Research Article

Assembly of 97 Novel Bacterial Genomes in the Microbial Community Affiliated with Polyvinyl Alcohol in Soil of Northern China

Jiaxing Liu,[1] Xun Gu [ID],[2] and Hui Li [ID][1]

[1]*MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200438, China*
[2]*Department of GDC Biology, Iowa State University, Ames, IA 50014, USA*

Correspondence should be addressed to Xun Gu; xgu@iastate.edu and Hui Li; leilyn102@88.com

Background. Undeveloped ecosystems belong to rich source of microbial population, of which resources remain unearthed. A kind of polymeric compound system with high polyvinyl alcohol (PVA) content has been reported and named *Taisui*. Marker gene amplification showed that *Taisui* harbored little-explored microbial communities. *Aim.* To address this issue, our study attempted to recover draft genomes and functional potential from microbial communities in *Taisui* using the metagenomic approach. *Material and Methods.* *Taisui* communities provided 97 novel bacterial genomes from 13 bacterial phyla, including bacteria candidate phylum. Two novel genus-level lineages were recovered from Planctomycetes and Chloroflexi. Based on the draft genomes, we expanded the number of taxa with potential productions of PKS and NRPS in phyla including Candidatus Dadabacteria, Chloroflexi, and Planctomycetes. *Results.* A rich diversity of PVA dehydrogenase genes from 4 phyla, involving Proteobacteria, Acidobacteria, Acitinobacteria, and Planctomycetes, were identified. The phylogenetic tree of PVA dehydrogenase showed the possibility of horizontal gene transfer between microbes. *Conclusion.* Our study underscores the substantial microbial diversity and PVA degradation potential in the previously unexplored *Taisui* system.

1. Introduction

Polyvinyl alcohol (PVA) is a common water-soluble polymer for farming, packing, fiber coating, etc. Due to the high mass of production and utilization, PVA pollution in the environment was considerable, especially in the major production regions: China, the USA, Western Europe, and Japan [1, 2]. For PVA degradation with less cost and sludge generation, bacterial PVA-degraders were isolated from PVA-containing systems [2]. However, symbiotic behaviors of PVA degrading microbes make the study of microbial community in PVA-containing systems be needed. A kind of underground polymeric compound (named "*Taisui*") system has been reported recently [3]. Chemical analysis showed that this system can be characterized by high level of PVA or a mixture of PVA and polyacrylic acid [3]. The distribution of *Taisui* has been analysed through finding reports from 1992 to 2015. More

than 60% of *Taisuis* existed in soil layers, and more than 60% of them were in northern China [4].

The PVA-containing system of *Taisui* can be a valuable object of metagenomic analysis. First, *Taisui* has stable morphological characteristics and internal structures [3]. Like a bioreactor, the structure of *Taisui* (Figure S1) separates microbial populations from surrounding environments. Second, *Taisui* has stable chemical compositions [3], which indicates stable microbial pattern. Third, *Taisui* systems harboring little-explored microbial communities. Previous research has shown that *Taisui* hosts a rich diversity of novel microbes [3]. More than 40% bacterial OTUs in 2 samples were unclassified at the phylum level, and more than 75% fungal OTUs in 8 samples were unclassified. Metagenomic sequencing can bypass the cultivation bottleneck by obtaining metagenome-assembled genomes (MAGs), leading to the discovery of novel microbial diversity and new metabolisms

from challenging systems [5, 6]. Based on these prior observations, we inferred that *Taisui* systems offered an opportunity for expand microbial diversity and PVA biodegradation. In this study, we reconstructed MAGs from metagenome of *Taisui* samples and investigated PVA degradation potential of *Taisui* microbiota. The results show the wealth of evolutionary diversity in unexplored systems, markedly expand the diversity of PVA dehydrogenase, and will contribute to future comparative studies of PVA polluted environments.

2. Material and Methods

2.1. Sample Collection. Four *Taisui* samples included in this study were collected from Jinzhou (Liaoning province), Baotou (Inner Mongolia), and Aksu (Xinjiang Province, Figure 1(a)). The *Taisui* individuals were washed using fresh water after being dug out from the soil layer, and the entire of them were delivered at room temperature (about 15°C). For each *Taisui* individual, four samples were collected in the meantime. The samples were collected both from the outside (the surface of *Taisui*) and the inside (about 5 cm depth from the outside) in equal quantity ($2 * 2 * 2$ cm of each). The samples, belonging to one *Taisui* individual, were pooled as one sample for future investigation. To observe the structure of *Taisui*, the TS.JZ1 sample was used as a representative sample for microscopic observations. Photographs were taken using a digital camera mounted on a light microscope.

2.2. DNA Extraction and Sequencing. Before DNA extraction, we used 1% dimethyl sulphoxide and dry bath to remove polyvinyl alcohol. DNA extraction referred to the phenol-chloroform extraction [7] was described in the supplement. After DNA purification, equal volumes of isopropanol at +4°C (100%) were added to the upper phase previously transferred into a clean 1.5 ml tube, then tubes were slowly mixed by inversion and kept overnight at -20°C, before another centrifugation (15 min at $16000 \times g$). Following removal of the supernatant, 1 ml of 70% ethanol at +4°C was added to the DNA pellets. These pellets were suspended by flicking the tubes, followed by inversion and centrifugation (15 min at $16000 \times g$), then pellets were dried for 10 min (V-AQ mode, Vacufuge plus, Eppendorf), and 50 μl of nuclease-free water was added. Samples were shotgun-sequenced for metagenomics on the Illumina HiSeq platform at Novogene (Tianjin, China), and paired-end reads were generated. For HiSeq sequencing, each sample can be barcoded (added during library preparation), and equal quantities of barcoded libraries can be multiplexed during sequencing. HiSeq reads are aligned to a cohort of nonredundant National Center for Biotechnology Information (NCBI) complete genomes using the Short Oligonucleotide Analysis Package (SOAP) alignment tool29, which is typically faster to run than the Basic Local Alignment Search Tool (BLAST) or the BLAST-like Alignment Tool (BLAT). Genome coverage is calculated using the SOAP.coverage package.

2.3. De Novo Assembly and Analysis. The quality control of the metagenomic reads was performed using FastQC [8]

and MultiQC [9]. Adapters and low-quality reads were removed. We set the minimum base quality score of 38. Bases with quality score < 38 were treated as low-quality bases. Reads were filtered as long as they contained more than 40 bp of low-quality bases. We also removed reads that contained more than 10 bp of N and reads that overlapped more than 15 bp with adapters to generate clean data. The high-quality reads were de novo assembled into contigs by using MEGAHIT [10] with default settings. We used QUAST [11] to evaluate metagenomic assemblies.
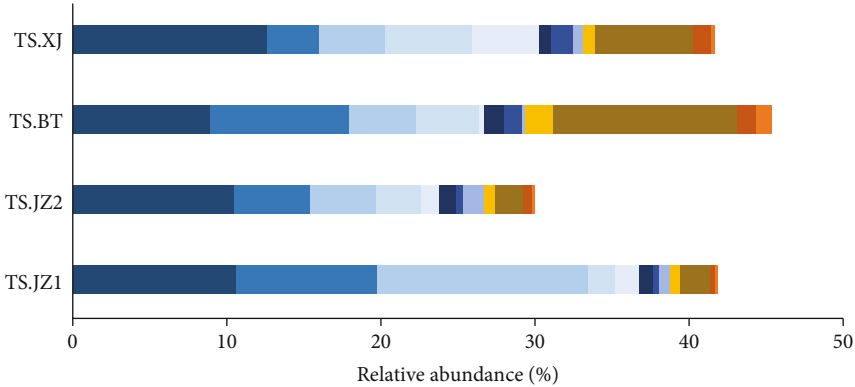
Assembled contigs were used to predict open reading frames (ORFs) using Prokka [12] with annotation mode included archaea, bacteria, mitochondria, and viruses. To get the nonredundant ORF dataset, the ORFs of four samples were merged, and the redundant ORFs were removed using CD-HIT [13, 14]. The abundance of ORFs in each sample was calculated using Salmon [15]. The ORFs were taxonomically annotated against the NCBI GenBank nonredundant protein sequence (nr) database using DIAMOND (*E*-value 1e-5) [16]. Taxonomic annotations were generated using MEGAN6 [17], with the maximum e-value cutoff as 1e-5. For KEGG and COG annotation, the ORFs were searched against eggNOG database using eggNOG-mapper [18]. Besides, the ORFs that encode carbohydrate-active enzyme (CAZyme) domain were profiled by mapping against CAZy database [19] using DIAMOND (*E*-value 1e-5).

Shannon-Weaver index, Simpson's index, Shannon evenness index, and Bray-Curtis index of all taxa at the species level were estimated using Vegan. UPGMA clustering of samples was obtained using SplitsTree4 [20]. The clustering tree and bar plots were edited in iTOL [21].

2.4. Meta-Pathway Reconstruction. The meta-pathways were the combination of metabolic pathways of multiple bacteria in the metagenomic dataset. The meta-pathways were reconstructed based on the functional annotation of ORFs (generated using eggNOG database and CAZy database). Also, the KEGG pathways were profiled using Pathview [22] for reconstruction. To evaluate the contribution of bacteria to a given enzyme in the metabolic pathways, the taxonomic annotation of ORFs (generated using NCBI-NR database) was combined with the functional annotation.
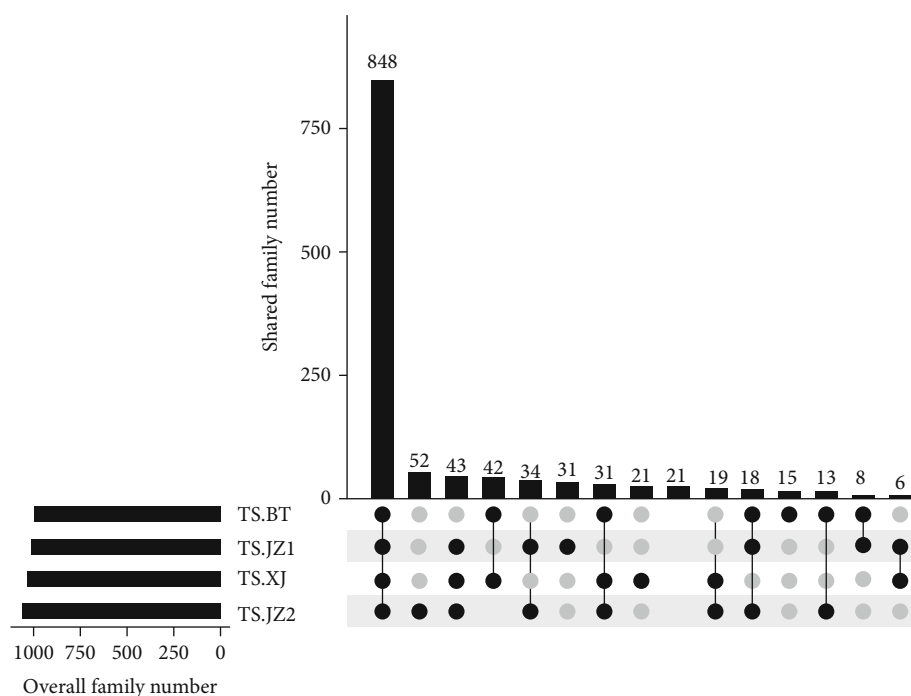
2.5. Metagenomic-Assembled Genome. Coassemblies were binned using MetaWRAP (parameters: -c 70 -x 5) [23], which called concoct, maxbin2, and metabat2 for binning at the same time. Bin refinement was performed based on the results of these binning software packages using "bin_refinement." Reads were aligned to contigs using BWA to generate separate files for each cluster. Separated reads were then reassembled using "reassemble_bins," and final beat bins were obtained based on the results of CheckM [24], which resulted in a final set of 97 bins.

Using CheckM, bins were filtered for completeness ≥ 70% and contamination ≤5%. Taxonomy was assigned to genome bins using GTDB-Tk [25]. The amino acid identity (AAI) between MAGs was calculated by CompareM v. 0.1.1 [26]. The replication of MAGs was checked at 95% average nucleotide identity (ANI) to generate the species number
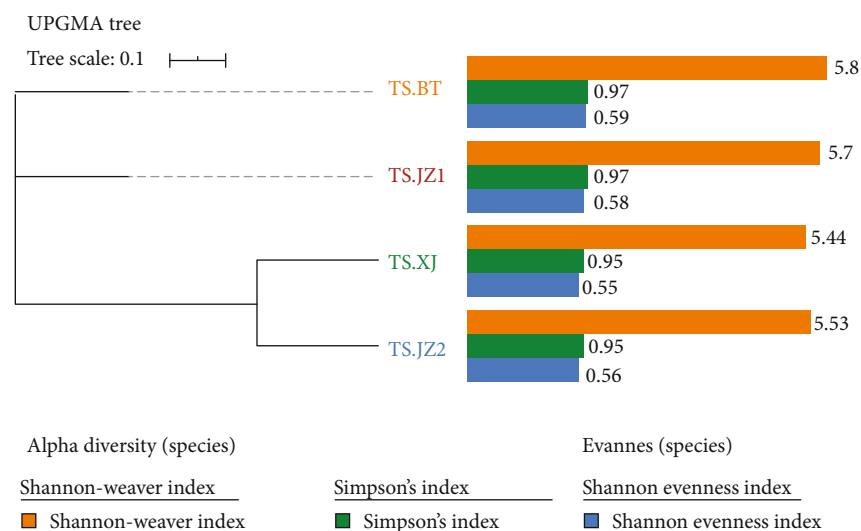
(a)

(b)

Figure 1: Continued.

(c)

Alpha diversity (species)

Shannon-weaver index
- 🟧 Shannon-weaver index

Simpson's index
- 🟩 Simpson's index

Evannes (species)

Shannon evenness index
- 🟦 Shannon evenness index

(d)

FIGURE 1: Taxonomic analysis of Taisui metagenome: (a) sample locations in northern China, (b) the relative abundance (%) of major families in Taisui samples, (c) the upset plot of the family count in each sample, and (d) alpha and beta diversities of Taisui metagenomes (the UPGMA clustering based on Bray-Curtis distances of species richness).

using pyani [27, 28]. According to the standards suggested by Glendinning et al. [29], genomes were defined as novel strains (GTDB-Tk ANI output < 99%), novel species (GTDB-Tk ANI output < 95%), and novel genera (all MAGs clustered at 60% AAI [30] without a genus GTDB-Tk assignment).

For MAG phylogenetic analysis, the concatenated sequence alignment of core marker genes was built by Up-to-date Bacterial Core Gene pipeline (UBCG) [31]. Conserved blocks were selected from multiple alignments by Gblocks (parameters: -b5 h) [32] for further analysis. The best-fit evolutionary model was selected using

ModelTest-NG [33]. The phylogenetic tree was inferred using RAxML-NG [34] with the GTR + I + G4 model. Node support was generated through 500 bootstrap replicates. The phylogenetic tree was finalized for publication using iTOL website [21].

For the phylogenetic placement analysis, the MAGs and reference genomes from NCBI database belonging to the phyla Planctomycetes and Chloroflexi were selected. To build the tree of Chloroflexi, we selected 15 genomes involving complete genomes of Anaerolineae, Ardenticatenia, and Dehalococcoidia (part of Dehalococcoidia complete

genomes were selected according to the results of GTDB-Tk), and closely related genome (UBA2991) of MAGs according to GTDB-Tk. And for the tree of Planctomycetes, we selected 48 genomes involving complete genomes of Planctomycetia, Phycisphaerales, and genomes of UBA1135, UBA1845, and UBA966 according to GTDB-Tk. All genomes used in phylogeny were evaluated by CheckM to confirm completeness≥70%. The concatenated core gene alignments were also performed through UBCG and were selected using Gblocks. Model selecting and phylogenetic tree bulling were performed as before with the GTR + I + G4 model and 100 bootstrap replicates.

2.6. The Phylogenetic Tree of PVA Dehydrogenase. We used our nonredundant protein sequences of PVA dehydrogenase and all known protein sequences of PVA dehydrogenase (AMG75031.1, Q588Z1.3, and P77931.1) from NCBI to build the phylogenetic tree. Only sequences with length more than 500 bp were involved. Sequences were aligned using MAFFT [35] and were trimmed using trimAL [36]. The best-fit evolutionary model was selected using ModelTest-NG [33]. A multiple sequence alignment program, MAFFT, includes two novel techniques: the progressive method (FFT-NS-2) and the iterative refinement method (FFT-NS-i). The MAFFT program package is freely available at http://www.biophys.kyoto-u.ac.jp/~katoh/programs/align/mafft. trimAl, a tool for automated alignment trimming in large-scale phylogenetic analyses, is freely available for download (http://trimal.cgenomics.org) and can be used online through the Phylemon web server (http://phylemon2.bioinfo.cipf.es/). ModelTest-NG is a reimplementation from scratch of jModelTest and ProtTest, two popular tools for selecting the best-fit nucleotide and amino acid substitution models, respectively. ModelTest-NG is available under a GNU GPL3 license at https://github.com/ddarriba/modeltest. The phylogenetic tree was inferred using RAxML-NG [34] with the LG + G4 model. RAxML-NG is a from-scratch reimplementation of the established greedy tree search algorithm of RAxML/ExaML. RAxML-NG offers improved accuracy, flexibility, speed, scalability, and usability compared with RAxML/ExaML. AxML-NG web service (maintained by Vital-IT) is available at https://raxml-ng.vital-it.ch/.Node support was generated through 100 bootstrap replicates. The phylogenetic tree was finalized for publication using iTOL website [21].

3. Results

We obtained shotgun metagenome sequence from four *Taisui* samples (located in Liaoning, Inner Mongolia, and Xinjiang province, Figure 1(a)). The metagenomic database provided 27.3 Gb clean data and 178 million high-quality reads in total for de novo assembling and binning (Table 1).

3.1. Taxonomic Annotation. The taxonomic annotation indicates the stabilization of four *Taisui* microbial communities. Proteobacteria was the dominant phylum with relative abundances from 59.55% to 71.82%. Five phyla, Planctomycetes, Acidobacteria, Chloroflexi, Actinobacteria, and Bacteroidetes, were subdominant populations (with abundances from 1% to 10%). The stabilization of *Taisui* communities was also found at other taxonomic levels. At the family level, most taxa (>80%) were common in all samples (Figure 1(c)). Twelve families—containing Caulobacteraceae, Sphingomonadaceae, Bradyrhizobiaceae, etc.—were enriched in samples with more than 1% relative abundance in at least one sample (Figure 1(b)). However, diversities emerged in family distribution. For example, TS.JZ1 had more than 2 fold abundance of Bradyrhizobiaceae compared with other samples; TS.BT and TS.XJ had more than 2 fold abundance of Sinobacteraceae compared with samples from Liaoning (Figure 1(b)).

Taisui from the same location (TS.JZ1 and TS.JZ2) had similar microbial composition. Simpson index in TS.JZ1 was the same as TS.JZ2, while Simpson index in TS.BT was the same as TS.XJ. Shannon index shows more diversity among samples, but indicates the same trend as Simpson index (Figure 1(d)). Besides, the UPGMA tree based on beta diversity illustrates TS.JZ1 and TS.JZ2 as one clade (Figure 1(d)).

3.2. Metagenome-Assembled Genomes

3.2.1. Assembly of 97 Draft Genomes. We generated 97 metagenome-assembled genomes (MAGs) with completeness ≥70% and contamination ≤5%. According to Bowers et al. [37], 65 MAGs were high-quality drafts. Among them, 43 bins were preferable with >95% completeness and <5% contamination, and 2 bins were almost complete with >97% completeness and 0% contamination. All MAGs had at least 6× average coverage depth, and 6 Proteobacteria MAGs had more than 100x coverage (Figure 2). Our MAGs were taxonomically assigned to 13 microbial phyla by GTDB-Tk (Figure 2). Notably, 2 MAGs belonged to the bacteria candidate phylum.

We obtained 95 putative novel species (ANI between MAG vs. reference genome < 95% and 67 MAGs without closest placement ANI), and 97 putative novel strains (ANI < 99%), according to the standard suggested by Glendinning et al. [29]. Besides, metagenomic databases provided 39 putative novel genera from 43 MAGs (MAGs which clustered at 60% mean AAI (amino acid identity) did not have a genus assignment by GTDB-Tk [30]). Putative novel genera belonged to 8 phyla involving Proteobacteria, Planctomycetes, Chloroflexi, Acidobacteria, Bacteroidota, Bdellovibrionota, Gemmatimonadota, and Verrucomicrobiota.

3.2.2. Novel Genus-Level Lineages. Two novel genus-level lineages were found in *Taisui* database, based on the standards: (1) they formed monophyletic lineages in the phylogeny, and (2) the average AAI was 60-80% between the genomes of such lineages and was <55% compared with known genera [30, 38].

In the tree of Planctomycetes members, we defined one genus-level novel lineage within the Planctomycetia class (Figure 3(a)). This lineage involved TS_28 (average coverage depth 24×) and Planctomycetales bacterium (GCF_009177095), with 66.4% AAI between members of it. The

TABLE 1: Summary of shotgun sequencing and assembly statistics.

|	TS.JZ1	TS.JZ2	TS.BT	TS.XJ
Clean data size (Gb)	6.4	6.7	7.1	7.1
Clean_Q20	98.66	97.48	97.09	97.52
Clean_Q30	96.55	93.37	92.36	93.41
Number of high quality paired-end reads	45,183,937	43,169,867	44,363,607	45,284,710
Total number of bases	6,777,590,550	6,475,480,050	6,654,541,050	6,792,706,500
Number of contigs (≥500 bp)	391,852	533,634	419,397	479,713
Total contig length (bp)	472,635,294	637,814,029	686,512,143	702,537,509
Average contig size (bp)	1206	1195	1637	1464
Contig N50 (bp)	1339	1287	2465	1910
Largest contig size (bp)	290,257	1,977,463	851,561	1,021,402
Number of transcripts (Prokka)	711,761	820,620	776,716	824,658
Average transcripts length (bp)	473	514	624	585
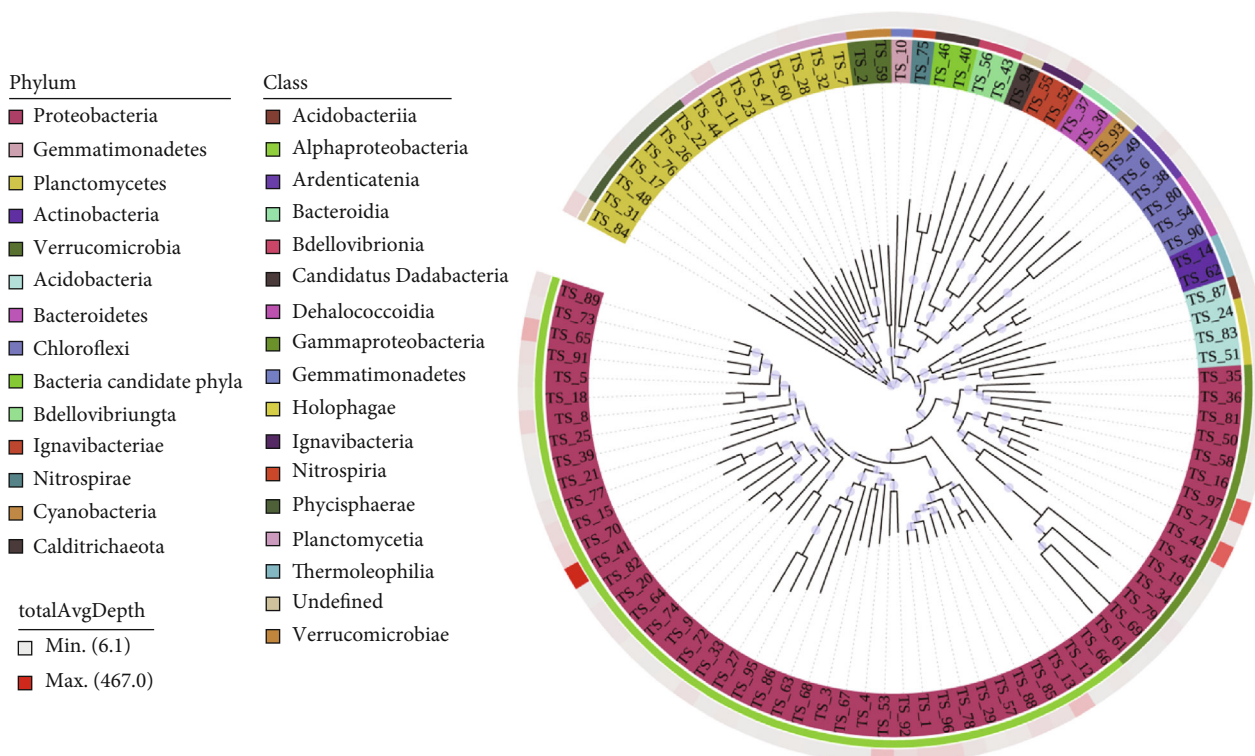
FIGURE 2: The maximum-likelihood tree among genomes recovered from Taisui samples. The MAGs were labelled by phylum and class, as assigned by GTDB-Tk. MAGs with "undefined" label were only able to be defined at phylum level. Branches with bootstrap that support more than 50% were denoted with a purple circle, which became larger as the bootstrap support value increases. The average coverage depth of all contigs in a bin was shown as heat map.

lineage—containing TS_28, GCF_009177095, TS_23 (7x), TS_47 (7x), and TS_60 (9x)—was a novel lineage at a higher taxonomic rank, because AAI between members ranged from 54.9% to 59.8%. Besides, the TS_84 (64x) may belong to a candidate novel class of Planctomycetes, which has shown AAI < 45% with other members of the Planctomycetes tree.

In the tree of Chloroflexi members, we defined one genus-level novel lineage within the Dehalococcoidia class (Figure 3(b)). It contained TS_80 (10x), TS_54 (13x), TS_90 (17x), and Dehalococcoidia bacterium UBA2991. The UBA2991 was identified in saline water as the unclassified Dehalococcoidia bacterium [39]. Members of this lineage show <49% AAI with genomes of Dehalococcoidales order and Dehalogenimonas order, which suggests this lineage was at least genus-level [30].

3.2.3. Novel Gene Clusters for Secondary Metabolite Biosynthesis. Among our MAGs, 503 biosynthetic gene clusters (BGCs) were identified from 12 microbial phyla
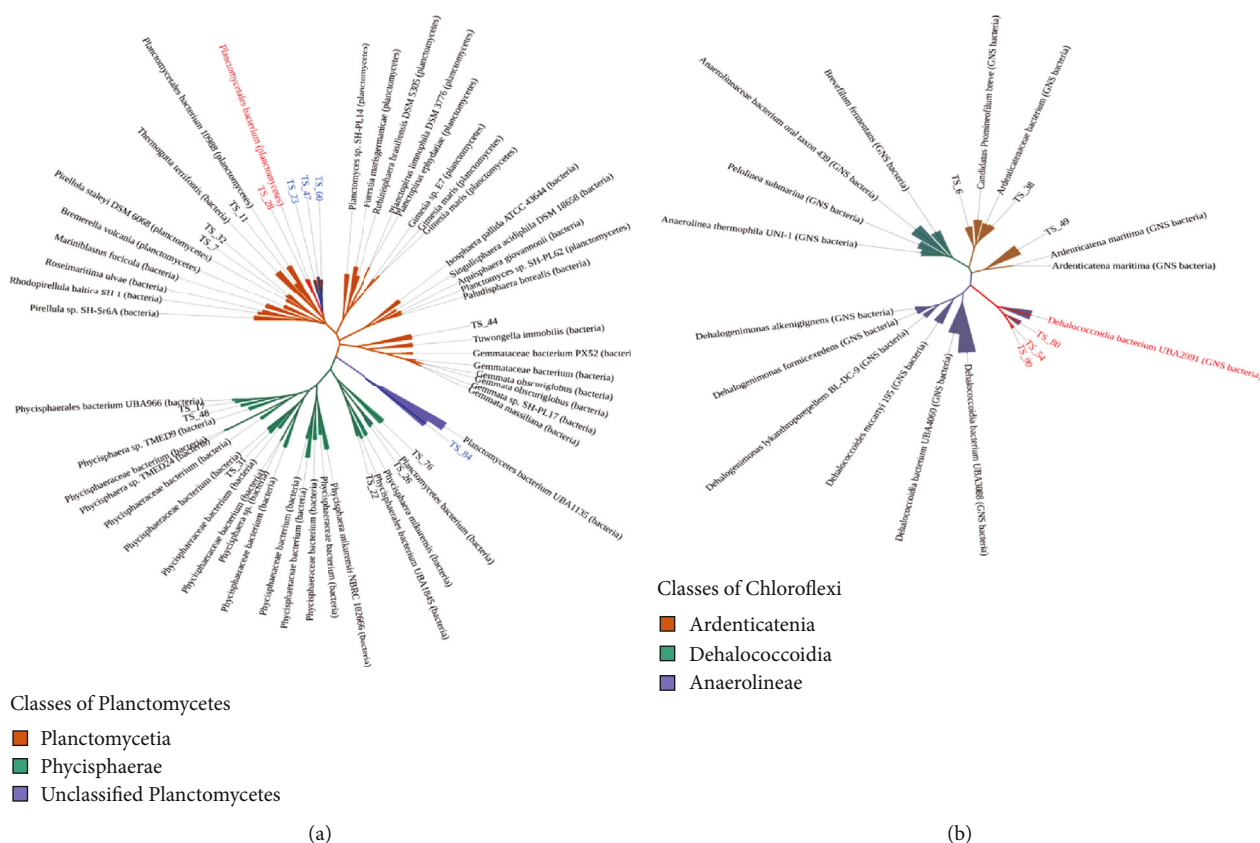
(a)

(b)

Figure 3: Placement of the Planctomycetes (a) and Chloroflexi (b) MAGs into their phylogenetic context. The trees included MAGs and published genomes with the completeness ≥ 70% and were inferred using RAxML-NG with 100 bootstrap replicates. Only branches with bootstrap support values ≥50% were shown. The MAGs were labelled at the class level. Genus-level novel branches were shown in red. Novel branches at higher level than genus were shown in purple.

(Figure 4(a)). Our BGCs included 22 kinds of types, containing polyketide synthases (PKSs), nonribosomal peptide synthases (NRPSs), synthases of linear azole-containing peptides (LAP), etc. Proteobacteria are linked with most types of BGCs ($n = 18$), followed by Planctomycetes ($n = 12$), Chloroflexi ($n = 10$), and Acidobacteria ($n = 10$).

MAGs provided BGCs that relate to new drug development. We identified 167 PKS (types I and III), NRPS, NRPS-like, and NRPS-PKS gene clusters from 8 phyla. Searching against the MIBiG database, 149 NRPS, NRPS-like, and PKS gene clusters had the potential structural divergence of their products with known biosynthetic genes, because these BGCs showed <60% similarity with known clusters [40].

In the Chloroflexi phylum, *Candidatus Promineofilum* sp. TS_38 had unusually large repertoires of NRPS (include NRPS-like clusters) or PKS regions. TS_38 contained 14 targeted biosynthetic loci (754.9 kbp in total length) in 7.3 Mbp contigs. The largest interested region was 108.2 kbp. Seven regions were complex NRPS-type I PKS systems (involving hybrid, neighboring, or interleaved) in TS_38 (Figure 4(b)).

In the Planctomycetes phylum, MAGs of novel lineage—including TS_28, TS_23, TS_47, and TS_60—were identified with PKS or NRPS loci. The phylogenetic tree reveals that the NRPS-like and T3PKS gene clusters were commonly observed in the closest reference genomes, but

the NRPS and T1PKS gene clusters were only observed in the *Taisui* MAGs (Figure 4(c)).

3.3. PVA Degradation

3.3.1. Novel Putative PVA Dehydrogenase Genes. *Taisui* metagenomic database provided 264 putative PVA dehydrogenase gene based on the KEGG database. Most of PVA dehydrogenase genes were derived from Proteobacteria (208), followed by Acidobacteria (29) and Actinobacteria (21). To understand the evolution of the PVA dehydrogenase genes, we built the phylogenetic tree using protein sequences of our putative genes (length > 500 bp) and all 3 PVA dehydrogenase sequences from NCBI. The phylogenetic tree shows that PVA dehydrogenase may have been ancestral in Proteobacteria, Acidobacteria, and Actinobacteria phyla (Figure 5). PVA dehydrogenase in Deltaproteobacteria members (including Deltaproteobacteria bacterium and *Phenylobacterium* sp.) may be gained via horizontal gene transfer from members of Actinobacteria. One species of Planctomycetes had PVA dehydrogenase, which was placed between clades of Actinobacteria.

3.3.2. Putative Pyrroloquinoline-Quinone Synthase Genes. In PVA degradation, pyrroloquinoline quinone (PQQ) is presumed to be needed for PVA dehydrogenase [2]. To identify
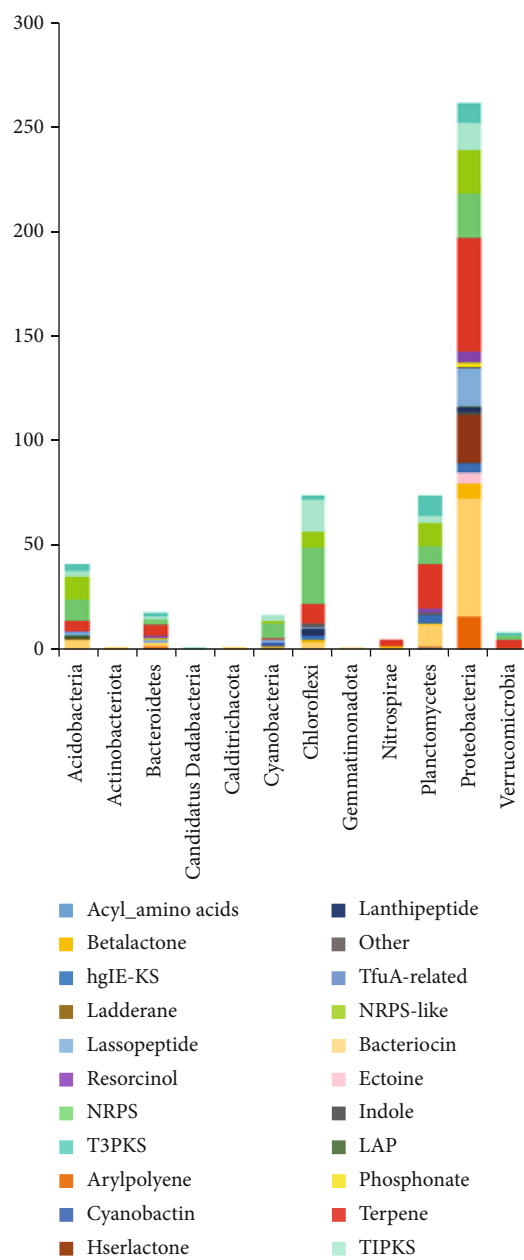
Acyl_amino acids · Lanthipeptide
Betalactone · Other
hgIE-KS · TfuA-related
Ladderane · NRPS-like
Lassopeptide · Bacteriocin
Resorcinol · Ectoine
NRPS · Indole
T3PKS · LAP
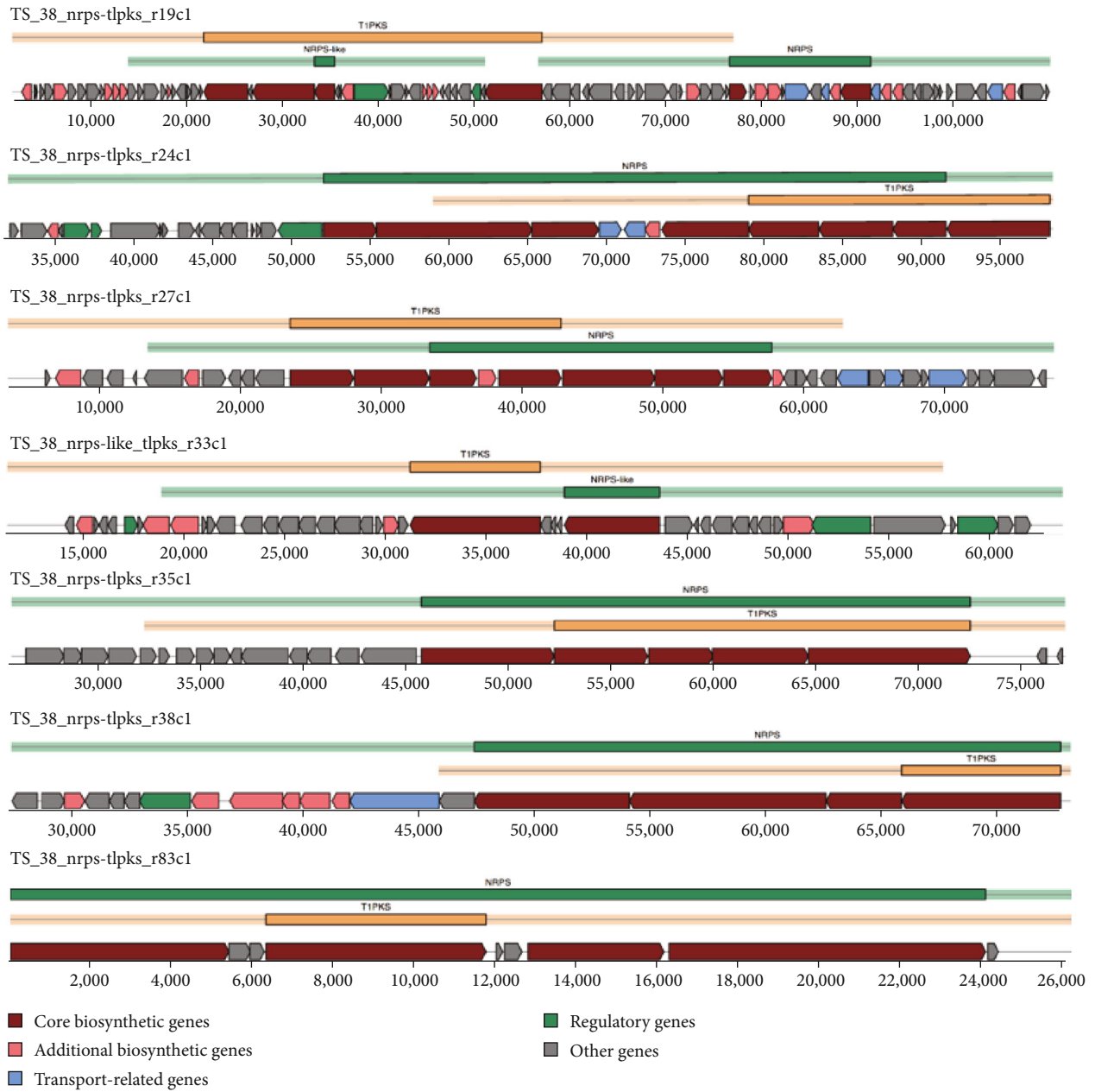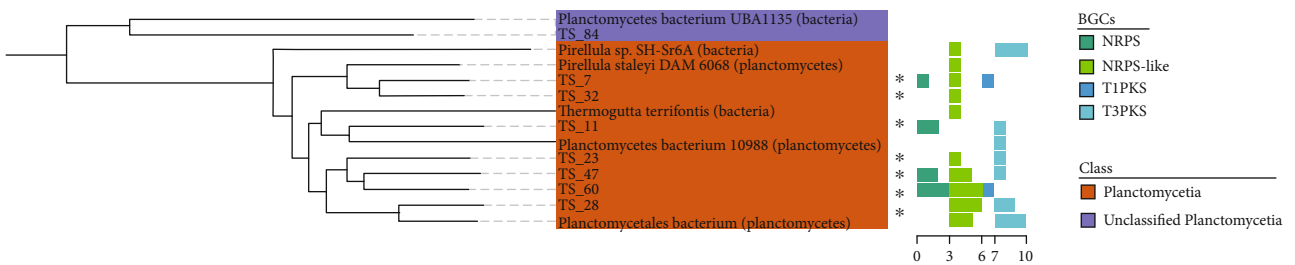Arylpolyene · Phosphonate
Cyanobactin · Terpene
Hserlactone · TIPKS

(a)

FIGURE 4: Continued.

FIGURE 4: Predicted biosynthetic gene clusters from the Taisui MAGs: (a) biosynthetic gene clusters (BGCs) identified on MAG sequences were summed as phyla on the y axis and were colored by product types as the results of antiSMASH, (b) seven large PKS-NRPS hybrid or neighboring or interleaved gene clusters in the Candidatus Promineofilum sp. TS_38 genome were shown, and (c) phylogenetic tree of Planctomycetia MAGs with predicted PKS or NRPS clusters (∗) and closest reference genomes of them (counts of PKS or NRPS clusters was shown in the bar plot).
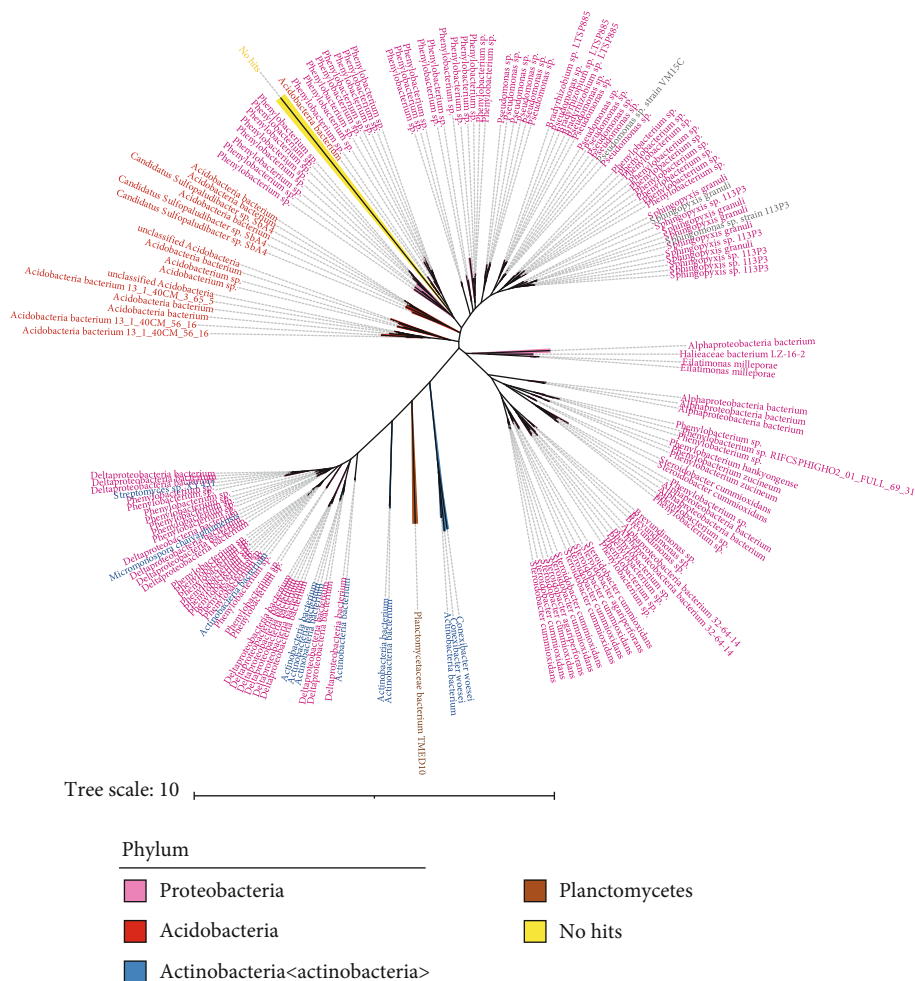
Figure 5: The maximum-likelihood tree of protein sequences of PVA dehydrogenase. Sequences from NCBI were shown in gray.

microbial symbiotic during PVA degradation, we assigned putative PQQ synthase genes to taxon. More than half of PQQ synthase genes were derived from Proteobacteria. *Bradyrhizobium icene* provided most amount of PQQ synthase (387), followed by *Chlorobi bacterium* OLB5 (164) and *Novosphingobium nitrogenifigens* (137). The result shows the difference of providers for PVA dehydrogenase and PQQ synthase in *Taisui* systems.

4. Discussion

4.1. The Relative Stable Microbial System in Taisui. In the present work, we gave the first standard of *Taisui* metagenome through shotgun sequencing of four samples. According to the analysis of 228 *Taisui* finding-reports, *Taisui* mainly exists in the soil layer of northern China [4]. Therefore, we chose four soil existed *Taisui* samples from northern China with stable morphology for better representation (microscopic observations in Figure S1). The taxonomic analysis illustrates a relatively consistent microbial community structure in different *Taisui* samples. *Taisui* samples from the same location had closer community structures. Biogeographic patterns of bacteria had been identified in soil metagenomes, which are more related to

environmental variation [41]. The taxonomic variation between *Taisui* samples may also be influenced by environmental variations, for example, precipitation (TS.JZ1 and TS.JZ2 were from subhumid regions; TS.BT was from the semiarid region; TS.XJ was from the arid region). Besides, dominant taxa in *Taisui* were common taxa in soil metagenomes. Marker gene amplification of *Taisui* also illustrates the same phenomenon [3]. It is reasonable because we used soil existed *Taisui*.

The functional annotation shows almost the same relative abundances of functional classifications of *Taisui* samples (Figure S2 and S3). The functional stability was also identified in human gut microbiota, which indicates the existence of a relatively stable ecological system [42]. Reconstruction of meta-pathway illustrated *Taisui* communities as carbon fixation, nitrogen fixation, and biosynthesis (Figure 6). At the family level, Caulobacteraceae, Sinobacteraceae, and Sphingomonadaceae contributed more than 30% related genes of the reductive citric acid cycle (rTCA cycle), while Burkholderiales, Xanthobacteraceae, Sphingomonadaceae, and Bradyrhizobiaceae contributed almost half of the N-fixation enzymes on average. The difference between functional providers of carbon fixation and nitrogen fixation may suggest the symbiosis between microbes
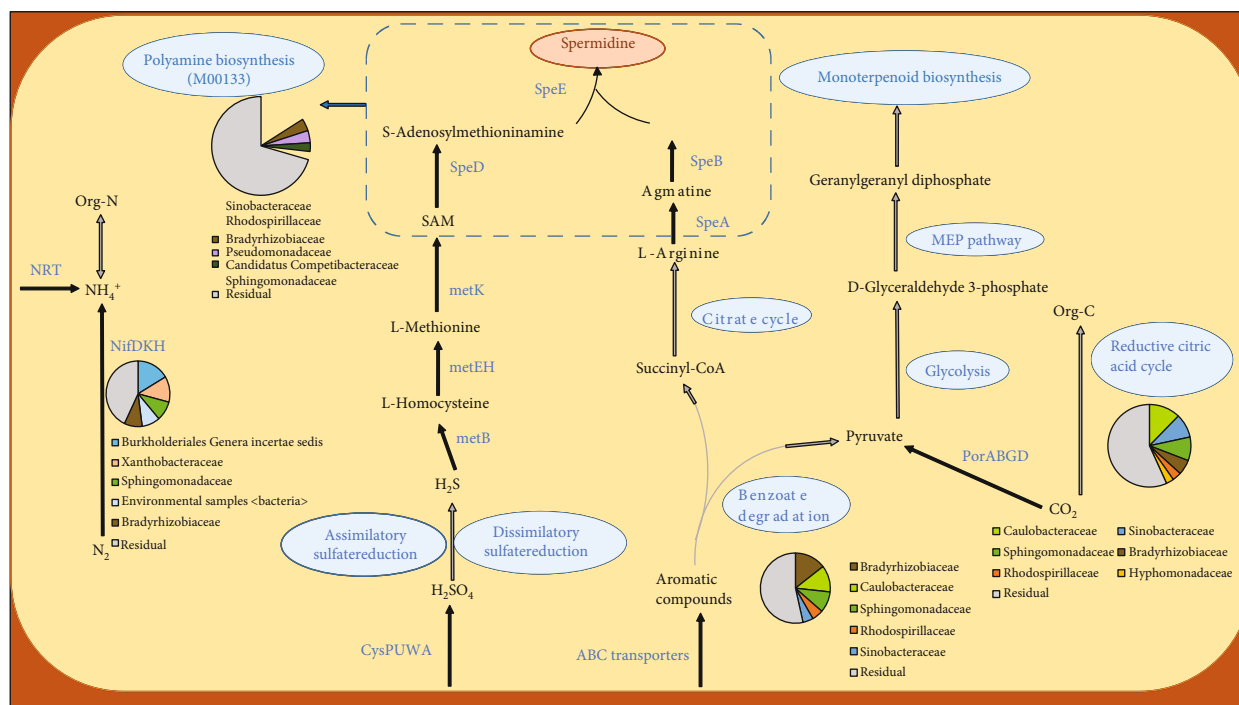
FIGURE 6: The sketch map of the metabolic pathways reconstructed from genes of Taisui samples (based on Figure S4–S7). The pie charts, which beside the names of enzymes or pathways, show the contribution ratio of taxa at family level. Abbreviations: NRT: ammonium transporter; NifDKH: nitrogenase; metB: cystathionine gamma-synthase; metE: 5-methyltetrahydropteroyltriglutamate–homocysteine methyltransferase; metH: 5-methyltetrahydrofolate–homocysteine methyltransferase; metK: S-adenosylmethionine synthetase; SepD: S-adenosylmethionine decarboxylase; SepA: arginine decarboxylase; SepB: agmatinase; SepE: spermidine synthase; PorABDG: pyruvate ferredoxin oxidoreductase; SAM: S-Adenosylmethionine; CysPUWA: sulfate transporters; ABC transporters: ATP-binding cassette transporters; MEP: 2-C-methyl-D-erythritol-4-phosphate.

in the *Taisui* community. However, more samples are needed for reliable cooccurrence network analysis [43].

4.2. Expanded Diversity of Microbial Genome and Biosynthetic Gene Cluster.

Historical changes in population size, such as those caused by demographic range expansions, can produce nonadaptive changes in genomic diversity through mechanisms such as gene surfing Microbial species and their populations exhibit remarkable genomic diversity [44]. While mutation and recombination promote genetic variation in all forms of life, the genomic diversity of Bacteria and Archaea is enhanced dramatically by their proclivity for Horizontal Gene Transfer (HGT). Genomic analyses of diverse microbes provide similar results and it seems that a majority of genes in any pan-genome will be comprised of either high-frequency core genes or low-frequency strain-specific genes. These patterns of genomic diversity reveal the fundamental impact of HGT on evolution, and they suggest that bacterial and archaeal genomes comprise a dynamic mosaic of horizontally acquired genes whose frequency fluctuates in the population in response to both selection and genetic drift.

Our MAGs filled in some phylogenetic gaps and could be valuable in the detail inferring of phylogenetic relationships in bacterial. The MAGs for novel genera include taxa with very little genomic information published. One novel genus belonging to Proteobacteria was assigned to the Steroidobacteraceae family, and this family has only 9 published

genomes in NCBI. In the Chloroflexi phylum, we identified one novel genus in each of the classes Dehalococcoidia and Aredenticatenia. The Dehalococcoidia contain only 2 formally published genera, involving Dehalococcoides and Dehalogenimonas [45]. And the Aredenticatenia class only had been identified in sludge and hydrothermal field [46–49] with 8 published genome on NCBI. In the Acidobacteria phylum, 3 novel genera of the Holophagales order were identified. The Holophagales order only contains the genera *Holophaga* and *Geotes* with a wide range of uncultured bacteria mainly from marine and soil [50].

Our MAGs also included novel gene clusters of NRPS, NRPS-like, and PKS (89.2% gene clusters < 60% similarity with known BGCs). Most of identified BGCs were linked with Proteobacteria, which is a common producer of wide bacterial natural products [51–53]. But few BGCs were reported in the published genomes of bacteria candidate phyla [54, 55]. We expand the number of taxa in Candidatus Dadabacteria phylum encoding BGCs. Besides, 2 MAGs of the Ardenticatenaceae family (belonging to Chloroflexi phylum) were detected with large NRPS or PKS loci, which were referred to as "*Candidatus Promineofilum* sp. TS_6" and "*Candidatus Promineofilum* sp. TS_38." The BGCs linked to Ardenticatenaceae were few (6 linked clusters in total), according to the IMG database [54]. We expand Ardenticatenaceae with potential productions of PKS and NRPS. The large NRPS and PKS loci may provide source for new drug finding.

According to the phylogenetic tree of Planctomycetes members, NRPS and T1PKS were only observed in *Taisui* MAGs (novel lineage including TS_23, TS_47, TS_60, and TS_28). This result suggests that these two types of BGCs were acquired independently in evolutionary time in these taxa. The products of BGCs were used for the competition or communication with the producers' environment, and the abundance of NRPS and PKS domains was influenced by environmental conditions, involving soil depth, latitude, moisture, etc. [55, 56]. Therefore, the unconventional environment offered by *Taisui* may cause the encoding of NRPS and T1PKS gene clusters in these MAGs.

4.3. Rich Diversity of PVA Dehydrogenase Genes in the Taisui Microbial Community. PVA degradation includes two steps: first, the conversion from the 1,3-glycol structure of two successive repeating units to the beta-diketone; second, the broken of the carbon-carbon bond and the conversion of ketone group to carboxylic group [57]. In *Taisui* systems, the first step may begin with PQQ and PVA dehydrogenase, according to the functional annotation. *Taisui* has high water content, so the carbon-carbon bond may be cleavage by oxidized PVA hydrolase with H_2O participation, like the model illustrated in methylotrophic yeast [58].

Microbes with PVA degradation ability are rare. Most PVA-degraders belong to the Pseudomonas and Sphingomonas genera. Although with novel degraders identified from grapes and marine bacterium [2], the taxonomic survey of PVA-degraders is not fruitful. Our results of putative PVA-degraders from 4 phyla can be a markable expansion for PVA biodegradation. Genera involving *Phenylobacterium*, *Conexibacter*, *Steroidobacter*, *Brevundimonas*, *Eilatimonas*, *Sphingopyxis*, and *Acidobacterium* may also be the source of PVA-degraders.

For PQQ-dependent PVA dehydrogenase, PVA degradation begins with the action of PQQ [2]. The production of PQQ by microbes (involving *Bradyrhizobium*, *Novosphingobium*, *Rhodopseudomonas*, etc.) can enhance the rate of PVA degradation. Besides, PVA dehydrogenase and PQQ synthase were abundant in different species, which suggests the microbial dependency on PVA degradation in *Taisui*.

5. Conclusion

According to metagenomic analysis, we concluded that the community structure of polymeric compound *Taisui* was relatively stable. The high abundances of Proteobacteria, Acidobacteria, Chloroflexi, Actinobacteria, and Bacteroidetes in *Taisui* can also be identified in the soil microbial community, suggesting the close relationship of *Taisui* and soil. As a unexplored system, *Taisui* communities provide genomes of previously poorly sampled microbial lineages, which is a valuable step for a comprehensive picture of the evolutionary history of life. For new drug development, BGCs of PKS and NRPS were expanded in phyla including Candidatus Dadabacteria, Chloroflexi, and Planctomycetes. The substantial putative PVA dehydrogenase genes were identified in 4 phyla, suggesting rich diversity of PVA dehydrogenase genes in *Taisui* communities. The gene of PVA

degradation in microbes may acquire independently or from horizontal gene transfer. And the PQQ providers may enhance the PVA degradation rate in *Taisui*.

Data Availability

Sequences in this paper have been submitted to NCBI with SRA accession numbers SRR8569139-SRR8569142.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Jiaxing Liu did the conceptualization, methodology, investigation, writing—original draft, and funding acquisition. Jiaxing Liu did the validation, writing—review and editing. Jiaxing Liu did the visualization. Jiaxing Liu did the supervision and writing—review and editing. Xun Gu did the conceptualization, supervision, and writing—review and editing. Hui Li did the conceptualization, writing—review and editing, supervision, and funding acquisition.

Acknowledgments

Supplementary Materials

Figure S1: microscopic observations of *Taisui* tissues. (a) The histological imprints of a piece of *Taisui* tissue. (b) The histological sections stained with hematoxylin and eosin (HE) of *Taisui* tissues. Figure S2: the relative abundance (%) of KEGG based annotations in *Taisui* samples. The relative abundances of the level 1 classification of KEGG in four samples. (b) The relative abundances of the level 2 classification of KEGG in four samples (classifications with relative abundance above 1% are shown). Figure S3: the relative abundance (%) of COG-based annotations in *Taisui* samples. The relative abundances of the level 1 classification of COG in four samples. (b) The relative abundances of the level 2 classification of COG in four samples. Figure S4: the enrichment of carbon fixation pathways in four *Taisui* samples. Each node was divided into four blocks vertically, representing TS.JZ1, TS.JZ2, TS.BT, and TS.XJ in order from left to right. The colors of blocks represent the count of transcripts that annotated to this function, and the red block represents the transcripts number ≥500. Figure S5: the enrichment of benzoate degradation pathways in four *Taisui* samples. Each node was divided into four blocks vertically, representing TS.JZ1, TS.JZ2, TS.BT, and TS.XJ in order from left to right. The colors of blocks represent the count of transcripts that annotated to this function, and the red block represents the transcripts number ≥500. Figure S6: the enrichment of nitrogen metabolism pathways in four *Taisui*

samples. Each node was divided into four blocks vertically, representing TS.JZ1, TS.JZ2, TS.BT, and TS.XJ in order from left to right. The colors of blocks represent the count of transcripts that annotated to this function, and the red block represents the transcripts number ≥500. Figure S7: the enrichment of terpenoid backbone biosynthesis pathways in four *Taisui* samples. Each node was divided into four blocks vertically, representing TS.JZ1, TS.JZ2, TS.BT, and TS.XJ in order from left to right. The colors of blocks represent the count of transcripts that annotated to this function, and the red block represents the transcripts number ≥500. *(Supplementary Materials)*

References

[1] M. Amann and O. Minge, "Biodegradability of poly(vinyl acetate) and related polymers," in *Synthetic Biodegradable Polymers, Advances in Polymer Science*, B. Rieger, A. Künkel, G. W. Coates, R. Reichardt, E. Dinjus, and T. A. Zevaco, Eds., pp. 137–172, Springer, Berlin, Heidelberg, 2011.

[2] N. B. Halima, "Poly(vinyl alcohol): review of its promising applications and insights into biodegradation," *RSC Advances*, vol. 6, no. 46, pp. 39823–39832, 2016.

[3] E. Li, J. Ren, Q. Chen et al., "Uncovering the mysterious identity of Taisui—an old Chinese folk legend," *Science China. Life Sciences*, vol. 63, no. 12, pp. 1942–1945, 2020.

[4] C. Wang and S. Wang, "A research of the finding and distribution law of Taisui in modern China," *Agricultural Sciences*, vol. 6, no. 4, pp. 407–414, 2015.

[5] M. Albertsen, P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen, "Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes," *Nature Biotechnology*, vol. 31, no. 6, pp. 533–538, 2013.

[6] L.-X. Chen, K. Anantharaman, A. Shaiber, A. M. Eren, and J. F. Banfield, "Accurate and complete genomes from metagenomes," *Genome Research*, vol. 30, no. 3, pp. 315–333, 2020.

[7] M. R. Green and J. Sambrook, "Isolation of high-molecular-weight DNA using organic solvents," *Cold Spring Harbor Protocols*, vol. 2017, no. 4, 2017.

[8] R. K. Patel and M. Jain, "NGS QC toolkit: a toolkit for quality control of next generation sequencing data," *PLoS One*, vol. 7, no. 2, article e30619, 2012.

[9] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047-3048, 2016.

[10] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph," *Bioinformatics*, vol. 31, no. 10, pp. 1674–1676, 2015.

[11] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "QUAST: quality assessment tool for genome assemblies," *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, 2013.

[12] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, no. 14, pp. 2068-2069, 2014.

[13] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[14] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658-1659, 2006.

[15] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature Methods*, vol. 14, no. 4, pp. 417–419, 2017.

[16] B. Buchfink, C. Xie, and D. H. Huson, "Fast and sensitive protein alignment using DIAMOND," *Nature Methods*, vol. 12, no. 1, pp. 59-60, 2015.

[17] D. H. Huson, S. Beier, I. Flade et al., "MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data," *PLoS Computational Biology*, vol. 12, no. 6, article e1004957, 2016.

[18] J. Huerta-Cepas, D. Szklarczyk, K. Forslund et al., "eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences," *Nucleic Acids Research*, vol. 44, no. D1, pp. D286–D293, 2016.

[19] B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat, "The carbohydrate-active enzymes database (CAZy): an expert resource for glycogenomics," *Nucleic Acids Research*, vol. 37, no. Database, pp. D233–D238, 2009.

[20] D. H. Huson and D. Bryant, "Application of phylogenetic networks in evolutionary studies," *Molecular Biology and Evolution*, vol. 23, no. 2, pp. 254–267, 2006.

[21] I. Letunic and P. Bork, "Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees," *Nucleic Acids Research*, vol. 44, no. W1, pp. W242–W245, 2016.

[22] W. Luo and C. Brouwer, "Pathview: an R/Bioconductor package for pathway-based data integration and visualization," *Bioinformatics*, vol. 29, no. 14, pp. 1830-1831, 2013.

[23] G. V. Uritskiy, J. DiRuggiero, and J. Taylor, "MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis," *Microbiome*, vol. 6, no. 1, p. 158, 2018.

[24] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes," *Genome Research*, vol. 25, no. 7, pp. 1043–1055, 2015.

[25] P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks, "GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database," *Bioinformatics*, vol. 36, pp. 1925–1927, 2019.

[26] D. Parks, "CompareM," 2016, https://github.com/dparks1134/CompareM.

[27] R. A. Barco, G. M. Garrity, J. J. Scott, J. P. Amend, K. H. Nealson, and D. Emerson, "A genus definition forBacteriaandArchaeaBased on a standard genome relatedness index," *mBio*, vol. 11, no. 1, 2020.

[28] L. Pritchard, R. H. Glover, S. Humphris, J. G. Elphinstone, and I. K. Toth, "Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens," *Analytical Methods*, vol. 8, no. 1, pp. 12–24, 2015.

[29] L. Glendinning, R. D. Stewart, M. J. Pallen, K. A. Watson, and M. Watson, "Assembly of hundreds of novel bacterial genomes from the chicken caecum," *Genome Biology*, vol. 21, no. 1, p. 34, 2020.

[30] C. Luo, L. M. Rodriguez-R, and K. T. Konstantinidis, "MyTaxa: an advanced taxonomic classifier for genomic and

metagenomic sequences," *Nucleic Acids Research*, vol. 42, no. 8, pp. e73–e73, 2014.

[31] S.-I. Na, Y. O. Kim, S.-H. Yoon, S. Ha, I. Baek, and J. Chun, "UBCG: up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction," *Journal of Microbiology*, vol. 56, no. 4, pp. 280–285, 2018.

[32] J. Castresana, "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis," *Molecular Biology and Evolution*, vol. 17, no. 4, pp. 540–552, 2000.

[33] D. Darriba, D. Posada, A. M. Kozlov, A. Stamatakis, B. Morel, and T. Flouri, "ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models," *Molecular Biology and Evolution*, vol. 37, no. 1, pp. 291–294, 2020.

[34] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis, "RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference," *Bioinformatics*, vol. 35, no. 21, pp. 4453–4455, 2019.

[35] J. Rozewicki, S. Li, K. M. Amada, D. M. Standley, and K. Katoh, "MAFFT-DASH: integrated protein sequence and structural alignment," *Nucleic Acids Research*, vol. 47, pp. W5–W10, 2019.

[36] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, "trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses," *Bioinformatics*, vol. 25, no. 15, pp. 1972-1973, 2009.

[37] R. M. Bowers, N. C. Kyrpides, R. Stepanauskas et al., "Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea," *Nature Biotechnology*, vol. 35, no. 8, pp. 725–731, 2017.

[38] N. K. Dudek, C. L. Sun, D. Burstein et al., "Novel microbial diversity and functional potential in the marine mammal oral microbiome," *Current Biology*, vol. 27, no. 24, pp. 3752–3762.e6, 2017.

[39] D. H. Parks, C. Rinke, M. Chuvochina et al., "Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life," *Nature Microbiology*, vol. 2, no. 11, pp. 1533–1542, 2017.

[40] A. Crits-Christoph, S. Diamond, C. N. Butterfield, B. C. Thomas, and J. F. Banfield, "Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis," *Nature*, vol. 558, no. 7710, pp. 440–444, 2018.

[41] S. K. Bay, M. A. McGeoch, O. Gillor et al., "Soil bacterial communities exhibit strong biogeographic patterns at fine taxonomic resolution," *Msystems*, vol. 5, no. 4, pp. e00540–e00620, 2020.

[42] The Human Microbiome Project Consortium, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.

[43] D. Berry and S. Widder, "Deciphering microbial interactions and detecting keystone species with co-occurrence networks," *Frontiers in Microbiology*, vol. 5, pp. 1–14, 2014.

[44] M. J. Choudoir, K. Panke-Buisse, C. P. Andam, and D. Buckley, "Genome surfing as driver of microbial genomic diversity," *Trends in microbiology*, vol. 25, no. 8, pp. 624–636, 2017.

[45] Y. Yang, R. Sanford, J. Yan et al., "Roles of Organohalide-Respiring *Dehalococcoidia* in carbon cycling," *Msystems*, vol. 5, no. 3, pp. e00757–e00819, 2020.

[46] J. Hemp, L. M. Ward, L. A. Pace, and W. W. Fischer, "Draft genome sequence of Ardenticatena maritima 110S, a thermo-philic nitrate- and iron-reducing member of the Chloroflexi class Ardenticatenia," *Genome Announcements*, vol. 3, no. 6, pp. e01347–e01415, 2015.

[47] S. Kawaichi, N. Ito, R. Kamikawa, T. Sugawara, T. Yoshida, and Y. Sako, "Ardenticatena maritima gen. nov., sp. nov., a ferric iron- and nitrate-reducing bacterium of the phylum ' Chloroflexi ' isolated from an iron-rich coastal hydrothermal field, and description of Ardenticatena classis nov," *International Journal of Systematic and Evolutionary Microbiology*, vol. 63, Part 8, pp. 2992–3002, 2013.

[48] L. M. Ward, S. E. McGlynn, and W. W. Fischer, "Draft genome sequence of a divergent anaerobic member of the *Chloroflexi* class *Ardenticatena* from a sulfidic hot spring," *Genome Announcements*, vol. 6, no. 25, pp. e00571–e00618, 2018.

[49] M. Zhang, J. Yao, X. Wang, Y. Hong, and Y. Chen, "The microbial community in filamentous bulking sludge with the ultra-low sludge loading and long sludge retention time in oxidation ditch," *Scientific Reports*, vol. 9, no. 1, article 13693, 2019.

[50] Y. Fukunaga, M. Kurahashi, K. Yanagi, A. Yokota, and S. Harayama, "Acanthopleuribacter pedis gen. nov., sp. nov., a marine bacterium isolated from a chiton, and description of Acanthopleuribacteraceae fam. nov., Acanthopleuribacter-ales ord. nov., Holophagaceae fam. nov., Holophagales ord. nov. and Holophagae classis nov. in the phylum 'Acidobacteria'," *International Journal of Systematic and Evolutionary Microbiology*, vol. 58, no. 11, pp. 2597–2601, 2008.

[51] C. A. Brotherton, M. H. Medema, and E. P. Greenberg, "*luxR* Homolog-Linked biosynthetic gene clusters in *Proteobacteria*," *MSystems*, vol. 3, no. 3, pp. e00208–e00217, 2018.

[52] Y. Buijs, P. K. Bech, D. Vazquez-Albacete et al., "Marine Proteobacteria as a source of natural products: advances in molecular tools and strategies," *Natural Product Reports*, vol. 36, no. 9, pp. 1333–1350, 2019.

[53] P. Cimermancic, M. H. Medema, J. Claesen et al., "Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters," *Cell*, vol. 158, no. 2, pp. 412–421, 2014.

[54] K. Palaniappan, I.-M. A. Chen, K. Chu et al., "IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase," *Nucleic Acids Research*, vol. 48, pp. D422–D430, 2020.

[55] A. M. Sharrar, A. Crits-Christoph, R. Méheust, S. Diamond, E. P. Starr, and J. F. Banfield, "Bacterial secondary metabolite biosynthetic potential in soil varies with phylum, depth, and vegetation type," *mBio*, vol. 11, no. 3, pp. e00416–e00420, 2020.

[56] Z. Charlop-Powers, J. G. Owen, B. V. B. Reddy, M. A. Ternei, and S. F. Brady, "Chemical-biogeographic survey of secondary metabolism in soil," *Proceedings of the National Academy of Sciences*, vol. 111, no. 10, pp. 3757–3762, 2014.

[57] K. Leja and G. Lewandowicz, "Polymer biodegradation and biodegradable polymers - a review," *Polish Journal of Environmental Studies*, vol. 19, no. 2, pp. 255–266, 2010.

[58] D. Jia, J. Li, L. Liu et al., "High-level expression, purification, and enzymatic characterization of truncated poly(vinyl alcohol) dehydrogenase in methylotrophic yeast Pichia pastoris," *Applied Microbiology and Biotechnology*, vol. 97, no. 3, pp. 1113–1120, 2013.