Environmental Microbiome

RESEARCH

Open Access

Polycyclic aromatic hydrocarbon (PAH) biodegradation capacity revealed by a genome-function relationship approach

Yue Huang[1], Liguan Li[1], Xiaole Yin[1] and Tong Zhang[1*]

Abstract

Background Polycyclic aromatic hydrocarbon (PAH) contamination has been a worldwide environmental issue because of its impact on ecosystems and human health. Biodegradation plays an important role in PAH removal in natural environments. To date, many PAH-degrading strains and degradation genes have been reported. However, a comprehensive PAH-degrading gene database is still lacking, hindering a deep understanding of PAH degraders in the era of big data. Furthermore, the relationships between the PAH-catabolic genotype and phenotype remain unclear.

Results Here, we established a bacterial PAH-degrading gene database and explored PAH biodegradation capability via a genome-function relationship approach. The investigation of functional genes in the experimentally verified PAH degraders indicated that genes encoding hydratase-aldolase could serve as a biomarker for preliminarily identifying potential degraders. Additionally, a genome-centric interpretation of PAH-degrading genes was performed in the public genome database, demonstrating that they were ubiquitous in *Proteobacteria* and *Actinobacteria*. Meanwhile, the global phylogenetic distribution was generally consistent with the culture-based evidence. Notably, a few strains affiliated with the genera without any previously known PAH degraders (*Hyphomonas*, *Hoeflea*, *Henriciella*, *Saccharomonospora*, *Sciscionella*, *Tepidiphilus*, and *Xenophilus*) also bore a complete PAH-catabolic gene cluster, implying their potential of PAH biodegradation. Moreover, a random forest analysis was applied to predict the PAH-degrading trait in the complete genome database, revealing 28 newly predicted PAH degraders, of which nine strains encoded a complete PAH-catabolic pathway.

Conclusions Our results established a comprehensive PAH-degrading gene database and a genome-function relationship approach, which revealed several potential novel PAH-degrader lineages. Importantly, this genome-centric and function-oriented approach can overcome the bottleneck of conventional cultivation-based biodegradation research and substantially expand our current knowledge on the potential degraders of environmental pollutants.

Keywords PAH, Biodegradation, Database mining, Functional gene, Genome-centric analysis, Genotype–phenotype relationship, Random forest

*Correspondence:
Tong Zhang
zhangt@hku.hk
[1] Environmental Microbiome Engineering and Biotechnology Lab, Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong, China

Background

Polycyclic aromatic hydrocarbon (PAH) contamination has been a global environmental issue for decades. PAHs are a group of organic compounds composed of two or more fused aromatic rings with natural and anthropogenic sources [1–3], which are well-recognized as carcinogenic, teratogenic, and genotoxic compounds

[4–6]. They are ubiquitous in the environments [3, 7, 8] at relatively high concentrations, likely to accumulate in animal tissues and vegetation due to their high lipophilicity [9–11], and thus harmful to ecosystems and human health. Although PAHs could be eliminated by adsorption [12], volatilization, and photochemical degradation [13], microbial degradation is one of the dominant mechanisms of PAH removal in natural environments [7, 14, 15]. In the past half-century, a wide variety of PAH-catabolic bacteria, archaea, fungi, and microalgae have been isolated mostly from contaminated soils and sediments [16], among which bacteria-mediated biodegradation has been extensively studied. Currently, the identified PAH-degrading bacteria are distributed in diverse genera, such as *Pseudomonas* spp. [17, 18], *Sphingomonas* spp. [19, 20], *Mycobacterium* spp. [21, 22], *Rhodococcus* spp. [23, 24], *Burkholderia* spp. [25, 26], etc. Nevertheless, it is believed that most PAH-degrading bacteria still hide in plain sight due to the isolation bottleneck [27]. Therefore, there is a dire need for a new method to efficiently identify novel potential PAH degraders.

Basically, biological traits are developed based on their encoding genes. Traditional culture-based approaches have set a good foundation for understanding PAH biodegradation pathways, functional genes, and enzyme-catalyzed reactions [15]. Commonly, in the upper pathway, biodegradation of PAHs is initially catalyzed by ring-hydroxylating dioxygenases (RHDs) [28] and followed by other enzymes encoded by *nah* gene cluster (*nahBCDEF*), which is well characterized in naphthalene and phenanthrene aerobic biodegradation [29, 30], but with a broad substrate specificity to aromatic compounds [31]. Conventionally, the *nahAc* encoding α-subunit of RHDs was usually employed as the biomarker to demonstrate the diversity and abundance of RHDs in PAH-degrading isolates and multiplex systems by quantitative real-time PCR [32, 33]. Nevertheless, owing to its high specificity, the primers of *nahAc* only target a relatively narrow range of *nahAc*-like sequences and result in an underestimated PAH-degrading consortia [34, 35]. Moreover, other PAH-catabolic gene clusters also exist in Gram-negative bacteria, including *nag* [36], *pah* [18], *ndo* [37], and *phn* [25] gene clusters, as well as in Gram-positive bacteria, including *nar* [16, 38], *phd* [39], *nid* [40], and *pdo* [41] gene clusters. However, a unified database integrating the diverse PAH-degrading genes for exploring potential novel PAH degraders is still lacking.

In the era of high-throughput sequencing, access to the genome information of currently uncultivable microbes has opened a new window to explore this topic. Concurrent with the advance of long-read sequencing technologies, the number of high-quality genomes increased exponentially. The current technology improvements make it possible to interpret biological traits based on their whole genomes instead of single or multiple biomarkers. Meanwhile, the biodegradation processes of PAHs and functional enzymes are very diverse and complicated in different species (i.e., *Pseudomonas* spp., *Mycobacterium* spp., and *Rhodococcus* spp.) and habitats (i.e., aerobic, anoxic, and anaerobic) [16, 42]. The unknown alternative genes or enzymes for individual steps may generally exist, which are not represented in the currently available gene database. Hence, it remains a big challenge to properly identify these genetic hints based on similarity search, not to mention their functional potentials. Fortunately, the introduction of the Hidden Markov Model (HMM) [43] has made it possible to detect remote homology between proteins with high efficiency and accuracy. This is a popular method predicting biological functions based on conserved protein domains, and has been widely applied to gene identification [44], phylogenetic analysis [45], and database construction [46]. Therefore, HMM was employed in the present study aiming to improve the accuracy of functional gene identification and profile the functional genes at scales ranging from a single isolate to the whole genome database.

Herein, we collected the protein sequences of key enzymes responsible for the upper pathway of PAH metabolism and established a dedicated database for similarity- and HMM-based searches. By investigating the distribution of PAH-degrading genes in the known degraders with complete genomes, we evaluated the performance of two alignment methods, paving the way for the identification of novel PAH degraders on a large scale. Then, a genome-centric interpretation of PAH-catabolic genes was conducted in the NCBI genome database, aiming to depict a phylogenetic distribution of PAH-degrading genes and discover novel lineages containing potential PAH degraders. Finally, a preliminary exploration to link the PAH-catabolic genotypes to phenotypes via a random forest analysis was performed and applied to predict the PAH biodegradation trait in the complete genome database. In general, this study based on the genome-function relationship represents a paradigm shift and provides a novel insight into conventional biodegradation research.

Methods

Construction and validation of PAH-degrading gene database

The PAH-degrading gene database was constructed following the workflow depicted in Fig. 1. The seed protein sequences were collected based on naphthalene and phenanthrene aerobic biodegradation pathways in the Kyoto Encyclopedia of Genes and Genomes

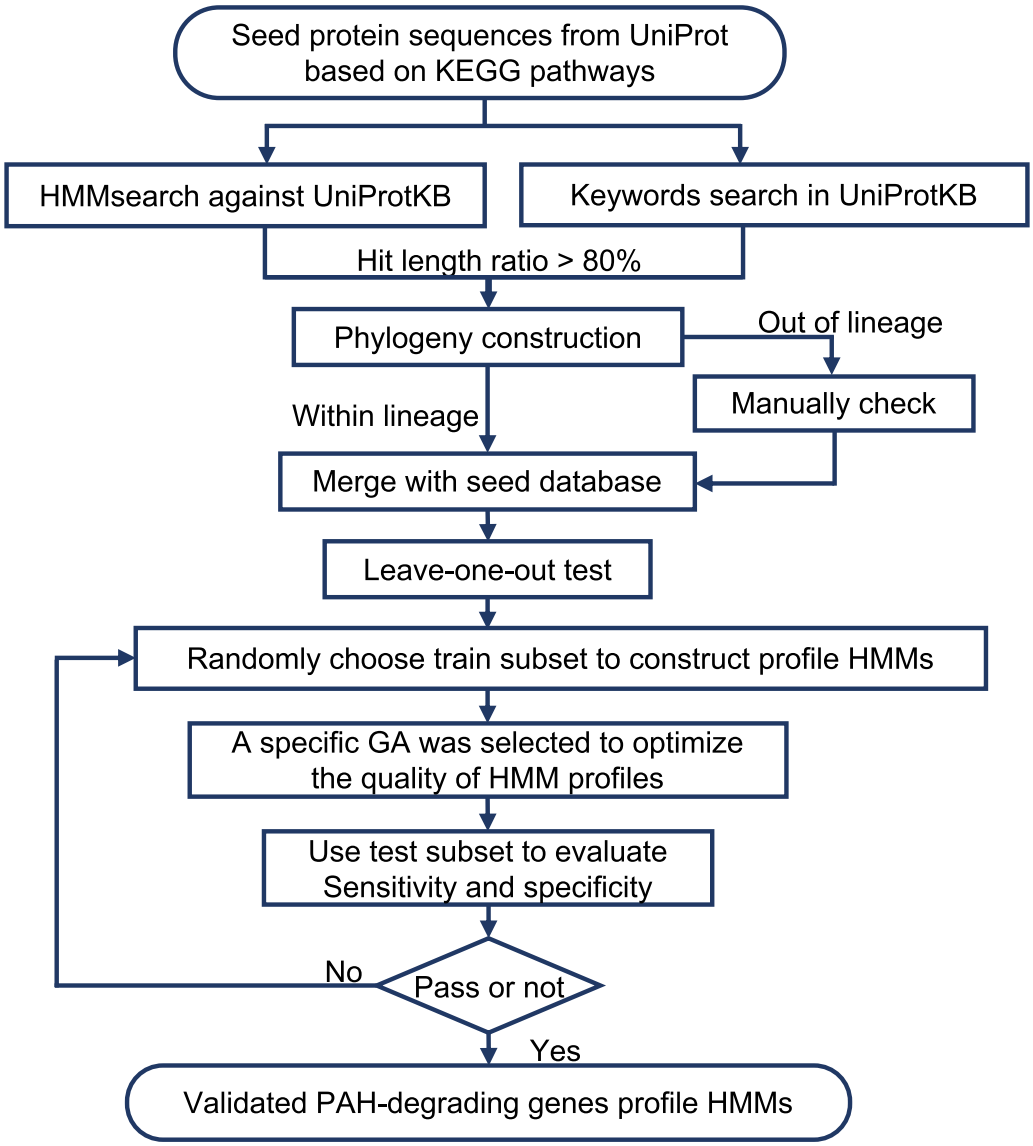

Fig. 1 The workflow of PAH-degrading gene database construction

(KEGG) pathway database. Those PAH-degrading genes with more than three protein sequences were used to construct initial profile HMMs. To retrieve more PAH-degrading protein sequences, a database search against the UniProtKB was performed by two strategies, namely HMMsearch and keywords. Then, a phylogenetic tree topology was constructed using MEGAX (v10.0.5) [47], and the questionable sequence, which was distinct from the core cluster, was filtered out after manually checking. Subsequently, a leave-one-out test was applied for fine adjustment of the protein database, where the sequence leading to significantly low sensitivity and specificity values was marked. Next, a training subset (two-thirds of sequences) was randomly chosen from the enriched database to construct profile HMM, while the rest were used as the test dataset for validation. A specific gathering threshold (GA) was selected for each profile HMM, and the optimal GA value was obtained according to the sensitivity and specificity using a bash script. The profile HMMs with both sensitivity and specificity values exceeding 90% were retained as the validated models. It is a loop that will stop when there is no further addition to this expanded database to form the final version.

Genome and protein sequence

The GenBank flat file (.gbff) of 22,507 bacteria with complete genomes, 263,643 bacteria with draft genomes (updated on March 2021), and 7045 archaeal genomes (updated on Apr 2023) were downloaded from the NCBI database. Their nucleotide sequences were extracted by an in-house Python script. Then, their open reading frames (ORFs) were predicted using prodigal (v2.6.3) [48].

PAH-degrading genes identification and PAH-degrading bacteria prediction

Two different methods (similarity- and HMM-based pipelines) were adapted to identify PAH-degrading genes in 47 experimentally confirmed PAH degraders with complete genomes. The similarity-based search was performed using DIAMOND (v2.0.8.146) [49] with an identity of over 70% and a hit length ratio of over 70%. For the HMM-based search, MAFFT (v7.310) [50] and hmmbuild from HMMER 3.0 suite [51] were used to align sequences and generate the profile Hidden Markov models. PAH-catabolic genes were identified using the profile HMMs and hmmsearch at -cut_ga mode. After comparing the accuracy and efficiency of the two methods, only the HMM-based approach was employed to interpret the distribution of PAH-degrading genes in the public genome database. The phylogenetic trees were visualized using iTOL (v6.6) [52].

To further investigate the genotype–phenotype relationships in the PAH-degrading bacteria, a supervised learning algorithm, Random Forest, was applied in this study. The analysis was performed with the R package 'randomForest' [53] using the maximum GA bit score of each gene in the genome. In addition to the numbers of variables at each node (m_{try}) and trees in the forest (n_{tree}), the ratio of True/False in the training dataset was also considered since the majority of bacteria were not PAH degraders. Basically, two-thirds of genomes were randomly chosen from the dataset for training the model, while the rest were utilized for verification. Four standard metrics are used to evaluate the quality of the proposed model, consisting of sensitivity (Sn), specificity (Sp), overall accuracy (Acc), and Mathew's correlation coefficient (MCC) with the following definitions:

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}}$$

where *TP* (true positive) and *TN* (true negative) are correctly predicted PAH-degrading positive and negative bacteria, respectively. *FP* (false positive) and *FN* (false negative) indicate falsely predicted PAH-degrading positive and negative bacteria, respectively. Among these metrics, *MCC* is the most stringent one, as it takes into account both accuracy and error rates.

Results and discussion

Experimentally verified PAH-degrading strains

To date, numerous PAH-degrading strains (more than 200) have been isolated from various habitats as aforementioned. More than 95% of them were from the domain of bacteria, but limited information was available regarding their genome. Therefore, the PAH-degrading strain database only comprised genome sequences of 95 reported PAH-degrading bacterial strains. The detailed information was summarized in the supplementary material (Additional file 1: Table S1). The profiling of these known PAH-degrading strains (Fig. 2) demonstrated that they were phylogenetically diverse owing to the evolutions in different habitats as well as horizontal gene transfer [54, 55]. Notably, 95% of those degradation strains were affiliated with two phyla of *Proteobacteria* and *Actinobacteria*. Of the 42 genera represented, *Pseudomonas* (14%) constitutes most of the reported PAH-degrading strains, followed by *Rhodococcus* (10%), *Mycobacterium* (9%), and *Sphingobium* (7%). Additionally, most of the identified PAH-degrading strains could catabolite multiple PAHs, such as *Pseudomonas putida* OUS82 and *Mycolicibacterium vanbaalenii* PYR-1, supporting that the functional enzymes have a broad substrate specificity to multiple aromatic compounds (Additional file 1: Table S1).

PAH-degrading protein sequence database

In addition to the PAH-degrading strain database, it is critical to construct a protein database for the annotation of related gene clusters. Enzymes related to the aerobic biodegradation pathways of six common PAHs (naphthalene, phenanthrene, anthracene, fluorene, pyrene, and benzo[*a*]pyrene) have been well archived in the KEGG database, as well as their associated enzymes. The protein sequences were retrieved based on the naphthalene and phenanthrene biodegradation pathways to form a seed database. A comprehensive database was prepared by expanding the seed database following the workflow described in the Methods. After expanding, the *nar* gene cluster was included to

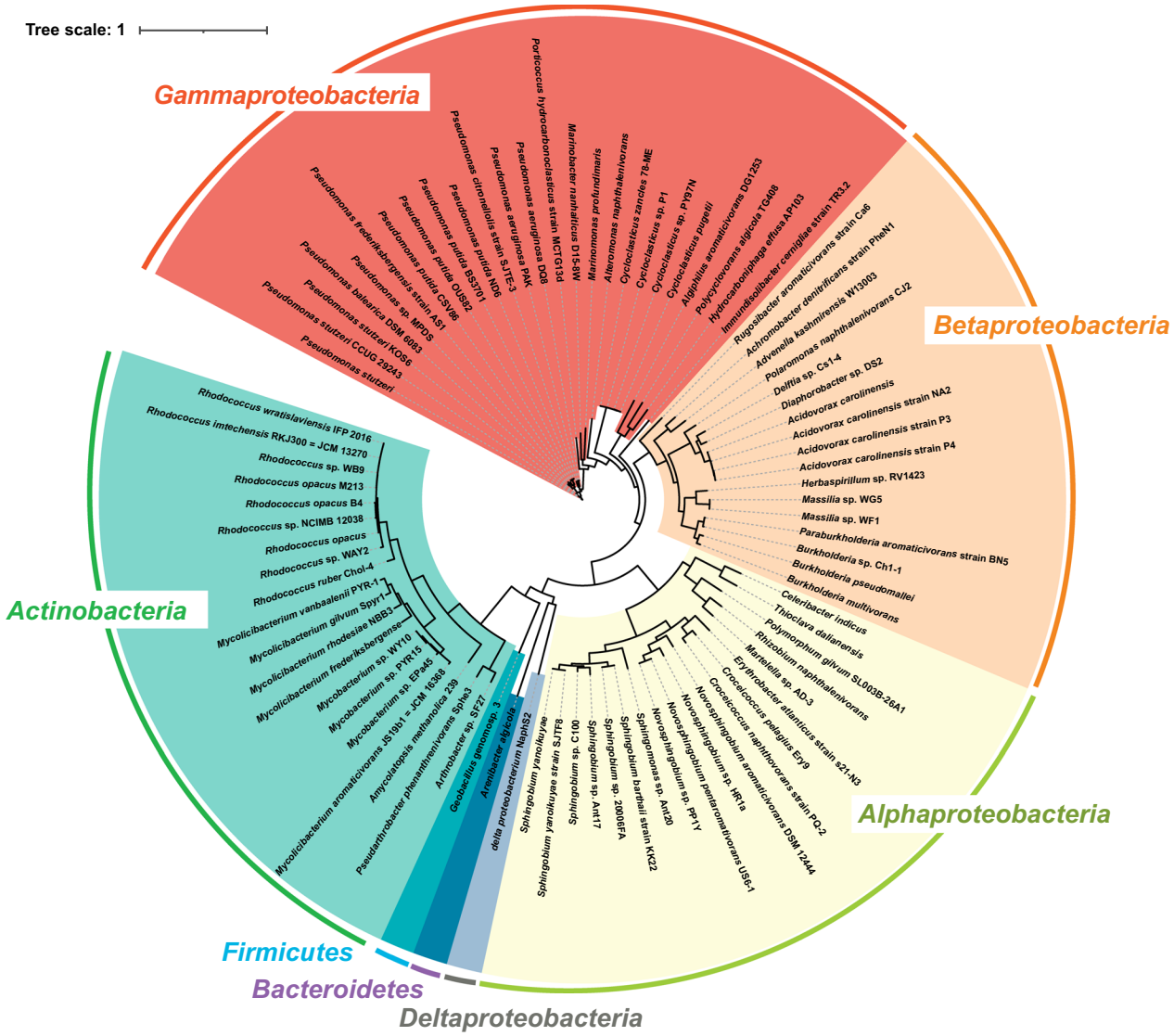

Fig. 2 A phylogenetic tree of the reported PAH-degrading bacterial strains based on their genomes. The tree was constructed using gtdb-tk (v1.7.1) [91]. The details of these PAH-degrading strains are summarized in Additional file 1: Table S1

improve the sensitivity for Gram-positive bacteria, such as *Rhodococcus* spp. Simultaneously, for RHDs, only *nahAc* genes encoding the ion-sulfur subunit were retained in the database because they were conserved and could serve as a biomarker for RHDs [32, 56, 57]. Likewise, *nahF* and *phdK* genes were excluded due to their poor phylogenetic conservation, which is hard to choose a suitable GA cut-off to ensure both sensitivity and specificity. Notably, *nag*, *ndo*, *pah*, *phn*, *dox*, and *bph* gene clusters were also included in the current database since they were homologous to the *nah* genes cluster. Eventually, a total of 1,191 manually checked PAH-degrading protein sequences were included in the comprehensive database, a twice-fold increase in the number of sequences compared with the seed database (Additional file 3: Figure S1). These reference sequences were from 17 different degradation genes, which could be classified into three types based on degradation mechanisms, namely (1) *nah*, (2) *nid* and *phd*, and (3) *nar* gene clusters (Additional file 3: Figures S2–4). The detailed information on each protein sequence was summarized in the Supplementary Information (Additional file 2: Tables S2–8). In the database, the *nah* gene cluster (843 protein sequences, 71% of total sequences in the database) was the most dominant type, followed

by the *nid* and *phd* gene cluster (26%) and the *nar* gene cluster (3%).

Similarity- and HMM-based searches for verified PAH-degrading strains

In the present study, similarity- and HMM-based methods were employed to identify functional genes in the 47 experimentally verified PAH-degrading strains (NCBI assembly level = Complete) (Additional file 1: Table S1 and Fig. 3). Both approaches could accurately identify most PAH-degrading genes, whereas the HMM-based method allows us to retrieve the potential PAH-degrading genes which, however, cannot be identified by the similarity-based strategy, such as *nahAc* genes in *Cycloclasticus* and *Acidovorax carolinensis*. Because HMM captures conserved protein domains necessary for the protein function, the HMM-based method thus is more sensitive and rapid in detecting remotely homologous sequences on a large scale. Notably, not all the PAH-degrading strains have a complete pathway of PAH biodegradation, such as *Archomobacter denitrificans* PheN1, *Celeribacter indicus* P73, and *Martelella* sp. AD-3

Fig. 3 The PAH-degrading gene distribution in 47 experimentally validated PAH-degrading bacteria. The heatmap compares similarity- and HMM-based searches. The right column demonstrates the location of PAH-degrading genes, and detailed information is depicted in Additional file 3: Figure S6. Notably, the identifications of *nahE* in *Mycobacteriaceae* and *Rhodococcus* were *phdJ* and *narC*, respectively

(Fig. 3). A parsimonious interpretation of gene deletion was the existence of alternative genes/enzymes for the individual steps in these strains, which have not yet been reported.

Furthermore, both methods can accurately distinguish the *nahAc*, *nidA*, and *narAa* encoding the large subunits of RHDs even though they showed significant but moderate sequence homology to each other [21, 34]. However, both approaches cannot alleviate the misclassification issue on *nahE*, *phdJ*, and *narC* in Gram-positive strains. The misclassification was defined as the single protein sequence being classified into multiple gene types under the optimized cut-off parameters. For example, the *phdJ* gene was identified as the *nahE* gene by similarity- and HMM-based methods in *Mycobacteriaceae* spp. Because both NahE and PhdJ were *N*-acetylneuraminate lyase subgroup members with a conserved $(\beta/\alpha)_8$ barrel structure, two strictly conserved active site residues (tyrosine and lysine), and a GXXGE motif (Gly-61, Thr-62, Phe-63, Gly-64, and Glu-65) [58]. It is not easy to distinguish them based on either similarity of the whole sequence or protein domains. In addition, the *narC* gene encoding aldolase in *Rhodococcus* spp. was classified into the *nahE* group, which was located near the *narB* gene and involved in the biodegradation of PAH compounds [24, 38, 59]. The misclassification suggested that hydratase-aldolase-coding genes were more conservative than RHDs-coding genes, consistent with the phylogenetic analysis of these PAH-degrading genes (Additional file 3: Figure S2). Meanwhile, hydratase-aldolase-coding genes were identified in 46 strains (98%), and, therefore, genes encoding hydratase-aldolase may be a superior biomarker for PAH degraders. Moreover, the primers to amplify *nahE*, *phdJ*, or *narC* have been well designed and evaluated in the previous studies [34, 58], providing a rapid way to initially explore the ecological role and degradation potential of PAH-catabolic bacteria in the natural environment. Notably, the conclusion from genome-centric interpretation was in agreement with the results based on the phylogenetic analysis of PAH-catabolic enzymes, which proposed *pahE* (including *nahE*, *phdJ*, and *narC*) as a functional marker because all the enzymes encoded by *pahE* clustered in an independent clade [34].

Intriguingly, *narAa* and *narAb* were also identified in *Mycobacteriaceae* spp., which were Gram-positive strains and contained a complete *nid* and *phd* gene cluster as well. We suspected the involvement of the enzyme encoded by *narA* during the initial attack of PAH biodegradation, but scientific evidence is still lacking. It requires more experimental validation conducted by transcriptomics to further investigate the expression of *narA* during the biodegradation process. Meanwhile, the location and gene arrangement of PAH-catabolic genes were investigated in the 47 identified PAH-degrading strains, showing similar gene arrangements in *Pseudomonadaceae* spp., *Comamonadaceae* spp., and *Mycobacteriaceae* spp. (Additional file 3: Figure S5). Interestingly, plasmid-bearing catabolic genes were detected in *Pseudomonas* spp., *Sphingobium* spp., and *Rhodococcus* spp. with a high frequency (8 out of 13 strains) (Fig. 3 and S6). This result was consistent with numerous studies that characterized the functional genes in PAH degraders in the genera of *Pseudomonas* [54, 60], *Sphingobium* [61, 62], and *Rhodococcus* [24, 59]. The presence of degradation genes on the transmissible plasmids, such as NAH7 [63], pKS14 [64], and pNL1 [65], has indicated easy spreading of PAH-catabolic ability via horizontal gene transfer in contaminated sites [54, 66].

The distribution of PAH-catabolic genes and strains in the public database

Subsequently, a large-scale survey in the NCBI database, including all complete-, scaffold-, contig-, and chromosome-level bacterial assemblies, was conducted to investigate the genome-centric portrait of PAH-degrading genes (Fig. 4). At the phylum level, the PAH-degrading genes were ubiquitous in *Proteobacteria* and *Actinobacteria*, in agreement with the result of our collected PAH-degrading strain database, proving the representativeness of our genomic database. Meanwhile, they were also found in other phyla, such as *Firmicutes*, and *Chloroflexi*, implying a phylogenetic diversity of PAH-degrading strains. Significantly, *nah* genes were the most widely distributed degradation genes in the public database, especially in Gram-negative strains of *Proteobacteria*. In *Actinobacteria*, three types of PAH-catabolic genes were observed, where the enzymes encoded by *nid* and *phd* genes were a conservative trait for PAH-catabolic strains in the family of *Mycobacteriaceae*. Moreover, the enzymes encoded by *nar* gene cluster were only identified in Gram-positive strains. At the family level, over 30% of strains with PAH-degrading genes were affiliated with *Sphingomonadaceae*, *Pseudomonadaceae*, *Nocardiaceae*, and *Mycobacteriaceae*, indicating that these strains constituted the majority of PAH degraders. Generally, these genome-centric results were consistent with cultivation-based experimental data (Fig. 2).

Additionally, there were 173 strains with a complete PAH-catabolic gene cluster and 52 with a near-complete PAH-catabolic gene cluster (one gene missing). The phylogenetic tree and details of these 225 strains were summarized in Additional file 3: Figure S7 and Additional file 2: Table S19, respectively. Intriguingly, in addition to those strains phylogenetically close to the well-known PAH degraders, nine strains in the genera

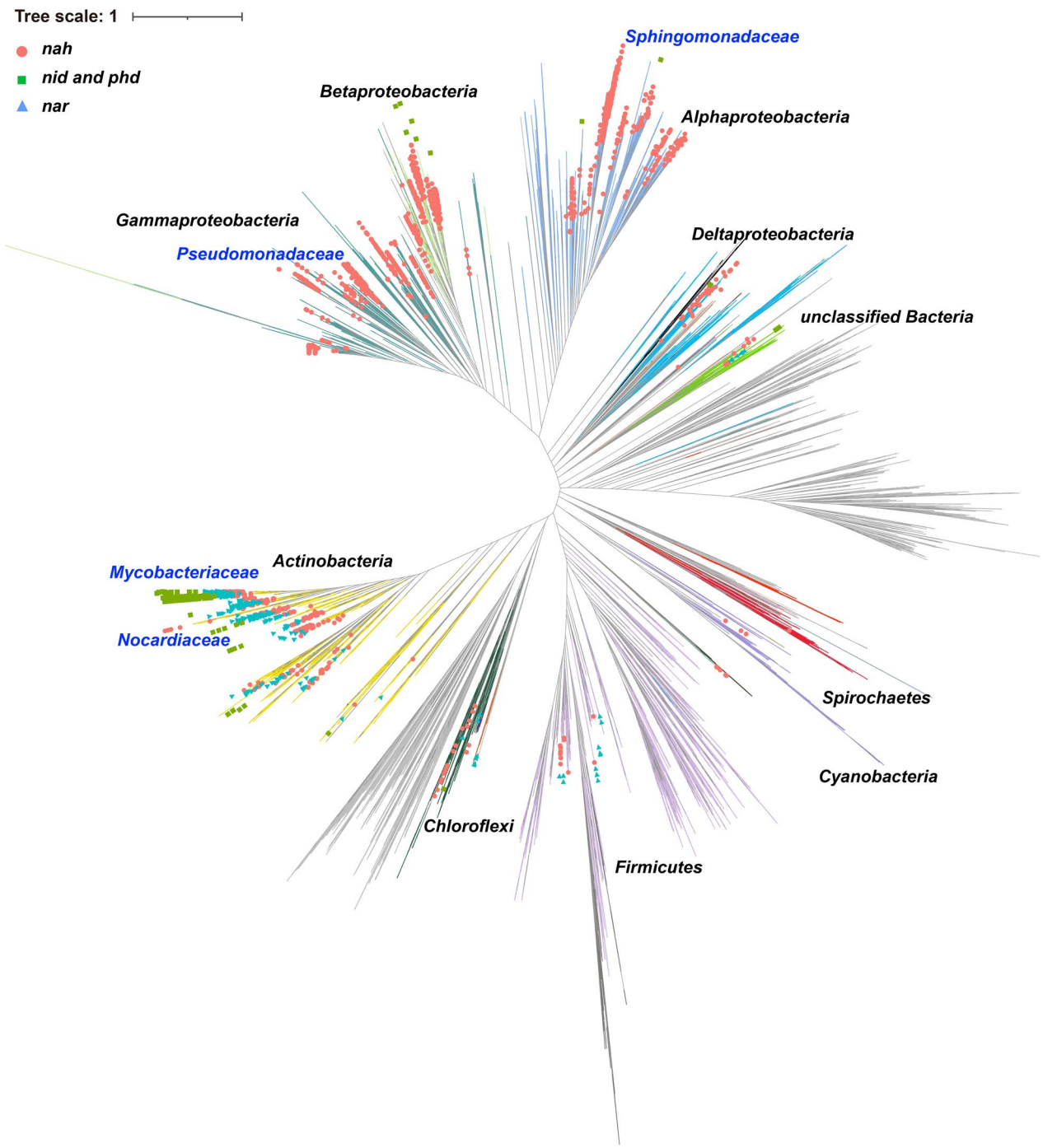

Fig. 4 The phylogenetic distribution of PAH-degrading genes. *nah* genes (*nahAc*, *nahB*, *nahC*, *nahD*, and *nahE*) are shown in pink circle plots. *nar* genes (*narAa*, *narAb*, *narB,* and *narC*) are shown in blue triangle plots. *nid* and *phd* genes (*nidA*, *nidB*, *nidD*, *phdE*, *phdF*, *phdG*, *phdI,* and *phdJ*) are shown in green square plots. The top four families with the most PAH-catabolic genes, *Pseudomonadaceae*, *Sphingomonadaceae*, *Mycobacteriaceae*, and *Nocardiaceae*, are highlighted in blue color

of *Hyphomonas*, *Hoeflea*, *Henriciella*, *Saccharomonospora*, *Sciscionella*, *Tepidiphilus*, and *Xenophilus* also bore the potential of PAH biodegradation owing to their possession of a complete PHA-degrading gene cluster (Additional file 3: Figure S8). Seven of them were isolated from marine water, and the rest two were from production water (*Tepidiphilus* sp. J18 [67]) and soil (*Xenophilus azovorans* DSM 13,620 [68]). Despite

the enrichment of these genera observed in PAH-contaminated sites [69, 70], related PAH-degrading isolates have not yet been reported, probably owing to the isolation bottleneck. Therefore, these seven genera were potential novel PAH-degrader lineages.

Furthermore, we performed a preliminary investigation in archaeal assemblies (data not shown), revealing that two halophilic archaea, namely *Halopenitus malekzadehii* and *Halobellus rufus*, contained a *nahE* gene. They were affiliated with *Halorubraceae* and *Haloferacaceae*, respectively, and phylogenetically related to the identified PAH-degrading archaea (at the family level) [71, 72]. Interestingly, a gentisate-1,2-dioxygenase-like gene (*gdoA*) was also identified in both archaeal assemblies based on a similarity search (>75% identity), which was homologous to bacterial dioxygenases and involved in the aromatic degradation in *Haloferacaceae* sp [73]. Nevertheless, the exploration of archaea-mediated PAH biodegradation is still in its infancy, and archaeal PAH-degrading genes were not included in the present PAH-degrading gene database, requiring more studies to pave the way for investigating biological traits on a genome scale.

Prediction of PAH-degrading strains in the complete genome database

In the random forest algorithm, the three most important parameters were the number of trees (n_{tree}), variables randomly chosen at each node split (m_{try}), and the composition of the training dataset. When the number of non-degraders was 3 to 8 folds larger than the number of PAH degraders in the dataset, the classifiers achieved a high accuracy with average *MCC* values of ~ 0.982 (Additional file 3: Figure S8a). Theoretically, a higher value of n_{tree} will lead to better accuracy, but the computation time will increase simultaneously. Additionally, theoretical and empirical research has highlighted that classification accuracy is more sensitive to m_{try} than n_{tree} [74]. Therefore, n_{tree} was fixed at 2000 in the present study, and m_{try} was optimized from 1 to 17 with a step size of 1 to generate the prediction model. The error rate decreased with the increase of m_{try} value and leveled off at a low error rate of 0.056 after the m_{try} value was set as 3 (Additional file 3: Figure S8c). Then, under the optimized parameters, we noticed that *nahE*, *phdJ*, and *phdG* genes played a crucial role in the classifier based on the high mean decrease accuracy and mean decrease gini values [75, 76], supporting the proposal of genes encoding hydratase-aldolase as a new biomarker for PAH-degrading strains [34] (Additional file 3: Figure S9). In contrast, *nahF* with low mean decrease accuracy and gini values was, therefore, excluded from the PAH-degrading gene database.

The optimized model (random forest classifier) was applied to predict PAH-degrading bacterial strains based on the results from the HMM-based search in the complete genome database (Fig. 5). Given that the degradation mechanism in *Mycobacteriaceae* was only reported via the enzymes encoded by *nid* and *phd* genes, strains were divided into three groups before model construction and prediction, namely, Gram-negative, Gram-positive, and *Mycobacteriaceae*. Only the strain whose prediction was consistent with its group tag would be output. In total, 28 strains were newly predicted to be capable of PAH biodegradation, including 14 strains from the Gram-negative group, 7 strains from the Gram-positive group, and 7 strains from the *Mycobacteriaceae* group. Among these newly predicted PAH degraders, nine strains contained a complete PAH-catabolic gene cluster. In the Gram-negative group, most strains (12) were from *Proteobacteria* and phylogenetically related to the identified degraders. In addition, *Thermomicrobium roseum* DSM 5159 and *Sediminispirochaeta smaragdinae* DSM 11,293 only harbored the *nahE* gene and were affiliated with the phyla of *Chloroflexi* and *Spirochaetes*, respectively. However, no PAH-degrading strain has been isolated from these two phyla to date, and the prediction needs further evaluation. Probably, this hydratase-aldolase-coding gene was involved in other catabolic processes since both strains were isolated from oligotrophic environments [77, 78]. In the Gram-positive group, all potential PAH-degrading strains were affiliated with *Actinobacteria*, of which PAH degrader has been reported previously.

Apart from the nine strains with a complete PAH-degrading pathway which were not publicly available, we purchased six strains from DSMZ for further experimental verification, including two predicted PAH-catabolic strains. However, none of them showed a PAH biodegradation capability during 30-day incubation with naphthalene as the sole carbon source (Additional file 3: Figure S10, the degradation experiment was described in SI). But gene expression is an intricate process controlled by the joint effect of multiple aspects, such as environmental factors [79], quorum sensing [80], etc. In addition, promoters (e.g., P_{pahA} and P_{pahR} [81]) and regulators (e.g., *nahR* [82], *nagR* [83], and *narR* [59]) also play an essential role in the expression of PAH degradation. Moreover, the expression of regulatory genes requires induction of naphthalene [59, 84] or its degradation metabolite, salicylate [29, 85]. Hence, we could only reach a limited conclusion that these strains did not exhibit their PAH-catabolic trait under this specific experimental condition. Simultaneously, among these six strains, three contained a *nahE*

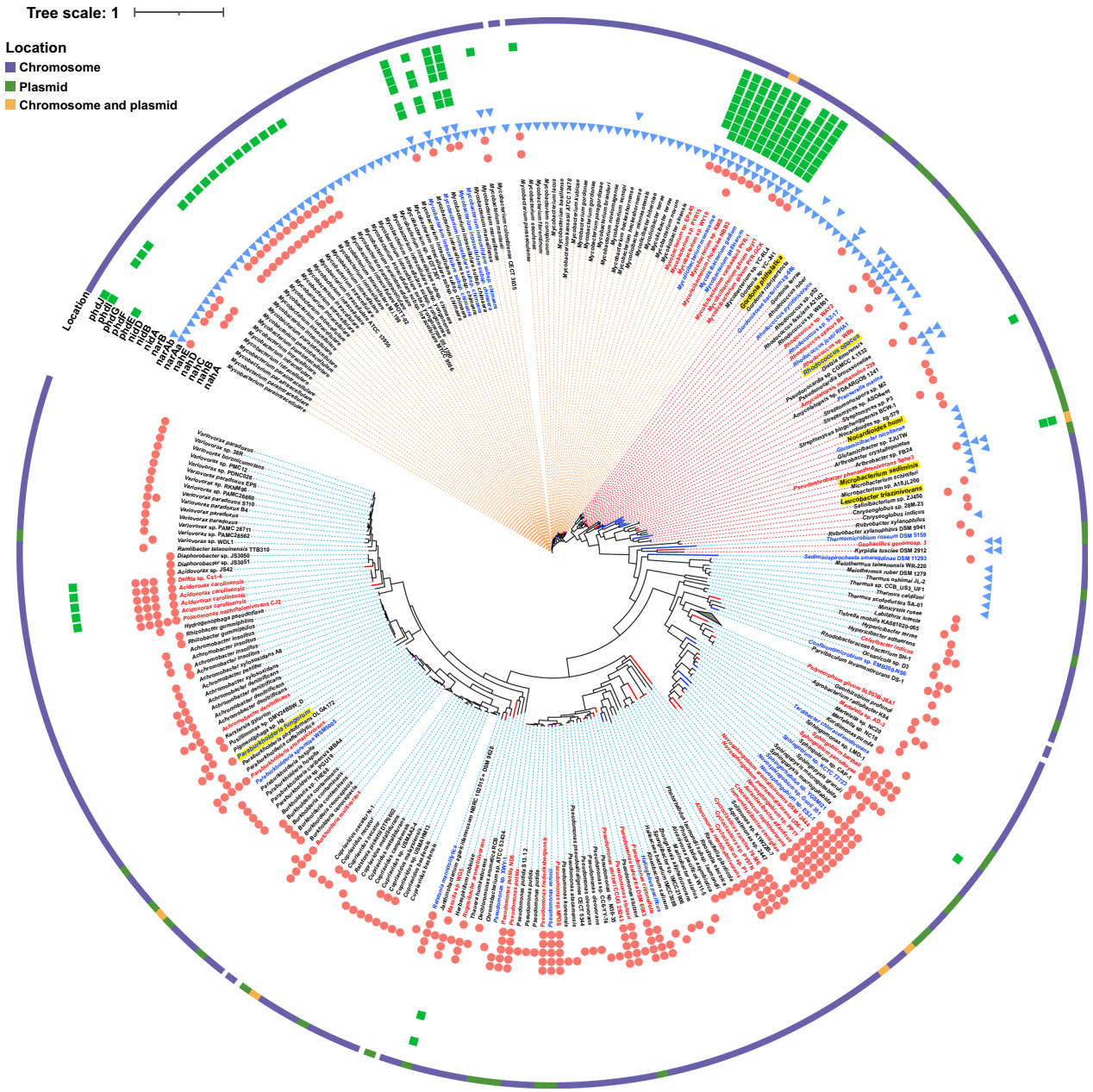

Fig. 5 The distribution of PAH-degrading genes in the NCBI complete genome database. From the outside to the inside: (1) The color in the circle indicates the location of identified PAH-degrading genes; (2) The pink circle, blue triangle, and green square indicate *nah*, *nar*, and *nid* and *phd* gene cluster, respectively. (3) The red tag represents the 47 identified PAH-degrading bacteria, while the blue tag denotes the 28 predicted PAH-degrading bacteria by random forest analysis. Six pure strains selected for experimental verification are depicted with yellow shades. (4) The species affiliated with the family *Mycobacteriaceae* are distinguished by orange branch dashed lines, whereas the branch dashed lines of Gram-negative and Gram-positive strains are blue and red, respectively. Notably, the *nahE* identified in *Mycobacteriaceae* was misclassification

gene and one carried a *nahAc* gene, demonstrating that a single biomarker, either *nahAc* or *nahE*, could not be an entirely reliable indicator for PAH degraders. Moreover, the two predicted PAH degraders indeed had an incomplete *nah* or *nar* gene cluster, hinting that the poor performance was probably due to the functional gene deletion. Meanwhile, the result also suggested that random forest analysis might be aggressive to some extent when applied to predict biological traits because every enzyme involved in PAH biodegradation was indispensable.

Conclusions

In the present study, a comprehensive bacterial PAH-degrading gene database was established, and a genome-centric portrait of bacterial PAH-degrading competency was depicted. Then, a global view of PAH-catabolic genes' phylogenetic distribution was investigated in the public database, showing a wide distribution in *Proteobacteria* and *Actinobacteria*. Simultaneously, seven potential novel PAH-degrader lineages were observed since a few strains from these genera born a complete PAH-catabolic gene cluster. Furthermore, random forest analysis was employed to predict potential PAH degraders in the complete genome database. In total, 28 strains were predicted as potential new PAH degraders, including nine strains encoding a complete PAH-degrading pathway.

Nevertheless, we have to keep in mind that gene expression involves the coordination of multiple biological traits, such as regulators, promoters, and genes encoding lower pathways. Meanwhile, compared to aerobic bacteria-mediated PAH biodegradation, it has been reported that PAHs can also be biodegraded by anaerobic bacteria [86–88], fungi [89], halophilic archaea [71], and microalgae [90] via significantly different pathways. These factors were not considered in the present study, requiring more experimental evidence and studies to move forward. Likewise, the accuracy of this machine learning-based and function-orientated method needs to be further evaluated by experiments. Nevertheless, we believe the method presented in this study could facilitate the exploration of alternative PAH-degrading genes, enzymes, and novel degradation mechanisms.

Abbreviations

GA	Gathering threshold
HMM	Profile hidden markov model
KEGG	Kyoto encyclopedia of genes and genomes
NCBI	National center biotechnology information
ORF	Open reading frame
PAH	Polycyclic aromatic hydrocarbon
RHDs	Ring-hydroxylating dioxygenases

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40793-023-00497-7.

Additional file 1. Details of 95 identified PAH-degrading bacterial strains.

Additional file 2. The summary of protein sequences in the PAH-degrading gene database. Detailed information of 225 strains with complete/near-complete PAH-degrading gene cluster.

Additional file 3. Supplementary method and figures.

Acknowledgements

We acknowledge the University of Hong Kong for the postgraduate studentship and postdoctoral fellowship. We thank the HKU Computer Center for providing the High Performance Computing System. YH wants to thank Dr. Zhong Yu for his insightful discussions and Dr. Yubo Wang for her inspiring study. All authors wish to appreciate the help of Miss Vicky Fung.

Author contributions

YH analyzed the data, prepared the figures and tables, and wrote the original manuscript. LL contributed to the methodology. XY collected the genomic resources. TZ conceived and supervised the study. All authors contributed to manuscript revision and editing. All authors read and approved the final manuscript.

Funding

This work was supported by Hong Kong RGC GRF (172061/20E).

Availability of data and material

The datasets analyzed during the current study are downloaded from the NCBI database (updated in March 2021). All in-house scripts used in this study are available at https://github.com/HuangYue-Emma/PAH-biodegradation.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 12 January 2023 Accepted: 26 April 2023
Published online: 30 April 2023

References

1. Reddy CM, Arey JS, Seewald JS, Sylva SP, Lemkau KL, Nelson RK, et al. Composition and fate of gas and oil released to the water column during the Deepwater Horizon oil spill. Proc Natl Acad Sci U S A. 2012;109:20229–34.
2. Manoli E, Samara C. Polycyclic aromatic hydrocarbons in natural waters: sources, occurrence and analysis. TrAC Trends Anal Chem. 1999;18:417–28.
3. Zhang Y, Tao S. Global atmospheric emission inventory of polycyclic aromatic hydrocarbons (PAHs) for 2004. Atmos Environ. 2009;43:812–9.
4. Boström C-E, Gerde P, Hanberg A, Jernström B, Johansson C, Kyrklund T, et al. Cancer risk assessment, indicators, and guidelines for polycyclic aromatic hydrocarbons in the ambient air. Environ Health Perspect. 2002;110:451–88.
5. Kim KH, Jahan SA, Kabir E, Brown RJ. A review of airborne polycyclic aromatic hydrocarbons (PAHs) and their human health effects. Environ Int. 2013;60:71–80.
6. Hylland K. Polycyclic aromatic hydrocarbon (PAH) ecotoxicology in marine ecosystems. J Toxicol Environ Health, A. 2006;69:109–23.
7. González-Gaya B, Martínez-Varela A, Vila-Costa M, Casal P, Cerro-Gálvez E, Berrojalbiz N, et al. Biodegradation as an important sink of aromatic hydrocarbons in the oceans. Nat Geosci. 2019;12:119–25.
8. Wilcke W. Global patterns of polycyclic aromatic hydrocarbons (PAHs) in soil. Geoderma. 2007;141:157–66.
9. Simonich SL, Hites RA. Importance of vegetation in removing polycyclic aromatic hydrocarbons from the atmosphere. Nature. 1994;370:49–51.
10. Orbea A, Ortiz-Zarragoitia M, Solé M, Porte C, Cajaraville MP. Antioxidant enzymes and peroxisome proliferation in relation to contaminant body burdens of PAHs and PCBs in bivalve molluscs, crabs and fish from the Urdaibai and Plentzia estuaries (Bay of Biscay). Aquat Toxicol. 2002;58:75–98.
11. Frapiccini E, Annibaldi A, Betti M, Polidori P, Truzzi C, Marini M. Polycyclic aromatic hydrocarbon (PAH) accumulation in different common sole (Solea solea) tissues from the North Adriatic Sea peculiar impacted area. Mar Pollut Bull. 2018;137:61–8.

12. Weissenfels WD, Klewer H-J, Langhoff J. Adsorption of polycyclic aromatic hydrocarbons (PAHs) by soil particles: influence on biodegradability and biotoxicity. Appl Microbiol Biotechnol. 1992;36:689–96.
13. Ge L, Na G, Chen C-E, Li J, Ju M, Wang Y, et al. Aqueous photochemical degradation of hydroxylated PAHs: kinetics, pathways, and multivariate effects of main water constituents. Sci Total Environ. 2016;547:166–72.
14. Yuan S, Chang J, Yen J, Chang B-V. Biodegradation of phenanthrene in river sediment. Chemosphere. 2001;43:273–8.
15. Haritash AK, Kaushik CP. Biodegradation aspects of polycyclic aromatic hydrocarbons (PAHs): a review. J Hazard Mater. 2009;169:1–15.
16. Ghosal D, Ghosh S, Dutta TK, Ahn Y. Current state of knowledge in microbial degradation of polycyclic aromatic hydrocarbons (PAHs): a review. Front microbiol. 2016;7:1369.
17. Simon MJ, Osslund TD, Saunders R, Ensley BD, Suggs S, Harcourt A, et al. Sequences of genes encoding naphthalene dioxygenase in *Pseudomonas putida* strains G7 and NCIB 9816–4. Gene. 1993;127:31–7.
18. Kiyohara H, Torigoe S, Kaida N, Asaki T, Iida T, Hayashi H, et al. Cloning and characterization of a chromosomal gene cluster, pah, that encodes the upper pathway for phenanthrene and naphthalene utilization by *Pseudomonas putida* OUS82. J Bacteriol. 1994;176:2439–43.
19. Coppotelli BM, Ibarrolaza A, Dias RL, Del Panno MT, Berthe-Corti L, Morelli IS. Study of the degradation activity and the strategies to promote the bioavailability of phenanthrene by *Sphingomonas paucimobilis* strain 20006FA. Microb Ecol. 2010;59:266–76.
20. Shi T, Fredrickson JK, Balkwill DL. Biodegradation of polycyclic aromatic hydrocarbons by *Sphingomonas* strains isolated from the terrestrial subsurface. J Ind Microbiol Biotechnol. 2001;26:283–9.
21. Khan AA, Wang RF, Cao WW, Doerge DR, Wennerstrom D, Cerniglia CE. Molecular cloning, nucleotide sequence, and expression of genes encoding a polycyclic aromatic ring dioxygenase from Mycobacterium sp strain PYR-1. Appl Environ Microbiol. 2001;67:3577–85.
22. Miller CD, Hall K, Liang YN, Nieman K, Sorensen D, Issa B, et al. Isolation and characterization of polycyclic aromatic hydrocarbon-degrading *Mycobacterium* isolates from soil. Microb Ecol. 2004;48:230–8.
23. Grund E, Denecke B, Eichenlaub R. Naphthalene degradation via salicylate and gentisate by *Rhodococcus* sp strain B4. Appl Environ Microbiol. 1992;58:1874–7.
24. Kimura N, Urushigawa Y. Metabolism of dibenzo-p-dioxin and chlorinated dibenzo-p-dioxin by a gram-positive bacterium, *Rhodococcus opacus* SAO101. J Biosci Bioeng. 2001;92:138–43.
25. Laurie AD, Lloyd-Jones G. Conserved and hybrid *meta*-cleavage operons from PAH-degrading *Burkholderia* RP007. Biochem Biophys Res Commun. 1999;262:308–14.
26. Vacca DJ, Bleam WF, Hickey WJ. Isolation of soil bacteria adapted to degrade humic acid-sorbed phenanthrene. Appl Environ Microbiol. 2005;71:3797–805.
27. Steen AD, Crits-Christoph A, Carini P, DeAngelis KM, Fierer N, Lloyd KG, et al. High proportions of bacteria and archaea across most biomes remain uncultured. ISME J. 2019;13:3126–30.
28. Mallick S, Chakraborty J, Dutta TK. Role of oxygenases in guiding diverse metabolic pathways in the bacterial degradation of low-molecular-weight polycyclic aromatic hydrocarbons: a review. Crit Rev Microbiol. 2011;37:64–90.
29. Peng RH, Xiong AS, Xue Y, Fu XY, Gao F, Zhao W, et al. Microbial biodegradation of polyaromatic hydrocarbons. FEMS Microbiol Rev. 2008;32:927–55.
30. Seo JS, Keum YS, Li QX. Bacterial degradation of aromatic compounds. Int J Environ Res Public Health. 2009;6:278–309.
31. Kanaly RA, Harayama S. Advances in the field of high-molecular-weight polycyclic aromatic hydrocarbon biodegradation by bacteria. Microb Biotechnol. 2010;3:136–64.
32. Park JW, Crowley DE. Dynamic changes in *nahAc* gene copy numbers during degradation of naphthalene in PAH-contaminated soils. Appl Microbiol Biotechnol. 2006;72:1322–9.
33. Cebron A, Norini MP, Beguiristain T, Leyval C. Real-Time PCR quantification of PAH-ring hydroxylating dioxygenase (PAH-RHD$_\alpha$) genes from Gram positive and Gram negative bacteria in soil and sediment samples. J Microbiol Methods. 2008;73:148–59.
34. Liang C, Huang Y and Wang H. *pahE*, a functional marker gene for polycyclic aromatic hydrocarbon-degrading bacteria. Appl Environ Microbiol. 2019;85.
35. Moser R, Stahl U. Insights into the genetic diversity of initial dioxygenases from PAH-degrading bacteria. Appl Microbiol Biotechnol. 2001;55:609–18.
36. Fuenmayor SL, Wild M, Boyes AL, Williams PA. A gene cluster encoding steps in conversion of naphthalene to gentisate in *Pseudomonas* sp strain U2. J Bacteriol. 1998;180:2522–30.
37. Gomes NCM, Borges LR, Paranhos R, Pinto FN, Krögerrecklenfort E, Mendonça-Hagler LC, et al. Diversity of *ndo* genes in mangrove sediments exposed to different sources of polycyclic aromatic hydrocarbon pollution. Appl Environ Microbiol. 2007;73:7392–9.
38. Di Gennaro P, Terreni P, Masi G, Botti S, De Ferra F, Bestetti G. Identification and characterization of genes involved in naphthalene degradation in *Rhodococcus opacus* R7. Appl Microbiol Biotechnol. 2010;87:297–308.
39. Goyal AK, Zylstra GJ. Molecular cloning of novel genes for polycyclic aromatic hydrocarbon degradation from *Comamonas testosteroni* GZ39. Appl Environ Microbiol. 1996;62:230–6.
40. Kweon O, Kim S-J, Freeman JP, Song J, Baek S, Cerniglia CE. Substrate specificity and structural characteristics of the novel Rieske nonheme iron aromatic ring-hydroxylating oxygenases NidAB and NidA3B3 from *Mycobacterium vanbaalenii* PYR-1. MBio. 2010;1:e00135-e1110.
41. Krivobok S, Kuony S, Meyer C, Louwagie M, Willison JC, Jouanneau Y. Identification of pyrene-induced proteins in *Mycobacterium* sp. strain 6PY1: evidence for two ring-hydroxylating dioxygenases. J Bacteriol. 2003;185:3828–41.
42. Meckenstock RU, Safinowski M, Griebler C. Anaerobic degradation of polycyclic aromatic hydrocarbons. FEMS Microbiol Ecol. 2004;49:27–36.
43. Eddy SR. Profile hidden Markov models. Bioinformatics (Oxford, England). 1998;14:755–63.
44. Yin X, Jiang X-T, Chai B, Li L, Yang Y, Cole JR, et al. ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. Bioinformatics. 2018;34:2263–70.
45. Darling AE, Jospin G, Lowe E, Matsen FA IV, Bik HM, Eisen JA. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ. 2014;2:e243.
46. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42:D222–30.
47. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35:1547.
48. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:1–11.
49. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
50. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 2005;33:511–8.
51. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res. 2018;46:W200–4.
52. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007;23:127–8.
53. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2:18–22.
54. Ma Y, Wang L, Shao Z. *Pseudomonas*, the dominant polycyclic aromatic hydrocarbon-degrading bacteria isolated from Antarctic soils and the role of large plasmids in horizontal gene transfer. Environ Microbiol. 2006;8:455–65.
55. DeBruyn JM, Mead TJ, Sayler GS. Horizontal transfer of PAH catabolism genes in *Mycobacterium*: evidence from comparative genomics and isolated pyrene-degrading bacteria. Environ Sci Technol. 2012;46:99–106.
56. Herrick J, Stuart-Keil K, Ghiorse W, Madsen E. Natural horizontal transfer of a naphthalene dioxygenase gene between bacteria native to a coal tar-contaminated field site. Appl Environ Microbiol. 1997;63:2330–7.
57. Lloyd-Jones G, Laurie AD, Hunter DW, Fraser R. Analysis of catabolic genes for naphthalene and phenanthrene degradation in contaminated New Zealand soils. FEMS Microbiol Ecol. 1999;29:69–79.
58. LeVieux JA, Medellin B, Johnson WH Jr, Erwin K, Li W, Johnson IA, et al. Structural characterization of the hydratase-aldolases, NahE and PhdJ: implications for the specificity, catalysis, and *N*-acetylneuraminate lyase subgroup of the aldolase superfamily. Biochemistry. 2018;57:3524–36.

59. Kulakov LA, Chen S, Allen CC, Larkin MJ. Web-type evolution of *Rhodococcus* gene clusters associated with utilization of naphthalene. Appl Environ Microbiol. 2005;71:1754–64.
60. Grimm AC, Harwood CS. Chemotaxis of *Pseudomonas* spp. to the polyaromatic hydrocarbon naphthalene. Appl Environ Microbiol. 1997;63:4111–5.
61. Ochou M, Saito M, Kurusu Y. Characterization of two compatible small plasmids from *Sphingobium yanoikuyae*. Biosci, Biotechnol, Biochem. 2008;72:1130–3.
62. Zhao Q, Yue S, Bilal M, Hu H, Wang W, Zhang X. Comparative genomic analysis of 26 *Sphingomonas* and *Sphingobium* strains: dissemination of bioremediation capabilities, biodegradation potential and horizontal gene transfer. Sci Total Environ. 2017;609:1238–47.
63. Dunn N, Gunsalus I. Transmissible plasmid coding early enzymes of naphthalene oxidation in *Pseudomonas putida*. J Bacteriol. 1973;114:974–9.
64. Cho J-C, Kim S-J. Detection of mega plasmid from polycyclic aromatic hydrocarbon-degrading *Sphingomonas* sp. strain KS14. J Mol Microbiol Biotechnol. 2001;3:503–6.
65. Romine MF, Stillwell LC, Wong K-K, Thurston SJ, Sisk EC, Sensen C, et al. Complete sequence of a 184-kilobase catabolic plasmid from *Sphingomonas aromaticivorans* F199. J Bacteriol. 1999;181:1585–602.
66. Johnsen AR, Wick LY, Harms H. Principles of microbial PAH-degradation in soil. Environ Pollut. 2005;133:71–84.
67. Wang X-T, Shan J-J, Li X-Z, Lin W, Xiu J-L, Li D-A, et al. *Tepidiphilus olei* sp. nov, isolated from the production water of a water-flooded oil reservoir in PR China. Int J Syst Evol Microbiol. 2020;70:4364–71.
68. Blümel S, Knackmuss H-J, Stolz A. Molecular cloning and characterization of the gene coding for the aerobic azoreductase from *Xenophilus azovorans* KF46F. Appl Environ Microbiol. 2002;68:3948–55.
69. Dong C, Bai X, Sheng H, Jiao L, Zhou H, Shao Z. Distribution of PAHs and the PAH-degrading bacteria in the deep-sea sediments of the high-latitude Arctic Ocean. Biogeosciences. 2015;12:2163–77.
70. Reyes-Sosa MB, Apodaca-Hernández JE, Arena-Ortiz ML. Bioprospecting for microbes with potential hydrocarbon remediation activity on the northwest coast of the Yucatan Peninsula, Mexico, using DNA sequencing. Sci Total Environ. 2018;642:1060–74.
71. Erdoğmuş SF, Mutlu B, Korcan SE, Güven K, Konuk M. Aromatic hydrocarbon degradation by halophilic archaea isolated from Çamaltı Saltern, Turkey. Water Air Soil Pollut. 2013;224:1–9.
72. Al-Mailem D, Sorkhoh N, Al-Awadhi H, Eliyas M, Radwan S. Biodegradation of crude oil and pure hydrocarbons by extreme halophilic archaea from hypersaline coasts of the Arabian Gulf. Extremophiles. 2010;14:321–8.
73. Fairley D, Wang G, Rensing C, Pepper I, Larkin M. Expression of gentisate 1, 2-dioxygenase (*gdoA*) genes involved in aromatic degradation in two haloarchaeal genera. Appl Microbiol Biotechnol. 2006;73:691–5.
74. Ghosh A, Fassnacht FE, Joshi PK, Koch B. A framework for mapping tree species combining hyperspectral and LiDAR data: Role of selected classifiers and sensor across three spatial scales. Int J Appl Earth Obs Geoinf. 2014;26:49–63.
75. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics. 2009;10:1–16.
76. Han H, Guo X and Yu H. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In: 2016 7th IEEE international conference on software engineering and service science (ICSESS): 2016. IEEE. pp. 219–224.
77. Shivani Y, Subhash Y, Sasikala C and Ramana CV. Description of '*Candidatus Marispirochaeta associata*'and reclassification of *Spirochaeta bajacaliforniensis*, *Spirochaeta smaragdinae* and *Spirochaeta sinaica* to a new genus *Sediminispirochaeta* gen. nov. as *Sediminispirochaeta bajacaliforniensis* comb. nov., *Sediminispirochaeta smaragdinae* comb. nov. and *Sediminispirochaeta sinaica* comb. nov. Int J Syst Evol Microbiol. 2016;66:5485–5492.
78. Wu D, Raymond J, Wu M, Chatterji S, Ren Q, Graham JE, et al. Complete genome sequence of the aerobic CO-oxidizing thermophile *Thermomicrobium roseum*. PLoS ONE. 2009;4: e4207.
79. Olson ER. Influence of pH on bacterial gene expression. Mol Microbiol. 1993;8:5–14.
80. Bassler BL. How bacteria talk to each other: regulation of gene expression by quorum sensing. Curr Opin Microbiol. 1999;2:582–7.
81. Segura A, Hernandez-Sanchez V, Marques S, Molina L. Insights in the regulation of the degradation of PAHs in *Novosphingobium* sp. HR1a and utilization of this regulatory system as a tool for the detection of PAHs. Sci Total Environ. 2017;590–591:381–93.
82. Schell MA. Transcriptional control of the *nah* and *sal* hydrocarbon-degradation operons by the *nahR* gene product. Gene. 1985;36:301–9.
83. Jones RM, Britt-Compton B, Williams PA. The naphthalene catabolic (nag) genes of *Ralstonia* sp. strain U2 are an operon that is regulated by NagR, a LysR-type transcriptional regulator. J Bacteriol. 2003;185:5847–53.
84. Larkin MJ, Kulakov LA, Allen CC. Biodegradation and *Rhodococcus*–masters of catabolic versatility. Curr Opin Biotechnol. 2005;16:282–90.
85. Yen K, Gunsalus IC. Regulation of naphthalene catabolic genes of plasmid NAH7. J Bacteriol. 1985;162:1008–13.
86. Zhang Z, Sun J, Guo H, Wang C, Fang T, Rogers MJ, et al. Anaerobic biodegradation of phenanthrene by a newly isolated nitrate-dependent *Achromobacter denitrificans* strain PheN1 and exploration of the biotransformation processes by metabolite and genome analyses. Environ Microbiol. 2021;23:908–23.
87. Kummel S, Herbst FA, Bahr A, Duarte M, Pieper DH, Jehmlich N, et al. Anaerobic naphthalene degradation by sulfate-reducing *Desulfobacteraceae* from various anoxic aquifers. FEMS Microbiol Ecol. 2015;91.
88. Kleemann R, Meckenstock RU. Anaerobic naphthalene degradation by Gram-positive, iron-reducing bacteria. FEMS Microbiol Ecol. 2011;78:488–96.
89. Kadri T, Rouissi T, Brar SK, Cledon M, Sarma S, Verma M. Biodegradation of polycyclic aromatic hydrocarbons (PAHs) by fungal enzymes: A review. J Environ Sci. 2017;51:52–74.
90. Semple KT, Cain RB, Schmidt S. Biodegradation of aromatic compounds by microalgae. FEMS Microbiol Lett. 1999;170:291–300.
91. Chaumeil P-A, Mussig AJ, Hugenholtz P and Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Oxford University Press; 2020.

Publisher's Note