferase (GT85), α-1,2-mannosyltransferase (GT87), and β-1,2-arabinofuranosyltransferase (GT89) are enzymes identified in *Corynebacterium* and *Mycobacterium*. These microbial species can indeed be detected in oral and the nasopharynx, and the present results support this fact [36–38].

Regarding the five GT families with high I_{ms} in gut (>0.7 of I_{ms}), α-1,6-L-fucosyltransferase (GT23), α-glucosyltransferase (GT44), Fuc4NAc transferase (GT56), α-1,3-galactosyltransferase (GT77), α-2,6-sialyltransferase (GT80), their species were investigated. It was found that α-glucosyltransferase (GT44) was identified in *Chlamydia* and pathogen-*Escherichia*, Fuc4NAc transferase (GT56) in *Salmonella* and *Klebsiella*, and α-2,6-sialyltransferase (GT80) in Pasteurellaceae and *Citrobacter*. In addition, α-1,6-L-fucosyltransferase (GT23) has been identified in *Bacteroides fragilis* [39]. More than 90% of the species in these families registered on CAZy were species known to reside

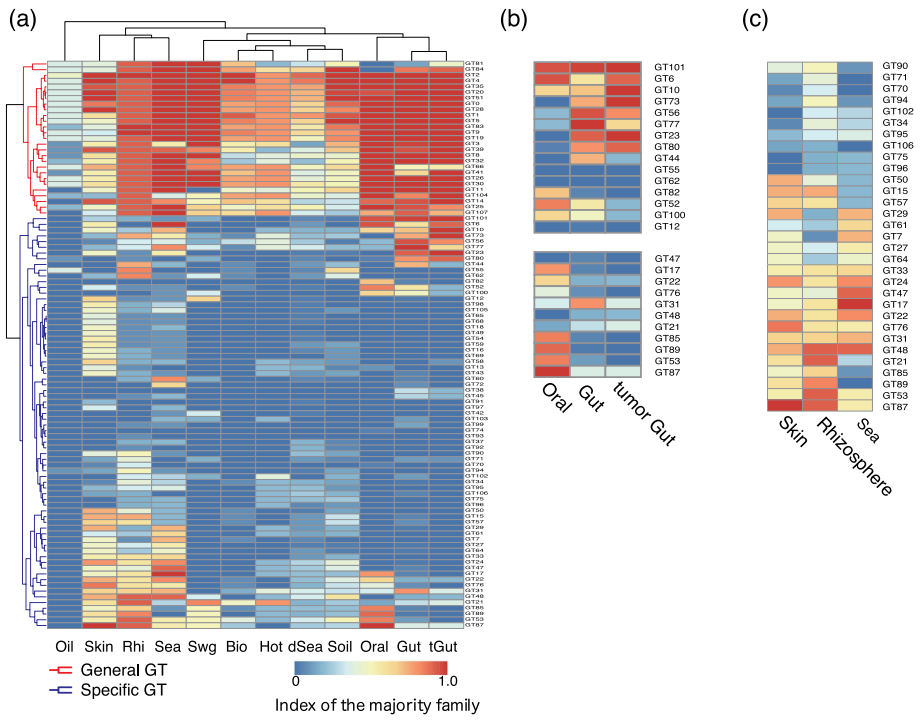

Fig. 5 Clustering of the families and distribution of the substrates in GT. **a** Cells in the heatmap show the I_{ms} of GT families (see Methods). **b** Remarkable family of specific-GT in oral, gut and tumor gut. This figure is an enlargement of (**a**). **c** Remarkable family of specific-GT in skin, rhizosphere and sea. This figure is an enlargement of (**a**). The abbreviation of the environmental names are *Bio* biofilm, *dSea* deep sea, *Gut* gut, *Hot* hot spring, *Oil* oil contaminated, *Oral* oral, *Rhi* rhizosphere, *Sea* sea, *Swg* sewage, *Skin* skin, *Soil* soil, *tGut* tumor gut

in the gut microbiome. It was reported that GT77 inactivated in the human intestine is carried by a group of microorganisms such as *Streptococcus* and *Escherichia* [40]. These results indicate that the CAZy families of gut and oral with reversed I_{ms} patterns is due to the species in the environment.

β-glucosyltransferase (GT21) showed high I_{ms} in rhizosphere (>0.7 of I_{ms}), sea (>0.4 of I_{ms}), sewage (>0.7 of I_{ms}), and hot spring (>0.7 of I_{ms}), which is an enzyme involved in pullulan synthesis and is known to be possessed by the genus *Aspergillus* [41]. β-1,4-*N*-acetylglucosaminyltransferase (GT17, MGAT3) has been identified in all metagenome samples of sea and in species such as *Acetobacter* and *Enterobacter* found in the ocean [42, 43].

Discussion

For functional metagenomics, it is desirable that databases used for gene annotation have sufficiency and little bias. In this study, strict search results showing complete homology suggest that many genes registered in CAZy and dbCAN were detected in human-related environments. Bias in the human-related environment as reference gene sequences used for gene function prediction could lead to missed prediction or mass production of hypothetical proteins. If the search is based on exact match only, this bias towards gene sequence data from the human-related environment is likely to have a

significant impact on the search results. However, we did not take this bias into account in the present analysis because the search results with the method we developed in this study (> 90% identity, > 25 amino acids alignment) showed sufficiently high accuracy in searching for homology with glycan-related genes.

For example, the number of glycan-related genes identified in the soil environments increased in more samples using our method than in the exact match criterion. An environment in which such new genes are often found is very likely to be a candidate environment for new gene discovery. In addition, the number of entries in CAZy and dbCAN has been increasing every year, and new families of each classification for new reaction modes and substrates continue to be established. GH157-161 was newly established in 2019, and GHs related to that family accounted for an average of 0.26% (0–2.6%) of the GHs identified in the present results. Therefore, it is highly possible that novel glycan-related genes with different substrates and functions could be identified in the soil environment, compared with the case of using the reference sequence database before that. Thus, it is highly likely that the opportunities for the discovery of novel glycan-related genes will increase as CAZy continues to be updated in the future.

The proportion of identified glycan-related genes between the environments showed a slight difference in intra-environments, but a large difference in inter-environments (see Additional file 1: Fig. S1). In particular, the proportion of the identified genes was higher in the human-associated environments, which is consistent with previous studies showing that the microbiome in human-associated environments harbor more genes for sugar-based metabolic systems (e.g., energy production) [16, 21, 22]. On the other hand, deep sea and oil contamination etc. detected a lower proportion of glycan-related genes, but in such an oligotrophic environment, there is less opportunity for sugar to be supplied to the environment due to less material circulation. Therefore, it is possible that the lower need for glycan-related genes in these environments compared to other genes may have led to the adaptation to a lower proportion of glycan-related genes.

Furthermore, the pattern of distribution was divided into two groups, one with a high GH content and the other with a high GT content. Surprisingly, the sum of both GHs and GTs remained constant at 80%, despite the fact that the ratio of the identified glycan-related genes differed between environments. Since the state of sugars may differ depending on an environment, gene organization in microbial species could be adaptable to substances in the environment. Specific enzymes are considered to play a role in the adaptation to the environment. In this study, we explored glycan-related genes, but genes related to other substances (organic compounds such as lipids and phosphate groups) may also show similar trends. Oil degrading bacteria such as *Pseudomonas* and *Rhodococcus* exist in the oil environment and can derive their energy from petroleum hydrocarbons by enzymes such as alkane hydrogenase [44, 45]. This suggests that microbes in the Rhizosphere may derive their energy from sugars, whereas in the oil environment they may derive it from oil.

General glycan-related genes for both GH and GT families were widely distributed in various environments. When these genes were mapped to "Starch and sucrose metabolism" in the KEGG PATHWAY [46–48], they accounted for 68% of the 76 EC numbers on that pathway (see Additional file 2: Table S10). From the mapped genes on the pathway, a series of metabolisms were linked from Amylose and GlcNAc to Glycolysis via

Glucose, Xylose, Arabinose and Mannose. Therefore, general-GH and general-GT could play an important role in energy acquisition. In the case of the genes in the specific-GH and specific-GT families, no enrichment to certain pathway maps was detected. Thus it can be expected that these genes are most likely involved in a specific metabolic pathway in each environment, rather than a common pathway across environments.

Among these specific-enzymes, many were enzymes that decompose polysaccharides. Certain distinctive specific enzymes were found such as plant-derived polysaccharides [49] (chitin, cellulose, lentinan, xylan) in rhizosphere, pullulan and laminaran in sea, and dextran and galactomannan in gut and oral [50]. These environment-specific glycans could be decomposed into more general glycan components by the specific enzymes in each environment. This would allow them to be further decomposed by a more common pathway of glycan metabolism. Finally, we surmise that general-enzymes play a role in the acquisition of energy from these glycan components (Fig. 6). According to this universal energy acquisition model, it is possible that environmentally-specific species and genes may play a role in regulating sugar metabolism in the environment.

Conclusions

It is important to evaluate the environment of the sample by predicting the function from metagenomic reads and inferring phenomena occurring in the sample. Our method should play an important role in establishing functional glyco-metagenomics to identify glycan-related genes in the environment from metagenomic data and to hypothesize the role of sugars by comparing them within and between environments. In addition, although we performed our analysis using metagenomic data already registered in repository databases, our re-analysis was able to illustrate new perspectives regarding sugars in the target environment. By applying our methods, we hope to find new perspectives and discoveries as we can now reanalyze large amounts of historical metagenomic data for comparison in various environments.

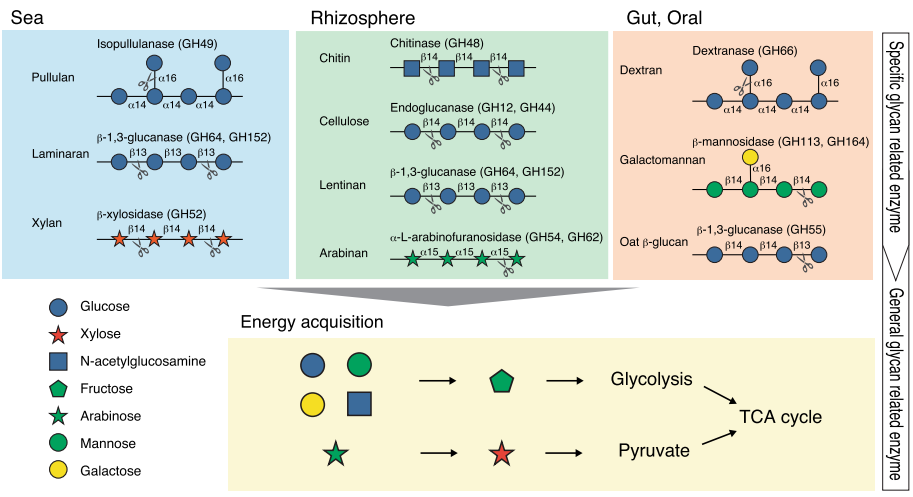

Fig. 6 A universal model of the roles of general and specific glyco-enzymes in environments for general and specific enzymes

Methods

Validation of identification glycan-related genes using sequence alignment

We developed a method for identifying glycan-related genes from nucleotide sequences such as short read sequences obtained from NGS using sequence alignment. The microbial genome of 39 genera (see Additional file 2: Table S1) belonging to multiple phyla was downloaded from KEGG [46–48]. Since these contained gene annotations, glycan-related genes could be found by referencing CAZy (http://www.cazy.org). If a detected gene was a glycan-related gene, the label "True" was added to the gene data, otherwise, the label "False" was added. The DNA sequences were cut every 100 bases from the 5' end, and three fragments were randomly selected for each gene. This process was performed with our ad hoc ruby scripts. The sequences of these fragments were converted to amino acid sequences using the standard genetic code. Six possible amino acid sequences were generated from a single fragment, due to the six frames possible on the forward and reverse strands and starting DNA position for each codon. This was used as the model dataset for evaluation of our method to identify glycan-related genes from metagenomic sequence data. The fragment peptide sequences obtained in these processes were generated using GhostX version 2.1 [31].

The 1.38 million amino acid sequences of proteins in FASTA format with respect to glycan-related genes were downloaded from the dbCAN meta server (http://bcb.unl.edu/dbCAN2/). After deleting approximately 550,000 redundant sequences, the resulting database consisted of approximately 830,000 sequences, for the reference data to identify glycan-related genes.

The alignments were carried out between all fragment peptide sequences and amino acid sequences of the reference database using GhostX. It is expected that in the case of glycan-related genes, alignments show high identity and long alignment length. Therefore, prediction of whether each fragment was a glycan-related gene was performed under the conditions of identity thresholds of 60–90% and an alignment length threshold of 5–25 aa using a validation dataset where each fragment is known to be glycan-related gene. The effectiveness of our prediction method was evaluated by computing the precision, recall and false discovery rates which were calculated according to following equations:

$$\begin{aligned}\text{Precision} &= \text{True Positive}/(\text{True Positive} + \text{False Positive})\\ \text{Recall} &= \text{True Positive}/\big(\text{True Positive} + \text{False Negative}\big)\\ \text{FDR} &= \text{False Positive}/(\text{True Positive} + \text{False Positive})\end{aligned}$$

for each genome.

Acquisition and sequence alignment of metagenomes in various environments

In total, 198 metagenomes were downloaded from the European Nucleotide Archive (https://www.ebi.ac.uk/ena). Each of these include over two million reads whose lengths are over 100 bases. Additional file 2: Table S2 shows the accession numbers, the environments from which the genomes were taken, and the total number of sequence reads. These 198 metagenomes were classified into 12 environments based on the Metagenome/Microbe Environmental Ontology [51]. According to the method described in

the previous section, sequence alignment of metagenome reads was performed using GhostX.

Identification of glycan-related genes and organization of gene information

For our 198 metagenomes, the populations of the reads of glycan-related genes against the total number of reads were calculated, and the averages and the standard deviations of the populations of glycan-related genes in the metagenomes were calculated for each of the 12 environments. The relative population of each class of CAZy against the glycan-related genes were also calculated.

The identification of glycan-related genes is based on > 90% of identity and > 25 aa of alignment length thresholds. This method was used as the optimal condition to search for a wide variety of glycan-related genes. However, it is not possible to distinguish whether the identified genes by our method are known genes or undiscovered genes. Of these identified glycan-related genes, genes showing 100% identity to the reference sequences were explored. Since these exact-matched sequences are no different from known gene sequences, at least with respect to their target sequence regions, they were considered as genes already identified and registered in the public databases. The processes in this section to identify glycan-related genes from the GhostX results were performed with our ad hoc ruby scripts.

Comparison of glycan-related genes for each environment

According to the CAZy functional classification, the glycoside hydrolase (GH) and the glycosyltransferase (GT) families, were classified into 166 and 109 families, respectively. The obtained reads of glycan-related genes were thus also classified by family. We defined the index of the majority family (I_{ms}) in an environment as the ratio of the number of metagenomes with the identified glycan-related genes to the number of metagenomes belonging to the environment. I_{ms} indicates how major the family of the glycan-related genes is in a given environment. If $I_{\mathrm{ms}} = 1$, it means that all metagenomes in the environment have glycan-related genes in the family, whereas $I_{\mathrm{ms}} = 0$ means that there are no glycan-related genes in the family. This process was performed using ad hoc ruby scripts. In order to visualize this data across each environment and family, Euclidean distances were calculated and hierarchical clustering analysis was carried out using the Ward's method with the pheatmap library (https://cran.r-project.org/web/packages/pheatmap/index.html) for R (https://www.R-project.org/) was used to produce the heatmap figure with the default color setting for this clustering.

The families of GHs were also classified according to substrate specificities, which are also in the CAZy database. The families of GHs were classified into monosaccharides, disaccharides, oligosaccharides, polysaccharides, peptidoglycan-related, and others, based on the substrate information provided by the CAZy database and the more detailed descriptions in KEGG COMPOUND [46–48]. In order to analyze the functions of the genes in the general-enzyme families, the EC numbers of general-enzyme genes were mapped to "Starch and sucrose metabolism" in the KEGG PATHWAY maps, and the mapped EC numbers were counted.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04425-9.

Additional file 1: Fig. S1. Percentage of predicted glycan-related genes in each metagenomic sample. The abbreviation of the environmental names are Bio: biofilm, dSea: deep sea, Gut: gut, Hot: hot spring, Oil: oil contaminated, Oral: oral, Rhi: rhizosphere, Sea: sea, Swg: sewage, Skin: skin, Soil: soil, and tGut: tumor gut.

Additional file 2: Table S1. List of bacterial taxonomy used in this study. **Table S2**. List of metagenomic samples used in this study. **Table S3**. Glycan-related genes predicted in each metagenome sample. **Table S4**. CAZy classification of predicted glycan-related genes in each metagenome sample. **Table S5**. Glycan-related genes perfectly matched to the dbCAN sequences. **Table S6**. GH families in each metagenome sample. A cell value is relative abundance in each sample. **Table S7**. GT families in each metagenome sample. A cell value is relative abundance in each sample. **Table S8**. Classification of substrates and properties of GH families. **Table S9**. Classification of specific- or general-properties of GT families. **Table S10**. Coverage of EC numbers on the KEGG pathway map "Starch and sucrose metabolism" (KEGG PATHWAY ko00500).

Acknowledgements
This study was supported by the Database Integration Coordination Program of the National Bioscience Database Center (NBDC), Japan Science and Technology Agency [17934031].

Authors' contributions
KK conceived the study, HT and SO designed the study and HT performed the analyses. HT, NM, and SO wrote the manuscript. All authors read and approved the final manuscript.

Funding
NBDC/JST(1793403117934031).

Availability of data and materials
All data generated or analyzed during this study are included in this published article and its supplementary information files. All sequence files used publicity available data in European Nucleotide Archive. Accession numbers were showed Additional file 2: Table S2.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Division of Bioinformatics, Niigata University Graduate School of Medical and Dental Sciences, 1-757 Asahimachi-dori, Chuo-ku, Niigata 951-8510, Japan. [2]Glycan and Life Systems Integration Center, Faculty of Science and Engineering, Soka University, 1-236 Tangi-machi, Hachioji, Tokyo 192-8577, Japan.

Received: 19 April 2021 Accepted: 4 October 2021
Published online: 18 October 2021

References

1. Venter JC, et al. Environmental genome shotgun sequencing of the Sargasso sea. Science (80-). 2004;304(5667):66–74. https://doi.org/10.1126/science.1093857.
2. Sunagawa S, et al. Structure and function of the global ocean microbiome. Science (80-). 2015;348(6237):1–10. https://doi.org/10.1126/science.1261359.
3. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC Biol. 2014;12(1):1–4. https://doi.org/10.1186/s12915-014-0069-1.
4. Muegge BD, et al. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science (80-). 2011;332(6032):970–4. https://doi.org/10.1126/science.1198719.
5. Methé B, et al. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science (80-). 2012;486(7402):215–21. https://doi.org/10.1038/nature11209.A.
6. T. H. M. P. Consortium. Structure, function and diversity of the healthy human microbiome. Science (80-). 2011;486(7402):207–14. https://doi.org/10.1038/nature11234.Structure.
7. Ngara TR, Zhang H. Recent advances in function-based metagenomic screening. Genomics Proteomics Bioinform. 2018;16(6):405–15. https://doi.org/10.1016/j.gpb.2018.01.002.
8. Wang WL, Xu SY, Ren ZG, Tao L, Jiang JW, Sen Zheng S. Application of metagenomics in the human gut microbiome. World J Gastroenterol. 2015;21(3):803–14. https://doi.org/10.3748/wjg.v21.i3.803.

9. Gordon J, et al. F1000Prime recommendations of: Human gut microbiome viewed across age and geography. Nature. 2012;486(7402):222–7. https://doi.org/10.1038/nature11053.Human.
10. Qin J, et al. Europe PMC Funders Group Europe PMC Funders Author Manuscripts A human gut microbial gene catalog established by metagenomic sequencing. Nature. 2010;464(7285):59–65. https://doi.org/10.1038/nature08821.A.
11. Helbert W, et al. Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space. Proc Natl Acad Sci USA. 2019;116(13):6063–8. https://doi.org/10.1073/pnas.1815791116.
12. Hanson AD, Pribat A, de Creécy-Lagard V. 'Unknown' proteins and 'orphans' enzymes: the mising half of the engineering part list—and how to find it. Biochem J. 2010;425(1):1–11. https://doi.org/10.1042/BJ20091328.
13. Roberts RJ. COMBREX: computational bridge to experiments. Biochem Soc Trans. 2011;39(2):581–3. https://doi.org/10.1042/BST0390581.
14. Hervé V, et al. Phylogenomic analysis of 589 metagenome-assembled genomes encompassing all major prokaryotic lineages from the gut of higher termites. PeerJ. 2020;2020(2):1–27. https://doi.org/10.7717/peerj.8614.
15. Tierney BT, et al. The landscape of genetic content in the gut and oral human microbiome. Cell Host Microbe. 2019;26(2):283-295.e8. https://doi.org/10.1016/j.chom.2019.07.008.
16. El Kaoutari A, Armougom F, Gordon JI, Raoult D, Henrissat B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. Nat Rev Microbiol. 2013;11(7):497–504. https://doi.org/10.1038/nrmicro3050.
17. Cantarel BI, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. 2009;37(SUPPL. 1):233–8. https://doi.org/10.1093/nar/gkn663.
18. Duarte M, Jauregui R, Vilchez-Vargas R, Junca H, Pieper DH. AromaDeg, a novel database for phylogenomics of aerobic bacterial degradation of aromatics. Database. 2014;1–12:2014. https://doi.org/10.1093/database/bau118.
19. Arango-Argoty GA, et al. ARGminer: a web platform for the crowdsourcing-based curation of antibiotic resistance genes. Bioinformatics. 2020;36(9):2966–73. https://doi.org/10.1093/bioinformatics/btaa095.
20. Zhang T, Miao J, Han N, Qiang Y, Zhang W. MPD: a pathogen genome and metagenome database. Database. 2018;2018(2018):1–6. https://doi.org/10.1093/database/bay055.
21. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol. 2019;37(8):953–61. https://doi.org/10.1038/s41587-019-0202-3.
22. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42(D1):D490–5. https://doi.org/10.1093/nar/gkt1178.
23. Ajit Varki PHS, Cummings RD, Esko JD, Stanley P, Hart GW, Aebi M, Darvill AG, Kinoshita T, Packer NH, Prestegard JH, Schnaar RL, Seeberger PH. Essentials of glycobiology, 3rd *edn.*, Cold Spring Harbor Laboratory Press, 2017.
24. Zhang H, et al. DbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46(W1):W95–101. https://doi.org/10.1093/nar/gky418.
25. Henrissat B. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J. 1991;280(2):309–16. https://doi.org/10.1042/bj2800309.
26. Henrissat B, Bairoch A. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J. 1993;293(3):781–8. https://doi.org/10.1042/bj2930781.
27. Henrissat BA. Updating the sequence-based classification of glycosyl hydrolases. Biochem J. 1996;316:695–6.
28. Campbell JA, Davies GJ, Bulone V, Henrissat B. Correction: A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities (Biochemical Journal (1997) 326 (929–939)). Biochem J. 1998;329(3):719. https://doi.org/10.1042/bj3290719.
29. Lombard V, Bernard T, Rancurel C, Brumer H, Coutinho PM, Henrissat B. A hierarchical classification of polysaccharide lyases for glycogenomics. Biochem J. 2010;432(3):437–44. https://doi.org/10.1042/BJ20101185.
30. Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol Biofuels. 2013;6(1):1. https://doi.org/10.1186/1754-6834-6-41.
31. Suzuki S, Kakuta M, Ishida T, Akiyama Y. GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. PLoS ONE. 2014;9(8):1–8. https://doi.org/10.1371/journal.pone.0103833.
32. ENA. European Nucleotide Archive. 2020. https://www.ebi.ac.uk/ena/browser/home
33. Thompson AJ, et al. The reaction coordinate of a bacterial GH47 α-mannosidase: a combined quantum mechanical and structural approach. Angew Chem Int Ed. 2012;51(44):10997–1001. https://doi.org/10.1002/anie.201205338.
34. Gao X, Wu Z, Liu R, Wu J, Zeng Q, Qi Y. Rhizosphere bacterial community characteristics over different years of sugarcane ratooning in consecutive monoculture. Biomed Res Int. 2019. https://doi.org/10.1155/2019/4943150.
35. Gupta R, Mukerji KG. Nigeran production in some Aspergillus and Penicillium species. Folia Microbiol (Praha). 1982;27(1):38–42. https://doi.org/10.1007/BF02883836.
36. Alderwick LJ, Seidel M, Sahm H, Besra GS, Eggeling L. Identification of a novel arabinofuranosyltransferase (AftA) involved in cell wall Arabinan biosynthesis in *Mycobacterium tuberculosis*. J Biol Chem. 2006;281(23):15653–61. https://doi.org/10.1074/jbc.M600045200.
37. Seidel M, Alderwick LJ, Birch HL, Sahm H, Eggeling L, Besra GS. "Identification of a novel arabinofuranosyltransferase AftB involved in a terminal step of cell wall arabinan biosynthesis in Corynebacterianeae, such as *Corynebacterium glutamicum* and *Mycobacterium tuberculosis*. J Biol Chem. 2007;282(20):14729–40. https://doi.org/10.1074/jbc.M700271200.
38. Morita YS, et al. PimE is a polyprenol-phosphate-mannose-dependent mannosyltransferase that transfers the fifth mannose of phosphatidylinositol mannoside in mycobacteria. J Biol Chem. 2006;281(35):25143–55. https://doi.org/10.1074/jbc.M604214200.
39. Huang HH, et al. Substrate characterization of bacteroides fragilis α1,3/4-fucosyltransferase enabling access to programmable one-pot enzymatic synthesis of KH-1 antigen. ACS Catal. 2019;9(12):11794–800. https://doi.org/10.1021/acscatal.9b04182.

40. Montassier E, et al. Distribution of bacterial α1,3-galactosyltransferase genes in the human gut microbiome. Front Immunol. 2020;10(January):1–9. https://doi.org/10.3389/fimmu.2019.03000.
41. Singh RS, Kaur N, Rana V, Kennedy JF. Pullulan: a novel molecule for biomedical applications. Carbohydr Polym. 2017;171:102–21. https://doi.org/10.1016/j.carbpol.2017.04.089.
42. Arciola CR, Campoccia D, Ravaioli S, Montanaro L. Polysaccharide intercellular adhesin in biofilm: structural and regulatory aspects. Front Cell Infect Microbiol. 2015;5(FEB):1–10. https://doi.org/10.3389/fcimb.2015.00007.
43. Band VI, Crispell EK, Napier BA, Herrera CM, Tharp GK, Vavikolanu K, Pohl J, Read TD, Bosinger SE, Stephen Trent M, Burd EM, Weiss DS. Antibiotic failure mediated by a resistant subpopulation in Enterobacter cloacae. Nat Microbiol. 2016;1(6):16053.
44. Xu J, Zhang Q, Li D, Du J, Wang C, Qin J. Rapid degradation of long-chain crude oil in soil by indigenous bacteria using fermented food waste supernatant. Waste Manag. 2019;85:361–73. https://doi.org/10.1016/j.wasman.2018.12.041.
45. Kumari S, Regar RK, Manickam N. Improved polycyclic aromatic hydrocarbon degradation in a crude oil by individual and a consortium of bacteria. Bioresour Technol. 2018;254(January):174–9. https://doi.org/10.1016/j.biortech.2018.01.075.
46. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28(1):27–30. https://doi.org/10.3892/ol.2020.11439.
47. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci. 2019;28(11):1947–51. https://doi.org/10.1002/pro.3715.
48. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021;49(D1):D545–51. https://doi.org/10.1093/nar/gkaa970.
49. Zhou Y, et al. Cyclodextrin glycosyltransferase encoded by a gene of Paenibacillus azotofixans YUPP-5 exhibited a new function to hydrolyze polysaccharides with β-1,4 linkage. Enzyme Microb Technol. 2012;50(2):151–7. https://doi.org/10.1016/j.enzmictec.2011.12.001.
50. Flint HJ, Scott KP, Duncan SH, Louis P, Forano E. Microbial degradation of complex carbohydrates in the gut. Gut Microbes. 2012;3(August):289–306.
51. MEO. Metagenome and microbes environmental ontology. 2020. https://bioportal.bioontology.org/ontologies/MEO.

Publisher's Note