



OPEN

A metagenomic catalog for exploring the plastizymes landscape covering taxa, genes, and proteins

Donya Afshar Jahanshahi^{1,2}, Shohreh Ariaeenejad³ & Kaveh Kavousi^{2✉}

There are significant environmental and health concerns associated with the current inefficient plastic recycling process. This study presents the first integrated reference catalog of plastic-contaminated environments obtained using an insilico workflow that could play a significant role in discovering new plastizymes. Here, we combined 66 whole metagenomic data from plastic-contaminated environment samples from four previously collected metagenome data with our new sample. In this study, an integrated plastic-contaminated environment gene, protein, taxa, and plastic degrading enzyme catalog (PDEC) was constructed. These catalogs contain 53,300,583 non-redundant genes and proteins, 691 metagenome-assembled genomes, and 136,654 plastizymes. Based on KEGG and eggNOG annotations, 42% of recognized genes lack annotations, indicating their functions remain elusive and warrant further investigation. Additionally, the PDEC catalog highlights hydrolases, peroxidases, and cutinases as the prevailing plastizymes. Ultimately, following multiple validation procedures, our effort focused on pinpointing enzymes that exhibited the highest similarity to the introduced plastizymes in terms of both sequence and three-dimensional structural aspects. This encompassed evaluating the linear composition of constituent units as well as the complex spatial conformation of the molecule. The resulting catalog is expected to improve the resolution of future multi-omics studies, providing new insights into plastic-pollution related research.

Polymer products are used worldwide, and at least 350–400 million tons are produced annually. Plastics are extensively consumed in the global economy for several reasons. Environmental pollution caused by plastics has become an intense ecological obstacle worldwide¹. Plastics affect ecosystems, and all types of plastics are of particular concern to the health of humans and other living organisms on Earth. Many microorganisms and enzymes have been characterized for plastic degradation using metagenomic approaches². The metagenomes of plastic-contaminated environments contain a wide variety of genomic contents, including bacteria, fungi, viruses, and other microorganisms. These organisms have developed adaptations that enable them to thrive in environments contaminated with plastic, and their genetic materials include genes and enzymes that facilitate the breakdown and metabolism of plastics. Uncultured techniques like shotgun metagenome sequencing have transformed pilot methods to identifying and investigating these communities^{3–5}. The environmental microbiome (such as soil and marine samples) is a highly variable microbial ecosystem. Numerous microbes, including bacteria and archaea, colonize plastic contaminated environmental samples (landfill, agricultural land covered with plastic mulch, wastewater treatment plants, and all places that are exposed to the plastic particles^{6–10}) and may play fundamental roles in degradation of recalcitrant synthetic polymers such as caprolactam (CPL is an organic compound with approximately five million tons global demand annually and is widely used to make plastics and nylon^{11,12}), polyethylene (PE)^{9,13}, polyethylene terephthalate (PET)^{14–17} polyamide (PA)^{18–20} and ester-based polyurethane (PUR)^{21,22}, etc.^{23–25}.

Enzymes represent vital biological catalysts with diverse characteristics, playing a pivotal role in various biotechnological applications^{8,26–28}. Analysis of the microbiome affected by plastic contamination is crucial in identifying microorganisms and enzymes that enhance plastic breakdown. Catalogs of genes, proteins, and

¹Department of Bioinformatics, Kish International Campus University of Tehran, Kish, Iran. ²Laboratory of Complex Biological Systems and Bioinformatics (CBB), Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran. ³Department of Systems and Synthetic Biology, Agricultural Biotechnology Research Institute of Iran (ABRII), Agricultural Research Education and Extension Organization (AREEO), Karaj, Iran. ✉email: kkavousi@ut.ac.ir; kkavousi@yahoo.com

taxa facilitate taxonomy and function categorization. The Carbohydrate-Active Enzymes (CAZy) database is a valuable and comprehensive source of enzyme information. As plastic-degrading enzymes primarily fall under the category of carbohydrate-active enzymes (CAZymes), investigation of these clusters could provide insights into plastic-degrading enzyme abundance. Enzymes identified in the degradation of polymers have belonged to PETase, MHETase, Cutinase, Laccase, Manganese Peroxidase, Lipase, Carboxylesterase, Hydrolase, and so on^{9,13,21,29–36}.

Despite our growing knowledge of plastic-contaminated environment microbiomes, the complex functions of plastic-contaminated environment microbiota are still unclear. These microbiome samples were due to the absence of reference genomes and gene catalogs.

Numerous reference gene catalogs of gut microbiomes have been reported for humans^{37–39} and animals such as mice⁴⁰, sheep⁴¹, chicken^{42,43}, soil⁴⁴ and camel rumen⁴⁵. In these catalogs, the distribution of microorganisms, genes, and metabolic pathways was analyzed and investigated, depending on the need for the study.

The Study of plastic-contaminated environments is a novel research field with limited understanding of microbial diversity. To date, no integrated gene and MAGs catalog has been developed. Although some studies have examined microbial diversity in these environments, an integrated gene and taxa catalog is essential to further our understanding of this system. This will enable both taxonomic and functional profiling of the plastic-contaminated environment.

In 2017, a study reported the soil microbial diversity and functional aspects using metagenomic and bioinformatic approaches in landfill Lysimeter soil of the Ghazipur Landfill Site, New Delhi, India⁴⁶. This study aimed to report the functional and taxonomical profiling of the Ghazipur Landfill. In 2021, Kumar et al. collected soil, leachate, and compost samples, from 10 different locations (height and depth) at the Pirana landfill site in Gujarat, India. They specified the relationship between microorganism diversity and the gene reservoirs involved in the plastic degradation process in landfill environments⁴⁷. Another study in 2020 determined the microbial communities associated with six samples of plasticized fabric materials exposed to a harsh tropical environment for 14 months in the Republic of Panama⁴⁸ and⁴⁹. Metagenomic analysis of 161 Gb sequence resulted in ~ 3 million contigs and 120 MAGs. In 2017, Chu's group revealed the effects of wastewater treatment plants on the scattering of microorganisms, genes, and antibiotic-resistance genes. In July 2015, 48 samples were collected from two wastewater treatment plants in Wisconsin, USA⁵⁰.

This study aimed to gain a better understanding of plastic-contaminated environments by combining four previously analyzed plastic-contaminated metagenomes, primarily environmental samples from Ghazipur, Gujarat, Panama, and Wisconsin, with a new soil sample collected from agricultural lands under mulch cultivation and agricultural lands irrigated with municipal wastewater.

This research presents a comprehensive overview of an integrated plastic-contaminated environment reference catalog that includes taxa, genes, proteins, and plastic-degrading enzymes (plastizymes). We developed this multi-functional catalog using bioinformatics tools and workflows; such as FastQC, SOAPaligner2, MEGAHIT, MetaBAT2, and MetaGeneMark. This resource is essential for gaining insights into the structure and functions of the plastic-contaminated environment. The generated catalogs can also be used to identify the characteristics of plastic-polluted environments and to attempt to remediate them for environmental conservation.

Methods

Metagenomic samples collection. The collection of metagenomic data from agricultural land under mulch cultivation and agricultural with municipal wastewater was important. The selected soil samples were contaminated with plastic residue for approximately 35 years. These lands are situated in Varamin and Ghaleh-No village in Tehran province with geographical coordinates (35.1848° N latitude, 51.4214° E) and (35.465° N latitude, 51.594° E).

Ten locations were sampled at various depths from the ground (5 to 20 cm), and the samples were stored in glass containers with dry ice to maintain required temperature. The collected samples were combined into one, filtered, and stored before DNA extraction. DNA was extracted from a membrane filter (cellulose ester, Millipore, Billerica, MA, United States) using a FastDNA Spin Kit (MP Biomedicals, Solon, OH, United States). The quality and quantity of the extracted DNA were assessed using agarose gel electrophoresis and a Nanodrop spectrophotometer (Thermo Scientific, Wilmington, DE, USA), respectively. For metagenome library preparation and sequencing, Illumina TruSeq DNA library preparation kit v2 (Illumina, San Diego, CA, USA) was utilized following the manufacturer's instructions. The quantity of each library was determined using a Qubit fluorimeter (Invitrogen, Carlsbad, CA, USA). All libraries were sequenced at Novogene Inc. (Beijing, China) using the Illumina HiSeq 2500 sequencing system.

We collected metagenomic data from other plastic-contaminated soil from four publicly available projects. Two samples from India SRX2861368 (n = 1) and PRJNA657696 (n = 10), one from Panama mgm4794685.3–mgm4794690.3 (n = 6), and one from the USA SRP107015 (n = 48).

Metagenome assembly, binning, genes prediction, and construction of gene and protein catalog. After Quality control of reads by FastQC, low-quality bases (Phred score < 20) and residual Illumina adapter contaminations were trimmed and filtered by Trimmomatic software, and reads were removed by SOAPaligner2⁵¹. We used MEGAHIT⁵² software to assemble high-quality short reads. Options used in MEGAHIT were `-kmin 1pass, -m 60e + 10, -k-min 27 -k-max 127 -k-step 10 -min-contig-len 300, -t 40`. After assembly, the resulting contigs were mapped to the primitive reads using BWA¹⁶ to determine contig coverage profiles, and Samtools⁵³ was used to convert to BAM format. Finally, we performed gene identification of the contigs from each sample by MetaGenMark⁵⁴.

We utilized CD-HIT⁵⁵ to construct a non-redundant gene catalog. The CD-HIT parameters were $-c\ 0.9 -M\ 0 -T\ 0$ to cluster the genes with the indicator of overlap $\geq 90\%$. At this step of the workflow, individual gene catalogs were constructed. These GCs were merged, and CD-HIT was used to make them unique again. Finally, we applied KofamKOALA⁵⁶ to assign the non-redundant gene catalog to KEGG orthology⁵⁷. eggNOG-mapper⁵⁸ was performed to allocate clusters of orthologous groups (COG) functional categories. Also, the standalone run_dbCAN2 was used to find the presence of CAZyme genes in the final GC. To identify protein-coding regions in metagenome sequences in all contigs and each bin, we used MetaGenMark software.

Taxonomic profiling and construction of taxa catalog. MetaPhlAn3 was used to perform taxonomic profiling of the raw data prior to assembly and contig creation. MetaPhlAn is a computational tool that assigns taxonomy to microbial communities from metagenomic shotgun sequencing. Raw reads were utilized to ascertain the abundance and diversity of the microbiome present in the sample, with respect to both bacteria and archaea. In addition, after assembly and generation of the contigs, genome bins were reconstructed using MetaBAT2 software with options $-\text{minContigLength}\ 2000, -\text{minContigDepth}\ 2$. The bins were merged and reduced to replicate dRep software. We first used the CheckM program⁵⁹ to check genome bins for completeness, contamination, and strain heterogeneity. Bins with $\geq 65\%$ completeness and $\leq 10\%$ contamination were retained. We applied GTDB-tk to map the taxonomy to the MAGs. Finally, we used the standalone run_dbCAN2 to determine the presence of CAZyme genes in the MAGs catalog.

Construction of plastizymes catalog. We collected a new dataset consisting of 158 unique enzyme sequences with different forms of plastic degrading enzymes by tracking the NCBI's protein database, BRENDA⁶⁰ and UniProt⁶¹, reported by literature⁶². CD-HIT was utilized to reduce redundancy of highly-homologous clusters with a cut-off value of 0.9. The collected dataset contained various forms of plastizymes, including Polyethylene (PE), Polyethylene Terephthalate (PET), Caprolactam, Nylon, and others, as illustrated in Fig. 1.

Within the scope of this investigation, the MeTarEnz⁶³ (metagenomic targeted enzyme miner) tool was utilized to identify putative plastic degrading genes from contigs. MeTarEnz is a multi-functional software that enables targeted screening of high-throughput metagenomic data with user-defined databases and bit-score cut-offs. A plastizyme database (158 enzyme sequences) was screened, with a minimum bit score of 200. Subsequently, the final metagenomic sequences were further analyzed using NCBI CDD⁶⁴, AlphaFold 2⁶⁵, and TMalign⁶⁶ tools to predict and compare the 3D structure of the predicted plastizymes.

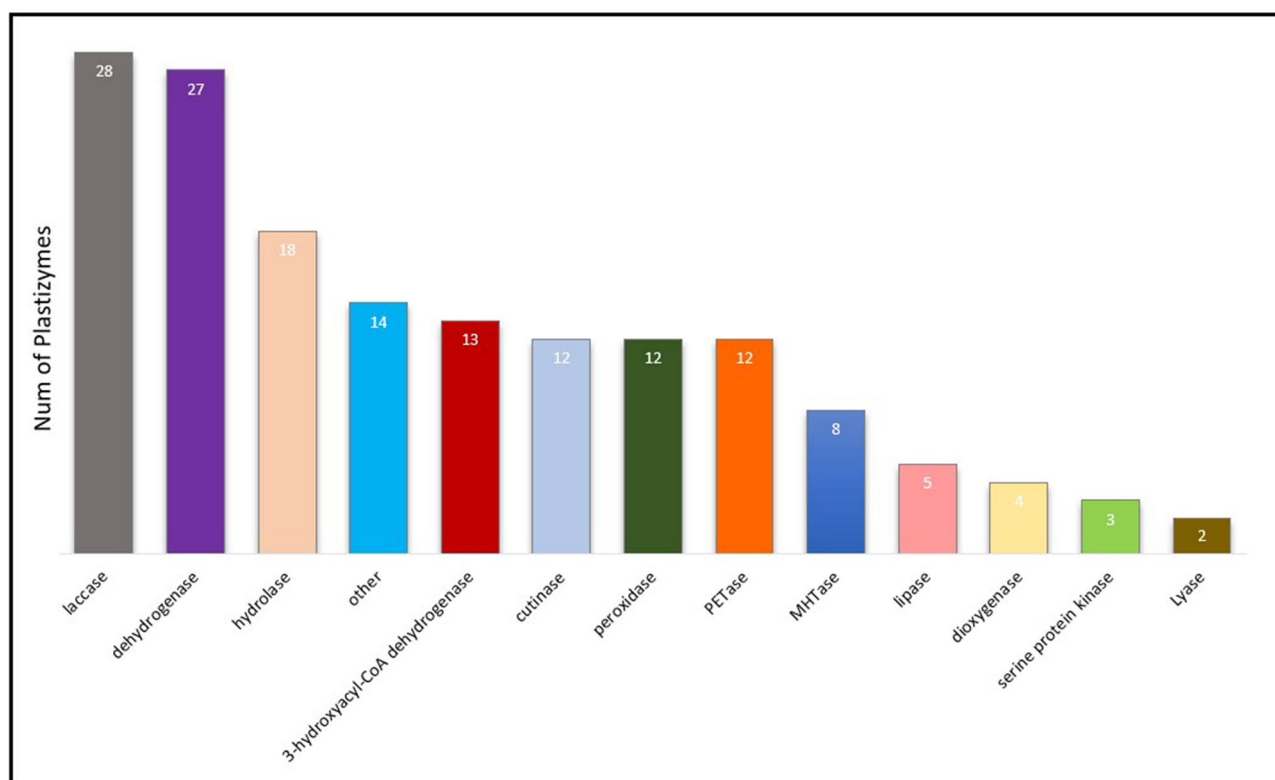


Figure 1. Number of plastizymes in the dataset.

Results and discussion

Construction of integrated PCEG and PCEP. To the best of our knowledge, the present study is the first to develop a comprehensive catalog of plastic-contaminated soils using an integrated bioinformatic workflow. Five datasets were used to produce 53,300,583 unique genes and proteins. Quality control of reads, trimming, assembly, and binning of contigs were all components of the workflow. The results of this study suggest that this workflow can be used to generate a catalog of plastic-contaminated soils and provide a basis for further investigation into the composition and function of these samples. (Fig. 2).

The proposed workflow was successfully applied to a newly plastic-contaminated soil sample combined with four previously published whole metagenome datasets. The different acquisition processes (i.e., taxonomical and functional profiling) contributed to more comprehensive results; a total of ~365 Gb of high-quality metagenome raw read data was obtained from the merged five datasets; this data size was greater than that of all plastic-contaminated samples. Raw reads of 66 plastic-contaminated metagenomic samples from five different locations (one sample from Varamin-Iran, one sample from Ghazipur-India, ten samples from Gujarat-India, six samples from PFHT-Panama, and 48 samples from WWTP-USA) were assembled using MEGAHIT⁶⁷ software⁶⁸. The assembly of short reads resulted in the identification of 44,115,301 contigs > 300 bp in length. The maximum contigs length was 954,897.

Table 1 shows the general features of each dataset. The results indicated that 53.3 million non-redundant genes and proteins were generated from the categorized complete genes. The first integrated plastic-contaminated environment gene catalog (PCEG) and integrated plastic-contaminated environment protein catalog (PCEP) were constructed. KEGG orthologs were mapped to proteins using the KofamKOALA⁵⁶ resulting in 3,370,057 genes. Additionally, the results of the evolutionary genealogy of genes in non-supervised orthologous groups (eggNOG)⁵⁸ mapper showed that 27,718,862 genes were annotated with COG functional categories in PCEG.

These two approaches were used separately for the five individual samples and PCEG. Figure 3a displays various genes annotated by eggNOG and KEGG for all five samples, whereas Fig. 3b illustrates the percentage of genes mapped by eggNOG and KEGG in PCEG.

The results of the present study showed that the number of genes mapped to eggNOG was higher than those mapped to KEGG in all samples. This finding was in line with those of other studies, such as the construction of a gene catalog of the chicken gut microbiome⁴². The highest frequency of unmapped genes (not mapped to any mapper database) was observed in the Panama sample (48%), followed by that in the Ghazipur sample (45%). The lowest frequency was observed in the Gujarat sample (approximately 30%).

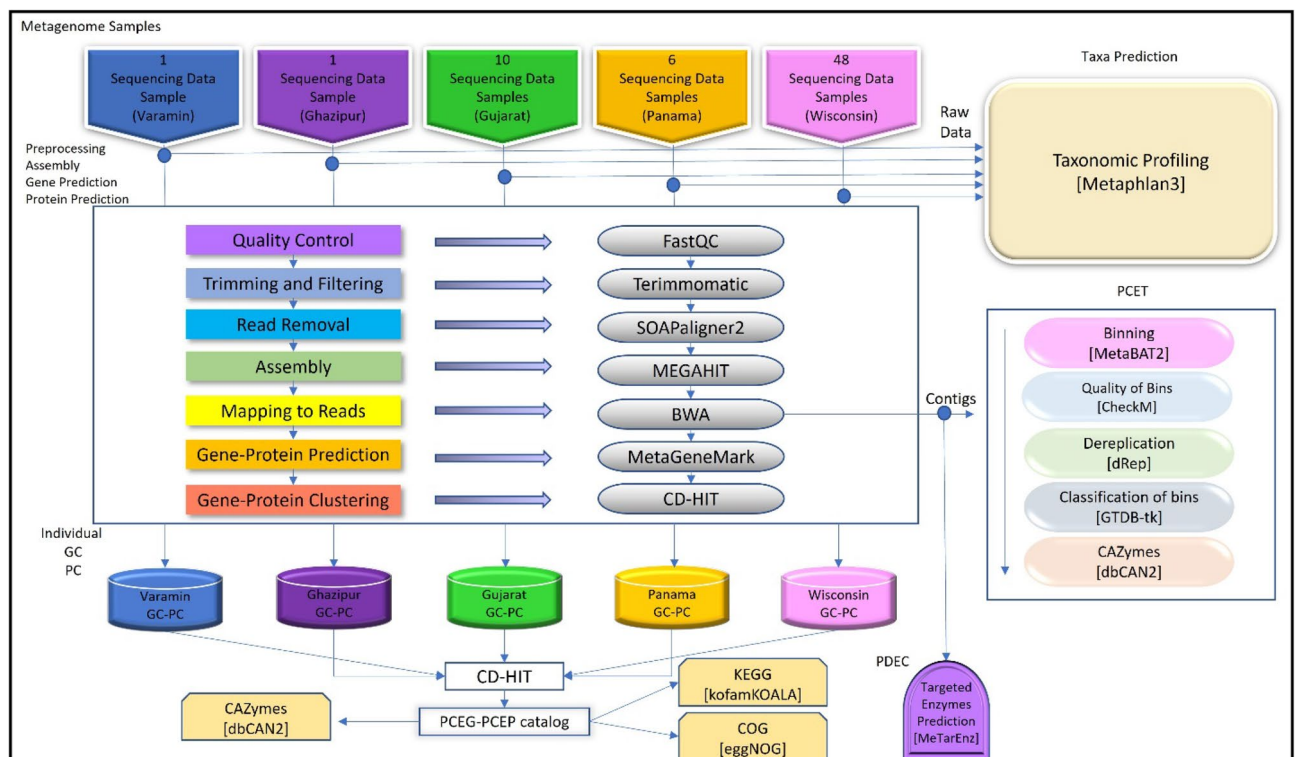


Figure 2. Bioinformatic workflow and tools used in data processing, analyzing, and integrating the metagenome samples. The workflow successfully processed 66 metagenome samples from five locations to generate their respective gene catalogs, which were merged into PCEG and PCEP. Taxonomic analysis with Metaphlan3 and GTDB-tk identified sequenced prokaryotic genomes or draft genomes of metagenome samples and constructed a taxa catalog. In addition, MeTarEnz was used to generate a plastzyme catalog from the contigs. The results from this workflow demonstrated the successful generation of gene and protein catalogs, taxa catalogs, and plastzyme catalogs from 66 metagenome samples from five locations.

Data set	Num of sample	Non redundant genes	Binned contigs	Unbinned contigs	Total length (bp)	Min length (bp)	Max length (bp)	Average length (bp)	References
Ghazipur	1	1,607,885	5,195	1,116,014	605,689,029	300	88,753	527	46
Gujarat	10	3,575,629	37,047	2,431,599	1,750,863,557	300	280,338	709	47
Panama	6	5,284,867	176,451	6,192,732	6,062,623,939	300	954,897	951	48
Wisconsin	48	41,544,363	770,174	29,972,297	16,428,178,966	300	184,708	546	50
Varamin	1	1,305,332	399,494	3,014,298	3,516,817,391	300	600,358	1411	In this study

Table 1. General feature of the individual gene catalogs.

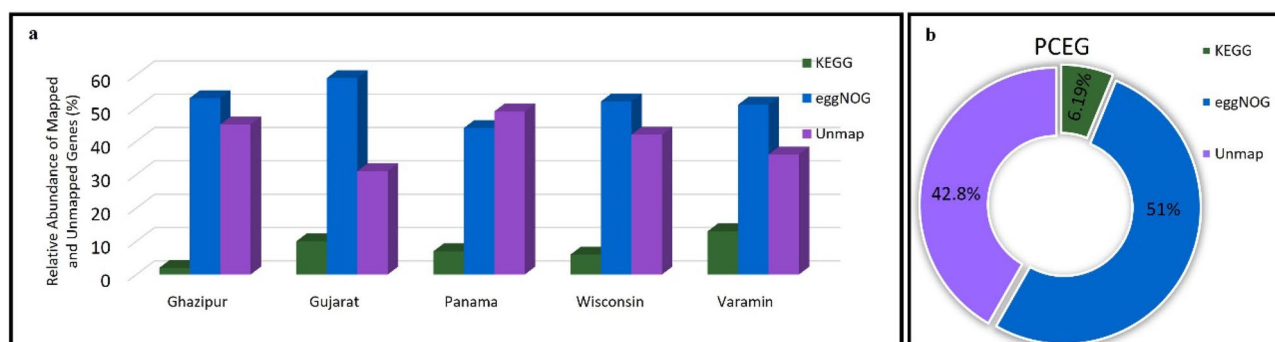


Figure 3. Metagenome-assembled 53,300,583 nonredundant genes from 66 plastic-contaminated soil metagenome samples. **(a)** Number of genes mapped to eggNOG⁵⁸ and KEGG^{56,57} in the individual gene catalog and PCEG. PCEG is an integrated gene catalog produced. **(b)** Distribution of the mapped gene by kofamKOALA (KEGG), eggNOG mapper, and unmapped genes in PCEG.

Furthermore, the Carbohydrate-Active Enzyme (CAZy) database is a powerful source of enzyme information. The findings of this research have the potential to advance the research in the field of plastic-contaminated gene catalogs⁶⁹. To the best of our knowledge, the scattering of different groups of plastizymes in CAZyme has not been extensively investigated. Plastizymes play an important role in the hydrolysis and decomposition of plastics; therefore, understanding the abundance of these enzymes in contaminated environments is essential. Therefore, we analyzed PCEG with dbCAN2⁷⁰ for CAZyme profiling in a sample from plastic-contaminated environment. dbCAN2 resulted in 205,066 CAZyme-encoding genes in PCEG, belonging to 51 CAZyme subclasses. The glycoside hydrolase (GH) class was the most abundant in the plastic-contaminated environment, followed by glycosyltransferase (GT) and carbohydrate-binding module (CBM). Notably, the AA, CBM, and GH groups with the AA1, AA3, and GH1 subclasses are known to be highly related to plastizymes.

Our results indicate that CAZymes were widely distributed among the five individual samples (Fig. 4). Of the 51 subclasses identified, GH1 and CBM were the most abundant among all samples (Fig. 4a). GH1 was particularly abundant in all samples except for Gujarat. AA3 was found to be the most distributed in the Varamin sample compared with the other samples.

Overall, GH1 and GT2 were the most abundant CAZymes among the different samples, as observed at the PCEG level (Fig. 4b).

The relative abundance of the six CAZyme categories (Auxiliary Activities (AA), Carbohydrate-Binding Module (CBM), Carbohydrate Esterase (CE), Glycoside Hydrolase (GH), Glycosyl Transferase (GT), Polysaccharide Lyase (PL)) among the plastic-contaminated samples is shown in Fig. 4c. CBM was the most abundant class, followed by GH and AA. These results suggest that plastic contamination affects the abundance of CAZymes in metagenome samples.

These three categories were highly abundant among the plastizymes. (In Fig. 4, categories and subclasses that include most plastizymes are illustrated by blue asterisks). These results indicate the differential properties of the plastic-contaminated environment metagenome in terms of plastic degradation.

Construction of integrated PCET. According to the results, microbial diversity in contaminated plastic samples is diverse, with the highest abundance of plastic degrading bacterial phylum belonging to *Actinobacteria*, *Firmicutes*, *Planctomycetes*, *Bacteroidetes* and *Pseudomonadota* categories. In addition, *Ideonella sakaiensis*, *Thermobifida fusca*, *Pseudomonas soli*, *Pseudomonas jessenii*, *Paenibacillus*, *Fusarium redolens*, *Fusarium spp*, *Penicillium*, and so on^{71,72} have been reported as the most popular plastic-degrading microorganisms. The PCET was investigated using two approaches. In the first approach, after quality control of reads, MetaPhlan3⁷³ was used for taxonomic profiling of the samples, and in the second approach, samples were analyzed by GTDB-tk⁷⁴. According to the MetaPhlan3 assignments, the most dominant phyla in the integrated catalog were *Proteobacteria*, *Actinobacteria*, *Firmicutes*, *Planctomycetes*, *Bacteroidetes*, and *Chloroflexi*. In total, 11 classes, 23 orders, 45 families, 42 genera, and 253 non-redundant species were predicted using the MetaPhlan3.

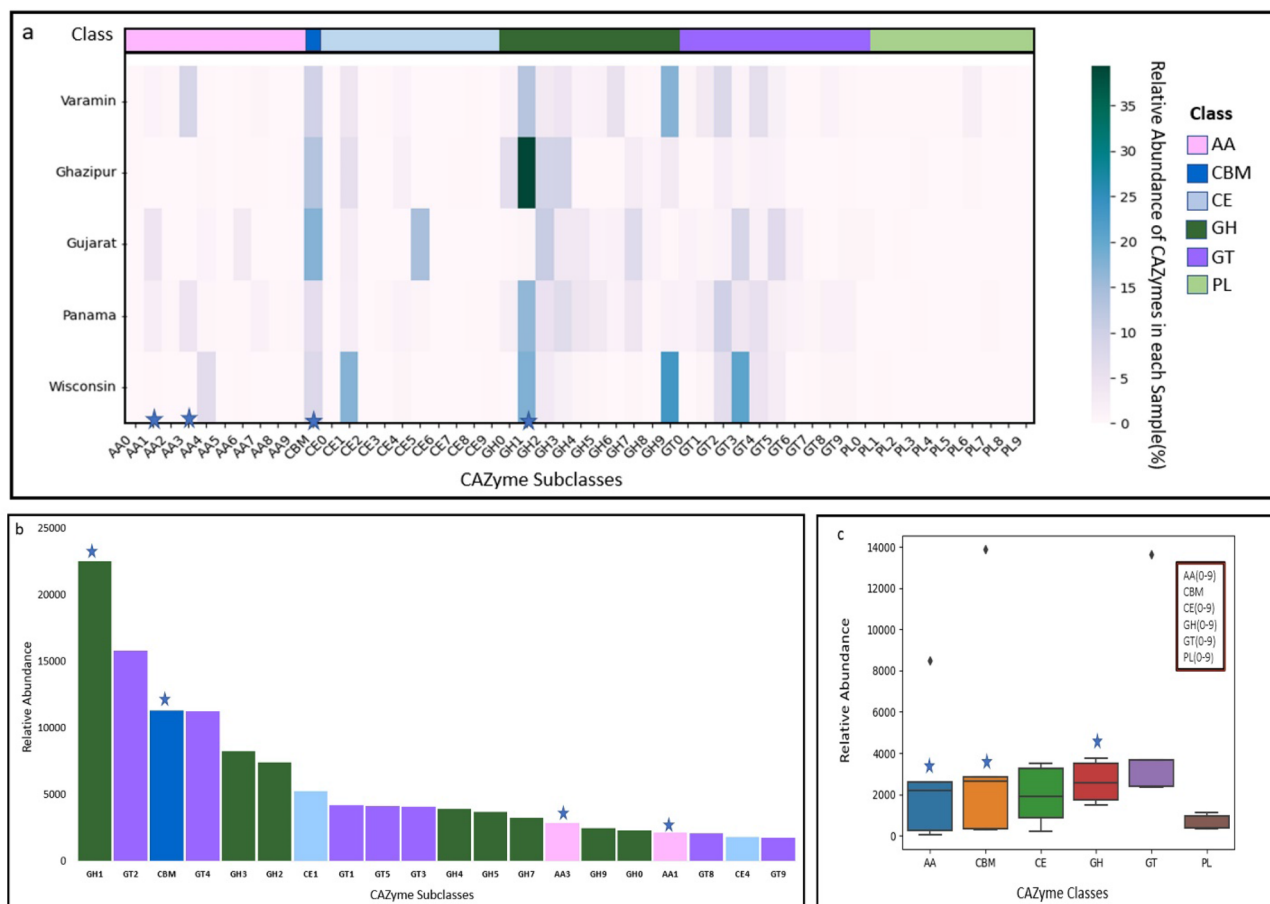


Figure 4. Abundance of CAZymes in plastic-contaminated soil metagenome samples. (a) Variations in abundance of all 51 CAZyme subclasses among the five samples. Columns in the heat map represent different CAZyme subclasses, and the color of each group is scaled from white-blue to green according to the relative abundance within the samples. (b) The 20 CAZyme subclasses with the most significant frequency found in PCEG. (c) Boxplot of relative abundance of the six CAZyme categories observed in the samples. Note: The CAZyme categories and subclasses that contained the most plastizymes are marked with a blue asterisk.

Figure 5a shows the relative abundance of the bacterial phyla in PCET, as determined by MetaPhlan3. *Proteobacteria* and *Chloroflexi* were the most and least abundant phyla, respectively. *Actinobacteria*, *Firmicutes*, *Planctomycetes*, and *Bacteroidetes* were the most abundant phyla, and these groups had the highest distribution of plastic-degrading bacteria. Figure 5b shows the distribution of the top 15 species level taxonomies with the most repetition among our plastic-contaminated soil taxa catalog. Three plastic-degrading species, *Thermobifida fusca*, *Paenibacillus* sp. AR247 and *Pseudomonas soli*, were identified in the top 15 species from taxonomical profiling by MetaPhlan3.

At this stage, several bacterial species, such as *Pseudomonas soli* and *Pseudomonas jessenii*, had the highest average integrity among all samples. (Fig. 5c). These species have been reported as the leading plastic-degrading bacteria. The results of taxonomic profiling suggested that *P. soli* was the most abundant plastic-degrading bacterium in all samples. Compared to other species, *Ideonella sakaiensis* had the lowest frequency. *Thermobifida fusca* and *Ideonella sakaiensis* are popular PET-degrading bacterial species. Both species had the highest frequencies in Varamin, Gujarat, and Ghazipur. The critical point is that these three environments were in direct contact with PET residues, so we can conclude that this is why these two bacterial species were more abundant compared to the other two environments. *Brucella intermedia* and *Novosphingobium*, which encode phenanthrene and caprolactam degradation enzymes, were highly abundant in the Wisconsin and Panama metagenomic samples.

In an additional analysis using the second approach, we used MetaBAT2⁷⁵ to identify 691 individual genome bins. Among them, 97 bins were associated with a completeness score of >65% and a contamination score of <10%, and individual genomes, including bacteria and archaea, remained. After the dereplication of bins resulting from MetaBAT2, the number of bins generated from Ghazipur, Gujarat, Panama, Wisconsin, and Varamin was 5, 62, 122, 127, and 75, respectively. All binned samples were analyzed using GTDB-tk⁷⁴. According to the GTDB-tk assignments, in PCET, 844 MAGs (691 non-redundant) were assigned to the 16S rRNA gene sequence. Figure 5d illustrates the distribution of all 37 phylum-level taxonomies in the prepared plastic-contaminated soil taxa catalog. *Proteobacteria* and *Calditrichota* had the highest and lowest frequencies in all five samples, respectively. *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, *Gemmatimonadota*, and *Planctomycetes* were the phyla most related to plastic digestion bacteria, with the highest amount found in Varamin and Panama (*Actinobacteria*), Varamin

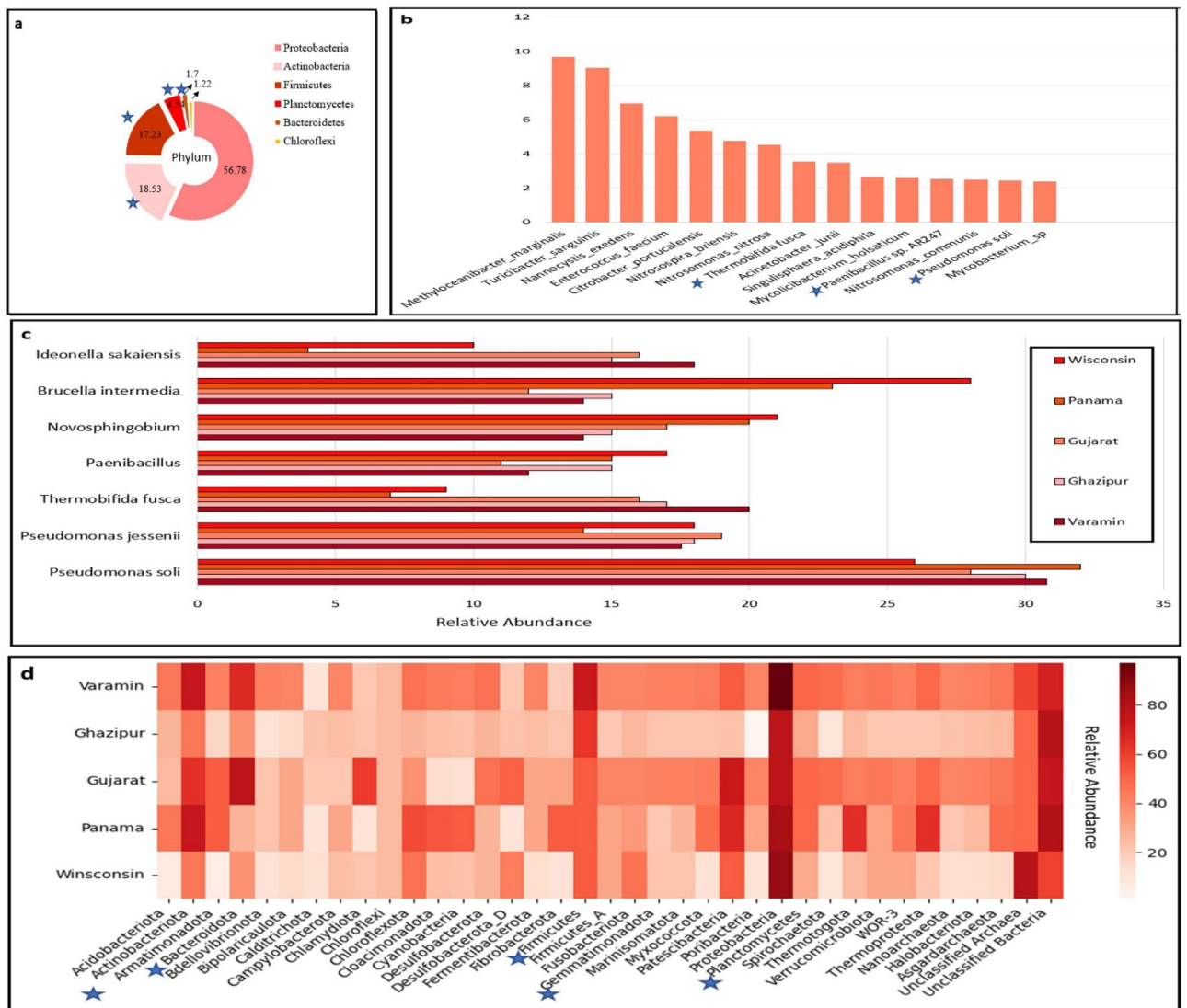


Figure 5. Phylum and species level taxonomic distribution. (a) The donut chart shows the abundance of phylum-level bacteria in PCET (MetaPhlAn3). (b) The bar chart shows the distribution of the top 15 species taxonomic levels among PCET (MetaPhlAn3). (c) The bar charts illustrate the distribution of the top seven plastic-degrading bacterial species in all five samples (GTDB-tk and NCBI-blast). (d) Distribution of all 37 phylum-level taxonomic classes among PCET predicted by GTDB-tk. Note: The phylum classes and species that play the most important role in plastic degradation are marked with blue asterisks.

and Gujarat (*Bacteroidota*), Varamin (*Firmicutes*), Varamin and Gujarat (*Gemmatimonadota*), and Varamin, Gujarat, and Panama (*Planctomycetes*).

The most abundant bacterial class, Order, Family, Genus, and Species-level taxonomic profiling were *Bacteroidia* (21.6%), *Steroidobacteriales* (20.11%), *Pseudomonadaceae* (12.22%), *Fermentibacter* (9.6%) and *Methyloceanibacter_marginalis* (9.7%). Additionally, 82 novel bacterial genomes were identified, belonging to novel species, respectively.

GTDB-tk analysis revealed 19 archaeal genomes, of which nine belonged to the phylum *Thermoproteota* (47%). *Asgardarchaeota*, *Halobacteriota*, and *Nanoarchaeota* were at the phylum level. Furthermore, 18 archaeal genomes were novel species. Figure 6a, b illustrate the relative abundance of the 410 most abundant OTUs bacterial and archaeal genomes in PCET.

To date, three species of sac fungi, *chrysosporium*, *Aspergillus niger*, and *fusarium solani* have been reported to be the leading plastic-degrading microorganisms. The vital point of our results is that a significant number of these microorganisms were identified in four samples Varamin (154), Gujarat (23), Panama (44), and Winsconsin (227).

The phylogenetic tree of 194 MAGs containing plastic-degrading bacteria is shown in Fig. 6c. The main species of identified plastic-degrading bacteria belong to *Pseudomonadota* and *Actinobacteria*. *Brucella intermedia*, *Novosphingobium*, *Pseudomonas soli*, *Ideonella sakaiensis*, and so on belong to members of *Pseudomonadota*. *Streptomyces clavuligerus*, *Paenarthrobacter ureafaciens*, etc., belong to members of *Actinobacteria*. (Fig. 6c; Table 2).

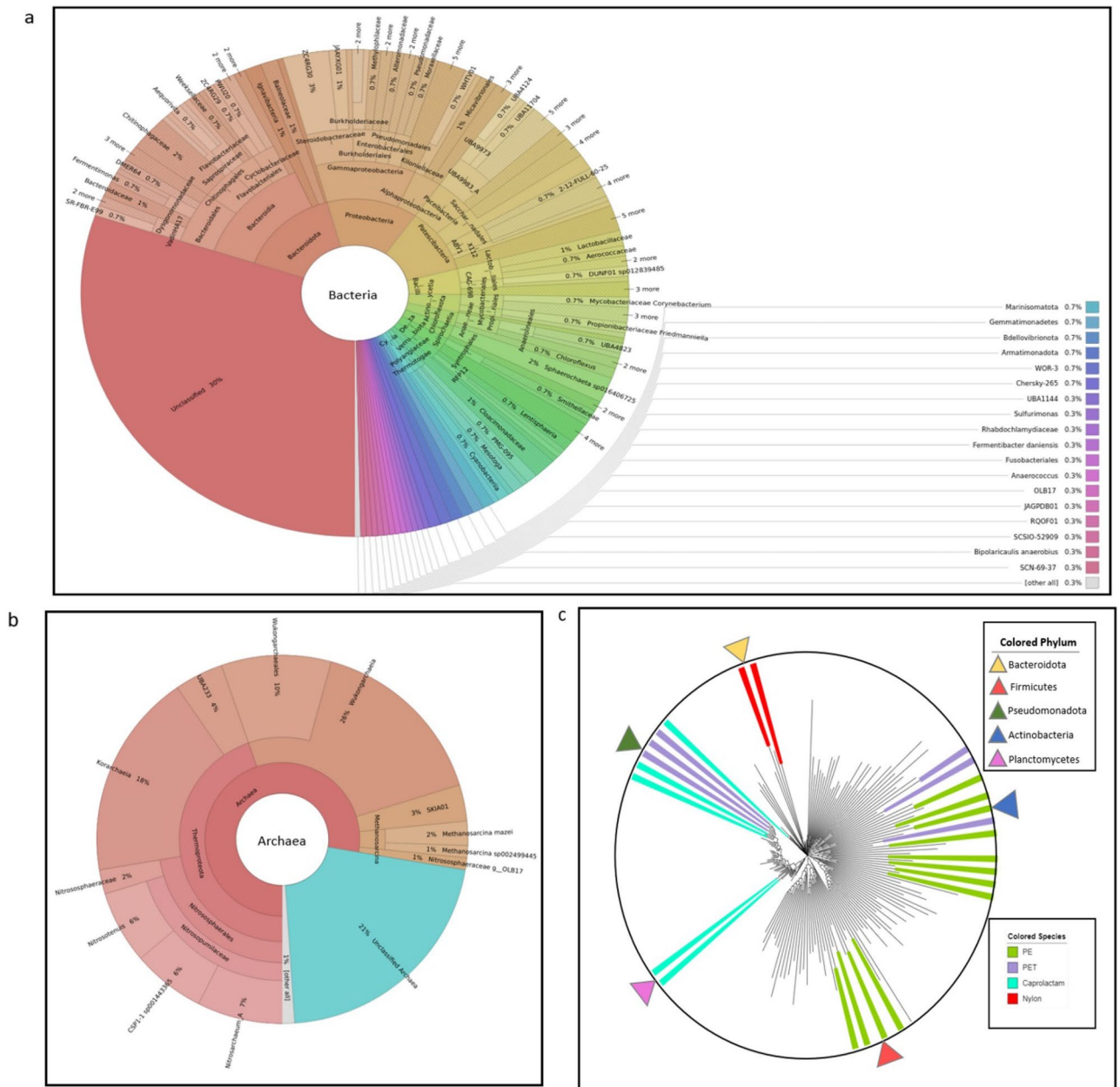


Figure 6. Diagrams showing the relative abundance diversity and phylogenetic tree of full-level taxonomic profiling of the 410 most abundant OTUs for the PCET. **(a)** Distribution of bacterial genomes in PCET. **(b)** Distribution of archaeal genomes in PCET. **(c)** Phylogenetic tree of 194 MAGs containing plastic-degrading bacteria. Colors inside the circle correspond to species with a role in caprolactam, nylon, PET, and PE-degrading bacteria. Colored triangles show the phyla of species that have the ability to digest plastic. Note: The diversity of taxa and taxonomic phylogenetic trees were visualized using the Krona chart-2.8.1⁷⁶ (<https://github.com/marbl/Krona/wiki>) and Interactive Tree Of Life (iTOL) v5⁷⁷ (<https://itol.embl.de/>), respectively.

In this study, we identified a remarkable number of bacterial species that were experimentally introduced as plastic-degrading bacteria in contaminated plastic metagenome samples collected by PCET. These findings suggest that these bacteria are capable of degrading plastic in the environment. Further research is needed to understand the role of these bacteria in plastics degradation.

Construction of integrated PDEC. Construction of the plastizyme catalog revealed a large number of enzymes with the highest similarity to the plastizymes database in both sequence and 3D structural dimensions. To identify possible novel plastizymes, after mapping the contigs against raw reads using BWA, the resulting contigs were screened using MeTarEnz⁶³ against the plastizymes’ dataset collected in the investigation with the maximum bit score and minimum E-value. Among the analyzed data, 136,654 plastic-degrading enzyme sequences were predicted. Furthermore, to validate the results of the identified plastizymes, AlphaFold2⁶⁵ was

Name	Phylum	Type of encoded plastizymes	Type of plastic
<i>Brucella intermedia</i>	<i>Pseudomonadota</i>	Dehydrogenase	Caprolactam
<i>Novosphingobium sp</i>	<i>Pseudomonadota</i>	Phenanthrene dioxygenase	Phenanthrene
<i>Pseudomonas soli</i>	<i>Pseudomonadota</i>	3-Hydroxyacyl-CoA dehydrogenase	Caprolactam
<i>Ideonella sakaiensis</i>	<i>Pseudomonadota</i>	PETase	PET
<i>Acidovorax carolinensis</i>	<i>Pseudomonadota</i>	Laccase	caprolactam
<i>Azoarcus</i>	<i>Pseudomonadota</i>	Laccase	Phthalate-PET
<i>Comamonas</i>	<i>Pseudomonadota</i>	Laccase	Phthalate-PET
<i>Massilia</i>	<i>Pseudomonadota</i>	Laccase	Phenanthrene-PET
<i>Paraburkholderia caribensis</i>	<i>Pseudomonadota</i>	Laccase	Phenanthrene-PET
<i>Streptomyces clavuligerus</i>	<i>Actinobacteria</i>	Hydroxylase	PE
<i>Paenarthrobacter ureafaciens</i>	<i>Actinobacteria</i>	6-Aminohexanoate-dimer hydrolase	PE-nylon
<i>Thermobifida alba</i>	<i>Actinobacteria</i>	Cutinase	PET
<i>Thermobifida fusca</i>	<i>Actinobacteria</i>	Cutinase	PET
<i>Streptomyces sp. SM14</i>	<i>Actinobacteria</i>	Alpha/beta hydrolase	PET-MHETase
<i>Thermomonospora curvata</i>	<i>Actinobacteria</i>	Lipase	PET-MHETase
<i>Thermobifida halotolerans</i>	<i>Actinobacteria</i>	Cutinase	PET
<i>Gordonia phthalatica</i>	<i>Actinobacteria</i>	Laccase	PE-phthalate
<i>Micrococcus luteus</i>	<i>Actinobacteria</i>	Hydrolase	PE
<i>Rhodococcus erythropolis</i>	<i>Actinobacteria</i>	Hydrolase	PE
<i>Nocardia asteroides</i>	<i>Actinobacteria</i>	Hydrolase	PE
<i>Mycobacterium sp</i>	<i>Actinobacteria</i>	Laccase	PE-phenanthrene
<i>Flavobacterium</i>	<i>Bacteroidota</i>	6-Aminohexanoate-cyclic-dimer hydrolase	Nylon
<i>Cytophagales</i>	<i>Bacteroidota</i>	Hydrolase	Nylon
<i>Bacillus subtilis</i>	<i>Firmicutes</i>	Para-nitrobenzylesterase	PE
<i>Sphaericus</i>	<i>Firmicutes</i>	Hydrolase	PE
<i>Lactobacillales</i>	<i>Firmicutes</i>	Esterase	PE
<i>Brevibacillus borstelensis</i>	<i>Firmicutes</i>	Hydrolase	PE
<i>rubripirellula</i>	<i>Planctomycetes</i>	Peroxidase	Caprolactam
<i>Pirellulase</i>	<i>Planctomycetes</i>	Peroxidase	Caprolactam

Table 2. The table shows the identified species associated with plastic degradation based on detailed reports in the phylogenetic tree shown in Fig. 6 (Section c).

used to predict the 3D structure and TM-align for 3D structure comparison of some of the obtained plastizymes (approximately 700 sequences). The AlphaFold predicted local distance difference test scores (pLDDT) were greater than 80% for all predicted enzymes, and the TM scores were between 70 and 98%. These scores indicated the high affinity of the identified enzymes for plastic degradation enzymes.

Putative enzymes belong to 12 different plastizyme families, including peroxidases, PETases, cutinases, and laccases. Table 3 shows details of the main plastizymes. After applying MeTarEnz and segregating all identified

Name of enzymes	EC No	Superfamily name	Top hit PDB code
3-Hydroxyacyl-CoA dehydrogenase	1.1.1.211	PRK08268	c3mogA_
Lyase	4.2.-	Lyase_I_like	d1gkma_
Serine protein kinase	2.7.11.1	PRK15455	c6iy8B_
Dehydrogenase	1.1.1.1	ALDH-SF	c4f9iA_
Hydrolase	3.1.1.-	Abhydrolase	c3visB_
Cutinase	3.1.1.74	Abhydrolase	d1cexa_
Laccase	1.10.3.2	Cu-oxidase_4	c5daoA_
Dioxygenase	1.14.-	Rieske	c2gbxE_
Peroxidase	1.11.1.-	Thioredoxin_like	c2v2gC_
Lipase	3.1.1.3	Abhydrolase	c3visB_
PETase	3.1.1.101	Abhydrolase	c5xg0A_
MHTase	3.1.1.102	Abhydrolase	c6qgbA_

Table 3. List of main enzymes involved in biodegradation of all types of plastics, EC number, superfamily name, and top-hit 3D structure PDB codes identified in within the scope of this research.

plastizymes from PDEC with the highest similarity to database sequences, NCBI-CDD was used to compare the conserved domains, superfamilies, and top-hit PDB structures.

The results indicate that the most abundant plastizymes belonged to Wisconsin samples, followed by Panama, Varamin, Gujarat, and Ghazipur samples. These plastizymes have been found to degrade caprolactam, nylon, PET, and PE. Other plastizymes include polyvinyl chloride (PVC), polyurethane (PUR), polyamide (PA), phenanthrene, and phthalates (Fig. 7a). The bit score and e-value of the selected putative plastizymes were also determined (Fig. 7b), with the highest bit score obtained for PETase (730), followed by cutinase (680) and dehydrogenase (620). These results suggest that PDEC are an excellent reference for PET and PE-degrading enzymes.

Focusing on the enzyme characteristics of the metagenomic sequences in plastic-contaminated environment, we first revealed the diverse distribution of plastizyme abundance (Fig. 7c) PETase had the highest abundance (average 2800) compared to other plastizymes, followed by hydrolases and peroxidases. In addition, we found that MHEase had the smallest redundancy (average of 500) among all five samples. Figure 7d illustrates the dispersion of the top six plastizymes in each sample. Hydrolase (nylon-PE), peroxidase, and cutinase enzymes were the most prominent plastizymes in the Varamin sample. In addition, hydrolase, peroxidase, and PETase



Figure 7. Profiling of predicted plastizymes in the PDEC. (a) Abundance of the predicted plastizymes in each sample. (Note that for each metagenome sample, the percentage of plastizymes was calculated for all samples.) (b) Reported bit score and e-value for putative plastizymes. (c) The violin plot shows the explainability of the samples in each province to interpret the first seven plastizymes. (d) Heatmap of the distribution of six plastizymes with high abundance in Varamin, Ghazipur, Gujarat, Panama, and Wisconsin samples separately. (e) Six plastizymes with the highest abundance in integrated PDEC. (f) Venn diagram analysis of five metagenome samples with different identified plastizymes groups.

were among the most abundant plastizymes in Gujarat and Ghazipur samples. In contrast to the Panama and Wisconsin samples, hydrolase was the main plastizyme with the highest quantity. Laccase and dehydrogenase were present in the lowest amounts in all samples. These results indicate that Varamin, Gujarat, and Ghazipur were more directly related to plastic particles than Wisconsin and Panama.

Also, Fig. 7e shows the six plastizymes with the highest abundance found in PDEC. The most abundant plastizymes were Proxidase, Dehydrogenase, Cutinase, PETase, PETase, Hydrolase, and Laccase. These plastizymes play important roles in Caprolactam, Nylon, PET, and PE.

Many plastizymes have rarely been described in metagenomic environments, because they have been identified in wet laboratories using traditional cultivation methods. In PDEC, we found all underrepresented plastizymes in the collected data. Strikingly, plastic-contaminated samples had 96 hydrolase plastizymes, 38 of which belonged to cutinase and PETase (PET-degrading enzymes), 15 to lipase (PE- and caprolactam-degrading), and the remaining enzymes belonged to nylon-, phenanthrene-, and phthalate-degrading enzymes. The second most crucial class of plastizymes is oxidoreductase. In PDEC, 27 different oxidoreductase plastizymes belong to laccases and dehydrogenases capable of PE, phenanthrene, and so on. In Fig. 7f the Venn diagram analysis revealed the presence of 63 plastizymes in all samples, as well as unique plastizymes in each sample, with the highest number of unique plastizymes found in Varamin (83), followed by Ghazipur (58), Gujarat (84), Panama (13), and Wisconsin (46). These results suggest that any plastic-contaminated site is a rich source of plastizymes and further research is needed to better understand their potential for plastic degradation.

The plastic-contaminated environments act as evolutionary repositories of plastic-degrading enzymes, as evidenced by the high diversity of plastizymes in the plastic-contaminated metagenome samples (Fig. 8). The depicted trees graphically elucidate the evolutionary connections among key plastizymes, encompassing the degradation of PE, caprolactam, nylon, PET, and phenanthrene, all within the framework of the comprehensive plastic-contaminated environment catalog (PDEC). In our analysis of metagenomic data, we observed many plastizymes previously identified experimentally. This could be attributed to the integrated different plastic-polluted

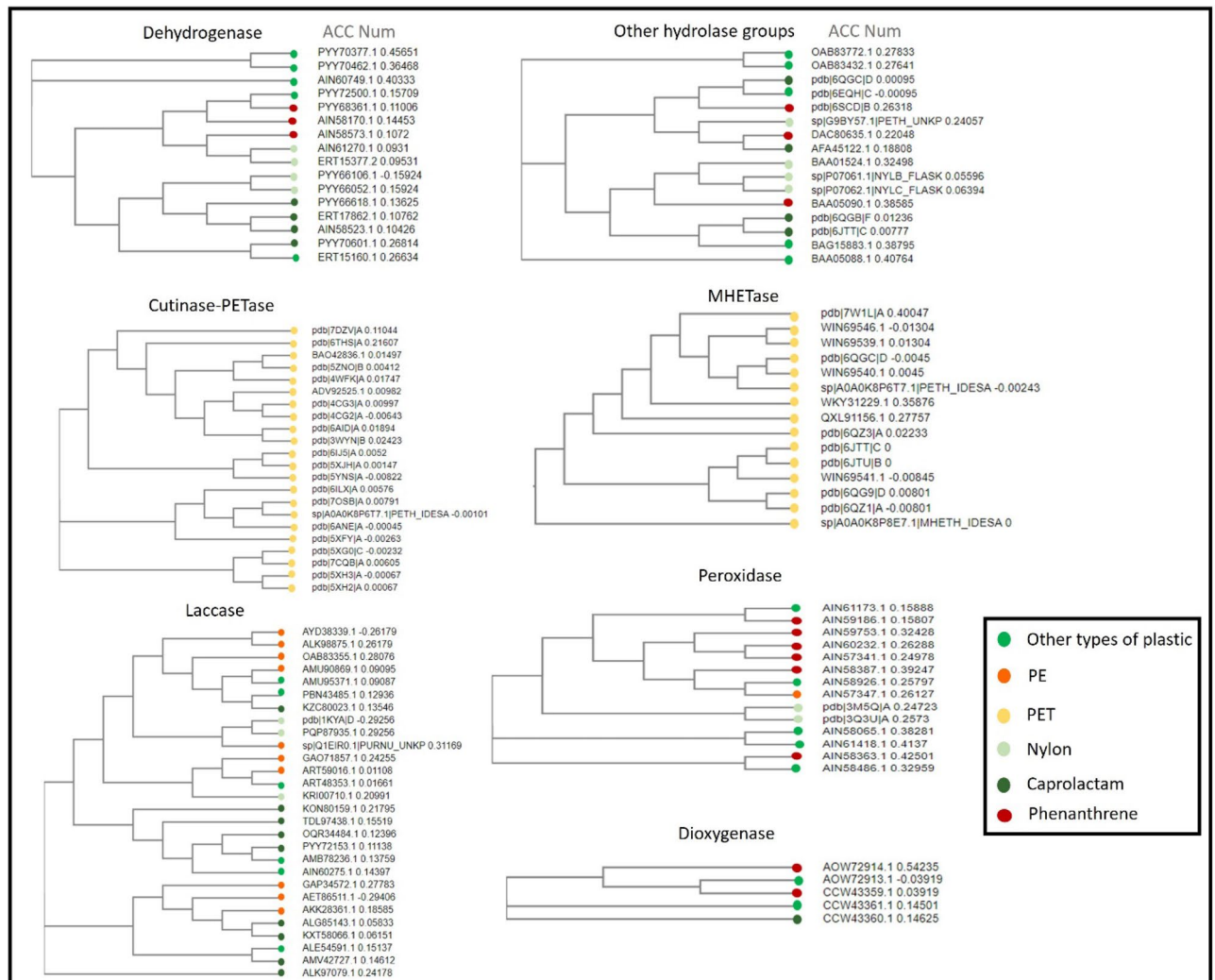


Figure 8. Seven phylogenetic trees of main Families of plastizymes that identified in PDEC. The hue of each node denotes the variety of plastic that the enzyme decomposes. They are downloaded from the NCBI RefSeq genomes.

environments exposure to a diverse array of plastics and chemicals. Further research on the underrepresented plastizymes in these samples could greatly expand our understanding of plastic contamination. Most of the enzymes identified were not previously identified by experimental approaches, suggesting that computational methods can be used to identify novel plastizymes. These findings have important implications for the field of metagenomics as they suggest that computational methods can be used to identify novel plastizymes and improve the accuracy of plastizyme identification.

Conclusions

The current plastic recycling process is inefficient and has detrimental effects on marine and terrestrial environments, as well as living organisms. To facilitate future studies related to the discovery of new plastizymes that can play an important role in the purification of plastic pollution, we developed a workflow and the first integrated catalog of plastic-contaminated samples to facilitate the exploration of genes, proteins, taxa, and enzymes associated with plastic degradation. We further conducted an in-depth analysis of the reconstructed genomes to identify novel plastizymes. Our findings suggest that plastic-contaminated environmental microbial communities have the potential to degrade plastic components and that these environments are full of microbial groups and plastizyme sequences that can be used to break down plastic. Further research in this area could revolutionize the plastic degradation industry by identifying novel plastic-degrading bacteria and plastizymes.

Statistical analysis. Statistical analyses and visualizations such as all heatmaps were conducted in Python 3.11.0 and PyCharm 2022.2.4 environment, utilizing the Matplotlib and Seaborn packages. Additionally, the diversity of taxa and taxonomic phylogenetic trees were visualized using the Krona chart-2.8.1⁷⁶ (<https://github.com/marbl/Krona/wiki>) and Interactive Tree Of Life (iTOL) v5⁷⁷ (<https://itol.embl.de/>), respectively. Further information can be obtained from the corresponding author upon request.

Data availability

The gene, taxa, and plastic degrading catalogs generated in the present study are available on <https://drive.google.com/drive/u/1/folders/1diw2DSZeuHFSM04jBG66YXMubcqds-Ul>. The obtained nucleotide sequences of agricultural land of Varamin were submitted to the NCBI Sequence Read Archive (SRA) under the accession number SRR23085642.

Received: 24 May 2023; Accepted: 18 September 2023

Published online: 25 September 2023

References

- Pathak, V. M. & Navneet. Review on the current status of polymer degradation: A microbial approach. *Bioresour. Bioprocess.* **4**, (2017).
- Visconti, A., Martin, T. C. & Falchi, M. YAMP: a containerized workflow enabling reproducibility in metagenomics research. 1–9 (2019) <https://doi.org/10.1093/gigascience/giy072>.
- Ariaeenejad, S., Kavousi, K. & Afshar, D. Enzymatically triggered delignification through a novel stable laccase: A mixed in-silico/ in-vitro exploration of a complex environmental microbiota. *Int. J. Biol. Macromol.* **211**, 328–341 (2022).
- Palomo, A. *et al.* Metagenomic analysis of rapid gravity sand filter microbial communities suggests novel physiology of *Nitrospira* spp. *ISME J.* **10**, 2569–2581 (2016).
- Bhaya, D. *et al.* Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME J.* **1**, 703–713 (2007).
- Li, Y. *et al.* The effects of straw incorporation with plastic film mulch on soil properties and bacterial community structure on the loess plateau. *Eur. J. Soil Sci.* <https://doi.org/10.1111/ejss.12912> (2019).
- Qi, Y. *et al.* Effects of plastic mulch film residues on wheat rhizosphere and soil properties. *J. Hazard. Mater.* **387**, 121711 (2020).
- Amobonye, A., Bhagwat, P., Singh, S. & Pillai, S. Plastic biodegradation: Frontline microbes and their enzymes. *Sci. Total Environ.* <https://doi.org/10.1016/j.scitotenv.2020.143536> (2020).
- Wei, R. & Zimmermann, W. Microbial enzymes for the recycling of recalcitrant petroleum-based plastics: How far are we?. *Microb. Biotechnol.* **10**, 1308–1322 (2017).
- Charnock, C. Norwegian soils and waters contain mesophilic, plastic-degrading bacteria. *Microorganisms* **9**, 1–18 (2021).
- Deguchi, T., Kitaoka, Y., Kakezawa, M. & Nishida, T. Purification and characterization of a nylon-degrading enzyme. **64**, 1366–1371 (1998).
- Ohki, T. *et al.* Two alternative modes for optimizing nylon-6 byproduct hydrolytic activity from a carboxylesterase with a β -lactamase fold: X-ray crystallographic analysis of directly evolved 6-aminohexanoate-dimer hydrolase. *Protein Sci.* **18**, 1662–1673 (2009).
- Sowmya, H. V., Ramalingappa, Krishnappa, M. & Thippeswamy, B. Degradation of polyethylene by *Penicillium simplicissimum* isolated from local dumpsite of Shivamogga district. *Environ. Dev. Sustain.* **17**, 731–745 (2015).
- Kawai, F., Kawabata, T. & Oda, M. Current knowledge on enzymatic PET degradation and its possible application to waste stream management and other fields. *Appl. Microbiol. Biotechnol.* **103**, 4253–4268 (2019).
- Palm, G. J. *et al.* Structure of the plastic-degrading *Ideonella sakaiensis* MHEase bound to a substrate. *Nat. Commun.* **10**, 1–10 (2019).
- Tournier, V. *et al.* An engineered PET depolymerase to break down and recycle plastic bottles. *Nature* **580**, 216–219 (2020).
- Herrero Acero, E. *et al.* Enzymatic surface hydrolysis of PET: Effect of structural diversity on kinetic properties of cutinases from *Thermobifida*. *Macromolecules* **44**, 4632–4640 (2011).
- Danso, D., Chow, J. & Streita, W. R. Plastics: Environmental and biotechnological perspectives on microbial degradation. *Appl. Environ. Microbiol.* **85**, (2019).
- Buchholz, P. C. F. *et al.* Plastics degradation by hydrolytic enzymes: The plastics-active enzymes database—PAZy. *Proteins Struct. Funct. Bioinforma.* **90**, 1443–1456 (2022).
- Otzen, M., Palacio, C. & Janssen, D. B. Characterization of the caprolactam degradation pathway in *Pseudomonas jessenii* using mass spectrometry-based proteomics. *Appl. Microbiol. Biotechnol.* **102**, 6699–6711 (2018).
- Shah, Z. *et al.* Degradation of polyester polyurethane by a newly isolated soil bacterium, *Bacillus subtilis* strain MZA-75. *Biodegradation* **24**, 865–877 (2013).

22. Peng, Y. H. *et al.* Degradation of polyurethane by bacterium isolated from soil and assessment of polyurethanolytic activity of a *Pseudomonas putida* strain. *Environ. Sci. Pollut. Res.* **21**, 9529–9537 (2014).
23. Ahmad, T., Singh, R. S., Gupta, G., Sharma, A. & Kaur, B. *Metagenomics in the search for industrial enzymes. Biomass, Biofuels Biochemicals: Adv. Enzyme Technol.* (2019). <https://doi.org/10.1016/B978-0-444-64114-4.00015-7>.
24. Ribitsch, D. *et al.* A new esterase from *Thermobifida halotolerans* hydrolyses polyethylene terephthalate (PET) and polylactic acid (PLA). *Polymers (Basel)*. **4**, 617–629 (2012).
25. Kawai, F. *et al.* A novel Ca²⁺-activated, thermostabilized polyesterase capable of hydrolyzing polyethylene terephthalate from *Saccharomonospora viridis* AHK190. *Appl. Microbiol. Biotechnol.* **98**, 10053–10064 (2014).
26. Ariaeenejad, S. *et al.* In-silico discovery of bifunctional enzymes with enhanced lignocellulose hydrolysis from microbiota big data. *Int. J. Biol. Macromol.* **177**, 211–220 (2021).
27. Foroozandeh Shahraki, M. *et al.* MCIC: Automated identification of cellulases from metagenomic data and characterization based on temperature and pH dependence. *Front. Microbiol.* **11**, 1–10 (2020).
28. Foroozandeh Shahraki, M. *et al.* A generalized machine-learning aided method for targeted identification of industrial enzymes from metagenome: A xylanase temperature dependence case study. *Biotechnol. Bioeng.* **118**, 759–769 (2021).
29. Chunfang Deng, Renxin Zhao, Zhiguang Qiu, Bing Li, Tong Zhang, Feng Guo, Rong Mu, Yang Wu, Xuejiao Qiao, Liyu Zhang, Jay J. Cheng, Jinren Ni, K. Y. Genome-centric metagenomics provides new insights into the microbial community and metabolic potential of landfill leachate microbiota. *Sci. Total Environ.* **816** (2022).
30. Yoshida S, Hiraga K, Takehana T, Taniguchi I, Yamaji H, Maeda Y, Toyohara K, Miyamoto K, Kimura Y, O. K. A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science* 2016;351(6278)1196–9. <https://doi.org/10.1126/science.aad6359>.
31. Müller, R. J., Schrader, H., Profe, J., Dresler, K. & Deckwer, W. D. Enzymatic degradation of poly(ethylene terephthalate): Rapid hydrolyse using a hydrolase from *T. fusca*. *Macromol. Rapid Commun.* **26**, 1400–1405 (2005).
32. Silva, C. *et al.* Engineered *Thermobifida fusca* cutinase with increased activity on polyester substrates. *Biotechnol. J.* **6**, 1230–1239 (2011).
33. Eberl, A. *et al.* Enzymatic surface hydrolysis of poly(ethylene terephthalate) and bis(benzoyloxyethyl) terephthalate by lipase and cutinase in the presence of surface active molecules. *J. Biotechnol.* **143**, 207–212 (2009).
34. Sulaiman, S. *et al.* Isolation of a novel cutinase homolog with polyethylene terephthalate-degrading activity from leaf-branch compost by using a metagenomic approach. *Appl. Environ. Microbiol.* **78**, 1556–1562 (2012).
35. Silva, C. M. *et al.* Cutinase: A new tool for biomodification of synthetic fibers. *J. Polym. Sci. Part A Polym. Chem.* **43**, 2448–2450 (2005).
36. Ronkvist, Å. M., Xie, W., Lu, W. & Gross, R. A. Cutinase-Catalyzed hydrolysis of poly(ethylene terephthalate). *Macromolecules* **42**, 5128–5138 (2009).
37. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. (2014) <https://doi.org/10.1038/nbt.2942>.
38. Qin, J. *et al.* ARTICLES A human gut microbial gene catalogue established by metagenomic sequencing. **464**, (2010).
39. Almeida, A. *et al.* from the human gut microbiome. *Nat. Biotechnol.* **39**, (2021).
40. Wild, S. P. F. V., Sczyrba, A., Mchardy, A. C. & Strowig, T. An integrated metagenome catalog reveals new insights into the murine gut microbiome. 2909–2922 (2020) <https://doi.org/10.1016/j.celrep.2020.02.036>.
41. Maman, L. G. *et al.* Co-abundance analysis reveals hidden players associated with high methane yield phenotype in sheep rumen microbiome. 1–12 (2020) <https://doi.org/10.1038/s41598-020-61942-y>.
42. Zhu, B. & Gao, G. F. Metagenome-assembled genomes and gene catalog from the chicken gut microbiome aid in deciphering antibiotic resistomes. 1–9 (2021) <https://doi.org/10.1038/s42003-021-02827-2>.
43. Yadav, S. & Jha, R. Strategies to modulate the intestinal microbiota and their effects on nutrient utilization, performance, and health of poultry. *J. Anim. Sci. Biotechnol.* **10**, 1–11 (2019).
44. Chen, Y., Neilson, J. W., Kushwaha, P., Maier, R. M. & Barberán, A. Life-history strategies of soil microbial communities in an arid ecosystem. *ISME J.* **15**, 649–657 (2021).
45. Gharechahi, J., Sarikhan, S., Han, J. L., Ding, X. Z. & Salekdeh, G. H. Functional and phylogenetic analyses of camel rumen microbiota associated with different lignocellulosic substrates. *npj Biofilms Microbiomes* **8**, (2022).
46. Site, G. L. & Delhi, N. crossm Metagenomic analysis of microbial diversity in landfill lysimeter soil of. *Genome Announc.* **5**, 51–52 (2017).
47. Kumar, R. *et al.* Land fill microbiome harbour plastic degrading genes: A metagenomic study of solid waste dumping site of Gujarat, India. *Sci. Total Environ.* **779**, 146184 (2021).
48. Radwan, O. & Ruiz, O. N. Shotgun metagenomic data of microbiomes on plastic fabrics exposed to harsh tropical environments. *Data Br.* **32**, 106226 (2020).
49. Radwan O, Lee JS, Stote R, Kuehn K, Ruiz ON. Metagenomic characterization of microbial communities on plasticized fabric materials exposed to harsh tropical environments. *Int. Biodeterior. Biodegrad.* **154** <https://doi.org/10.1016/j.ibiod.2020.105061>.
50. Chu, B. T., Petrovich, M. L., Chaudhary, A., Wright, D., Murphy, B., Wells, G., Poretsky, R. Metagenomics reveals the impact of wastewater treatment plants on the dispersal of microorganisms and genes in aquatic sediments. *Appl. Environ. Microbiol.* 2018;**84**(5):e02168–17.
51. Gu, S., Fang, L., Xu, X. Using SOAPaligner for short reads alignment. *Curr Protoc Bioinform.* 2013; <https://doi.org/10.1002/0471250953.bi11111s44>.
52. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
53. Li, H. *et al.* The sequence alignment/map format and SAMtools. **25**, 2078–2079 (2009).
54. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. **38**, 1–15 (2010).
55. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite : A web server for clustering and comparing biological sequences. **26**, 680–682 (2010).
56. Aramaki, T. *et al.* Genome analysis KofamKOALA : KEGG Ortholog assignment based on profile HMM and adaptive score threshold. **36**, 2251–2252 (2020).
57. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**(1), 27–30. <https://doi.org/10.1093/nar/28.1.27>. PMID:10592173;PMCID:PMC102409 (2000).
58. Huerta-cepas, J. *et al.* eggNOG 5.0 : A hierarchical , functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. **47**, 309–314 (2019).
59. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* <https://doi.org/10.1101/gr.186072.114> (2015).
60. Jeske, L., Placzek, S., Schomburg, I., Chang, A. & Schomburg, D. BRENDA in 2019 : A European ELIXIR core data resource. 1–8 (2019) <https://doi.org/10.1093/nar/gky1048>.
61. Consortium, T. U. UniProt : A worldwide hub of protein knowledge. 1–10 (2018) <https://doi.org/10.1093/nar/gky1049>.
62. Hiraga, K., Taniguchi, I., Yoshida, S., Kimura, Y. & Oda, K. Biodegradation of waste PET. *EMBO Rep.* **20**, 1–5 (2019).
63. Shahraki, M. F., Atanaki, F. F., Ariaeenejad, S., Ghaffari, M. R., Norouzi-Beirami, M. H., Maleki, M., Salekdeh, G. H., Kavousi, K. A computational learning paradigm to targeted discovery of biocatalysts from metagenomic data: A case study of lipase identification. *Biotechnol. Bioeng.*

64. Marchler-bauer, A. *et al.* CDD/SPARCLE : Functional classification of proteins via subfamily domain architectures. **45**, 200–203 (2017).
65. Ronneberger, O. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596** (2021).
66. Zhang, Y. & Skolnick, J. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
67. Li, D., Liu, C., Luo, R., Sadakane, K., & Lam, T. Sequence analysis MEGAHIT : An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. **31**, 1674–1676 (2015).
68. Luo, C., Tsementzi, D., Kyrpides, N. C. & Konstantinidis, K. T. Individual genome assembly from complex community short-read metagenomic datasets. *ISME J.* **6**, 898–901 (2012).
69. Cantarel, B. I. *et al.* The carbohydrate-active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res.* **37**, 233–238 (2009).
70. Zhang, H. *et al.* dbCAN2 : A meta server for automated carbohydrate-active enzyme annotation. **46**, 95–101 (2018).
71. Kumari, A., Bano, N., Chaudhary, D. R. & Jha, B. Draft genome sequence of plastic degrading *Bacillus* sp. AIIW2 isolated from the Arabian ocean. *J. Basic Microbiol.* **61**, 37–44 (2021).
72. Kanagawa, K., Oishi, M., Negoro, S., Urabe, I. & Okada, H. Characterization of the 6-aminohexanoate-dimer hydrolase from *Pseudomonas* sp. NK87. *J. Gen. Microbiol.* **139**, 787–795 (1993).
73. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. 1–42 (2021).
74. Chaumeil, P., Mussig, A. J., Parks, D. H. & Hugenholtz, P. Genome analysis GTDB-Tk : A toolkit to classify genomes with the Genome *Taxonomy Database*. **36**, 1925–1927 (2020).
75. Kang, D. D. *et al.* MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* <https://doi.org/10.7717/peerj.7359> (2019).
76. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser Interactive metagenomic visualization in a Web browser. **385** (2011).
77. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

Acknowledgements

This research was carried out by the support of the Institute of Biochemistry and Biophysics (IBB), University of Tehran and Department of Systems and Synthetic Biology, Agricultural Biotechnology Research Institute of Iran (ABRII).

Author contributions

K.K.: Methodology, contributed to the computational and bioinformatics data analysis, Writing—review & editing. S.A.: Conceptualization, Methodology, Writing—original draft, Resources, Project administration. D.A.J.: Contributed to the computational and bioinformatics data analysis, Writing—original draft.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023