Genome-resolved analyses show an extensive diversification in key aerobic hydrocarbon-degrading enzymes across bacteria and archaea

Maryam Rezaei Somee[1], Mohammad Ali Amoozegar[1], Seyed Mohammad Mehdi Dastgheib[2], Mahmoud Shavandi[2], Leila Ghanbari Maman[3], Stefan Bertilsson[4], Maliheh Mehrshad[4*]

[1]Extremophile Laboratory, Department of Microbiology, School of Biology, College of Science, University of Tehran, Tehran, Iran

[2]Biotechnology Research group, Research Institute of Petroleum Industry, Tehran, Iran

[3]Laboratory of Complex Biological Systems and Bioinformatics (CBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

[4]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences (SLU), Box 7050, SE75007 Uppsala, Sweden

Corresponding author: maliheh.mehrshad@slu.se

Abstract

Hydrocarbons (HCs) are organic compounds composed solely of carbon and hydrogen. They mainly accumulate in oil reservoirs, but aromatic HCs can also have other sources and are widely distributed in the biosphere. Our perception of pathways for biotic degradation of major HCs and genetic information of key enzymes in these bioconversion processes have mainly been based on cultured microbes and are biased by uneven taxonomic representation. Here we use Annotree to provide a gene-centric view of aerobic degradation of aliphatic and aromatic HCs in a total of 23446 genomes from 123 bacterial and 14 archaeal phyla. Apart from the widespread genetic potential for HC degradation in *Proteobacteria*, *Actinobacteriota*, *Bacteroidota*, and *Firmicutes*, genomes from an additional 18 bacterial and 3 archaeal phyla also hosted key HC degrading enzymes. Among these, such degradation potential has not been previously reported for representatives in the phyla UBA8248, Tectomicrobia, SAR324, and Eremiobacterota. While genomes containing full pathways for complete degradation of HCs were only detected in *Proteobacteria* and *Actinobacteriota*, other lineages capable of mediating such key steps could partner with representatives with truncated HC degradation pathways and collaboratively drive the process. Phylogeny reconstruction shows that the reservoir of key aerobic hydrocarbon-degrading enzymes in Bacteria and Archaea undergoes extensive diversification via gene duplication and horizontal gene transfer. This diversification could potentially enable microbes to rapidly adapt to novel and manufactured HCs that reach the environment.

Introduction

According to the biogenic (organic) theory, petroleum hydrocarbons originate from ancient remains of detrital matter buried and diagenetically modified in marine or freshwater sediments. This organic matter is then gradually converted to petroleum compounds enriched in aromatic and aliphatic hydrocarbons (HCs) via the sequential activity of aerobic and anaerobic microorganisms [1][2][3]. In addition to their role in the formation of oil HCs, microbes play a crucial role in the biological integration of these HCs into the actively cycled carbon pool [4]. Microbial HC degradation occurs through a cascade of enzymatic reactions in three main steps: (i) activation and attack of the HC-bond, producing signature intermediate compounds, (ii) conversion of signature degradation intermediates to central cell metabolites, followed by (iii) mineralization to CO2. Microorganisms must overcome and break the stability and energy in carbon-hydrogen bonds in order to degrade HCs. Since HCs are structurally diverse, a plethora of enzymes are involved in their activation and degradation, and consequently, the energy that needs to be invested in the initial degradation varies. Various microorganisms can degrade different HCs according to their enzymatic repertoire and available energy [5]. Microorganisms have evolved to degrade different HCs under both aerobic and anaerobic conditions. However, biodegradation typically occurs much faster under aerobic conditions, in part due to the availability of thermodynamically favorable electron acceptors that leads to higher energy yield [6], but also because of the action of some HC-degrading enzymes requires oxygen as substrate or cofactor. Similar to all biological pathways, rate-limiting key enzymes drive the main steps of HC degradation.

Under aerobic conditions, oxygenase enzymes initiate the degradation of different aliphatic or aromatic compounds by adding one (mono-oxygenase) or two (di-oxygenase) oxygen molecules. Saturated aliphatic compounds such as alkane and cycloalkane (studied here) are in this process converted to their corresponding carboxylic acid. Catechol/gentisate derivatives are intermediate compounds during aerobic degradation of aromatic mono- and polycyclic HCs. They are then de-aromatized via subsequent meta/ortho cleavage. Intermediate compounds produced during the degradation of aliphatic and aromatic HCs converge to the B-oxidation and tricarboxylic acid (TCA) cycle [7]. While enzymes involved in the downstream part of the degradation process are widespread across living cells shared by many metabolic pathways, the mono/di-oxygenase enzymes catalyzing the first hydroxylation of aliphatic/aromatic compounds are crucial for the initial step in the HC degradation process and likely rate-limiting. Accordingly, microorganisms carrying the enzymes for such initial degradation will be rate-controlling drivers of HC degradation.

The capacity of microbial isolates to metabolically degrade oil HCs have been frequently studied [8–11]. However, our knowledge has until recently been mainly limited to cultivated microorganisms. The present study provides a systemic and genome-resolved view of hydrocarbon degradation capability in the growing database of archaeal and bacterial genomes. To provide this extensive view, we compiled a database of enzymes involved in the aerobic degradation pathway of aliphatic and aromatic HCs (toluene, phenol, xylene, benzene, biphenyl, naphthalene). We then explored the distribution of these enzymes in 24692 publicly available archaeal (n=1246) and bacterial (n=23446) genomes via AnnTree [12] and manually confirmed all annotations. We focused on the microbial genomes containing enzymes for complete/near complete degradation of specific HCs and suggest that lineages with the great genetic potential to degrade a broad range of HC compounds can be exploited for bioremediation purposes. We also reconstructed the phylogenetic relationships of the recovered key HC degradation enzymes to investigate their evolution and explore the potential role of horizontal gene transfer. Several microorganisms contain multiple copies of key HC degrading genes across their genome. We thus explored whether these copies are likely to have been acquired through HGT or if they are likely to be paralogs. Having a genome-resolved view we also studied ecological strategies of these microbes to see whether all critical HC degraders adopt similar growth strategy in terms of the canonical r and k-strategists.

Results and discussion

HC degradation across domain Bacteria. alkB/M and almA/ladA genes are alkane mono-oxygenases that initiate the degradation of short (C5–C15) and long-chain alkanes (>C15), respectively. The alkB/M is rubredoxin-dependent, while almA and ladA are flavin-dependent mono-oxygenases. The genes pheA (phenol), xylM (xylene), xylX, todC1, and tmoA (benzene/toluene) for monoaromatic and bphA1(biphenyl), ndoB (naphthalene/phenanthrene) for polyaromatic compounds code for catalytic domains of ring hydroxylating oxygenases (RHOs) that add -OH group(s) to compounds undergoing degradation (**Supplementary Figure S1**). We explored the distribution of these enzymes and associated degradation pathways in a total of 23446 representatives out of 143512 bacterial genomes available in release 89 of the GTDB database that has been annotated via Annotree [12]. These annotated genomes are dominated by representatives of phyla *Proteobacteria* (32.5%), *Actinobacteriota* (13.3%), *Bacteroidota* (12.13%), and *Firmicutes* (8.01%) (**Supplementary Figure S2** and **Supplementary Table S4**). Among the 123 represented bacterial phyla, 58 phyla had ≤ five genomes available per phylum and combined only represented 0.57% of the explored genomes. To avoid misinterpretations due to this uneven taxonomic distribution of representative genomes, we explored the contribution of members of

each phylum in the HC degradation process by showing what proportion of microbes containing each HC degrading enzyme exist in each phylum (panel A of **Supplementary Figures S3-S6**). We also analyze the percentage of members of each phylum containing each HC degrading enzyme to ensure that we consider the contributions of underrepresented phyla in the HC degradation (panel B of **Supplementary Figures S3-S6**).

As expected, representatives of the phylum *Proteobacteria* (*Pseudomonadales* and *Burkholderiales* orders) presented the highest abundance of aliphatic and aromatic HC degrading enzymes, followed by *Actinobacteriota* and *Bacteroidota* for aliphatic and *Actinobacteriota* and *Firmicutes* for aromatic HC degrading enzymes (**Supplementary Figures S3** and **S4**, panel A).

Underrepresented phyla remain mainly uncultured and are notably underexplored for metabolic potential (58 of 123 phyla, n=131 genomes). Our analyses revealed that representatives of these taxa contain HC degrading enzymes involved in both the initiation and downstream steps of HC degradation processes. For example, phyla *Tectomicrobia* (*Entotheonella*), *Binatota*, *Firmicutes*_K, and *Firmicutes*_E contained mono-aromatic HC degradation enzymes (**Figure 2**). In addition to these phyla, we annotated enzymes involved in the degradation of aliphatic HC in representatives of phyla SAR324, *Eremiobacterota* (*Baltobacterales*), *Bdellovibrionota*_B, and *Chloroflexota*_B (**Figure 1**).

Other enzymes in the degradation pathways beyond the key genes for the initial degradation (**Supplementary Table S1**) are typically involved in several degradation pathways and are broadly distributed accordingly. As an example, the process of converting catechol to non-aromatic compounds with further conversion to intermediates of the TCA cycle (e.g., acetaldehyde and pyruvate) (**Supplementary Figure S1**) is shared among degradation pathways of xylene, naphthalene/phenanthrene, and phenol (blue color in **Figure 2**). These ring-cleavage enzymes are also involved in the degradation of aromatic amino acids. Our analysis showed that representatives of phyla Firmicutes (mainly from the orders *Bacillales* and *Staphylococcales*), *Firmicutes*_I, *Firmicutes*_K, *Firmicutes*_E, *Firmicutes*_G, F*irmicutes*_H, *Eremiobacterota*, *Deinococcota*, *Chloroflexota*, *Campylobacterota*, *Myxococcota* and *Bdellovibrionota* play a significant role in this part of HC degradation process (blue color in **Figure 2**).

Distribution of key genes involved in the degradation of Alkanes. At lower taxonomic rank, the alkB/M and ladA genes were differently distributed across members of phyla *Gammaproteobacteria*, *Alphaproteobacteria*, and *Actinobacteriota*, hinting at their capacity for degrading hydrocarbons of variable chain length. Altogether 2089 genomes in orders *Mycobacteriales* (23.95%), *Rhodobacterales* (20.46%), *Pseudomonadales* (17.13%), *Flavobacteriales* (8.3%), *Burkholderiales* (6.16%), *Cytophagales*

(3.66%), *Propionibacteriales* (2.47%), *Rhizobiales* (1.89%), and *Chitinophagales* (1.81%) contained alkB/M genes, while ladA was present in 2154 genomes from *Pseudomonadales* (21.05%), *Rhizobiales* (16.27%), *Burkholderiales* (14.44%), *Actinomycetales* (13.44%), *Mycobacteriales* (13.05%), *Bacillales* (4.74%), *Enterobacterales* (3.7%), *Acetobacterales* (2.31%), *Streptomycetales* (1.91%) (**Figures 3** and **4**, panel B, **Supplementary Table S6**).

An indirect role of *Cyanobacteria* in HC degradation, especially in microbial mats, has been previously reported. These primary producers often have the nitrogen-fixing ability and can fuel and promote aerobic and anaerobic sulfate/nitrate-reducing HC degrading microorganisms in microbial mats [13]. There are also reports of a minor role of some *Cyanobacteria* members like *Phormidium*, *Nostoc*, *Aphanothece*, *Synechocystis*, *Anabaena*, *Oscillatoria*, *Plectonema,* and *Aphanocapsa* in direct HC degradation [14][15]. In this study, we detected the presence of long-chain alkane degrading genes, ladA, in different members of *Cyanobacteria* with 0.31 and 12.54% of genomes in this phylum containing ladA (in *Elainella saxicola*, *Nodosilinea sp000763385*) and almA genes (in *Synechococcales*, *Cyanobacteriales*, *Elainellales*, *Phormidesmiales*, *Thermosynechococcales*, *Gloeobacterales*, *Obscuribacterales*), respectively.

Phylogenetic reconstruction of recovered alkB/M and ladA genes grouped them into five and nine main clades, respectively (**Figures 3** and **4**, panel A). The branching pattern of these clades partially followed the taxonomic signal of the genomes they were retrieved from, specifically for most dominant phyla; however, some branches also contained alkB/M and ladA genes originating from different and distantly related phyla. The placement of phylogenetically diverse groups in one branch is likely to result from the horizontal transfer of these genes between microbial taxa [16]. Additionally, apart from the chromosomal type, both alkB/M and ladA genes have previously been reported to be located on plasmids (OCT and pLW1071), corroborating their potential for horizontal transfer. For instance, there are reports on the intraspecies transfer of alkB/M among *Pseudomonas* members [17]. Placement of rare microbial groups harboring ladA gene among clusters V-IX further suggests a prominent role of Actinobacteriota and Firmicutes members in expanding the distribution of this gene (**Figure 4**).

We also detected several genomes with multiple copies of the alkB gene that were not necessarily branching together in the reconstructed alkB phylogeny, hinting at the probability of either gene duplication, paralogue occurrence or HGT. Examples of these genomes with several copies of alkB/M are *Polycyclovorans algicola* (10), *Nevskia ramose* (7), *Zhongshania aliphaticivorans* (7), *Solimonas aquatic* (7), *Immundisolibacter cernigliae* (6), and *Rhodococcus qingshengii* (6). Multiple copies have also been

detected in representatives of the genera *Nocardia*, *Rhodococcus,* and *Alcanivorax* (**Supplementary Table S6**).

Furthermore, the ladA gene was also detected in *Mycolicibacterium dioxanotrophicus*, *Cryobacterium*_A sp003065485, *Kineococcus rhizosphaerae*, *Microbacterium* sp003248605, *Paenibacillus*_S sp001956045, *Pararhizobium* polonicum, *Mycolicibacterium septicum*, and *Microbacterium* sp000799385 with six copies in each genome. Several examples were also present in genera *Pseudomonas*_E, *Bradyrhizobioum*, *Rhizobioum*, and *Paraburkholderia,* which had more than one copy (904 genomes)(**Supplementary Table S6**).

The presence of multiple copies of alkane hydroxylase genes has been hypothesized to enable cells to use an expanded range of n-alkanes or to adapt to different environmental conditions. However, the exact evolutionary rationale has not yet been established [18][19]. To evaluate this hypothesis, we compared different sequences of each gene in an individual genome (mentioned above for ladA and alkB) using BLAST (**Supplementary Table S7**). The results showed that the identity of multiple gene copies in a single genome was in the range of 30 to 70 percent, while they are still predicted to have the same function. This further supports the hypothesis that these genes originated from different sources and were transferred horizontally.

Distribution of key genes involved in the degradation of ring hydroxylating oxygenases (RHOs). Genomes containing RHOs (2761 genomes, 16 phyla) present an overall lower phylogenetic diversity than alkane mono-oxygenases (4669 genomes, 21 phyla for both alkB/M and ladA). In general, alkB/M and ladA enzymes consist of FA_desaturase (PF00487) and Bac_luciferase-like mono-oxygenase (PF00296) domains, respectively (**Supplementary Table S5**). They act non-specifically on a wide range of alkanes of different chain lengths. Therefore, they are likely to be more widespread in genomes, especially because alkane compounds do not exclusively originate from petroleum. For instance, in pristine marine ecosystems, primary producers such as *Cyanobacteria* can release long chain-length aliphatic compounds (e.g., pentadecane, heptadecane). Alkane-producing *Cyanobacteria* include prominent and globally abundant genera such as *Prochlorococcus* and *Synechococcus*. Therefore, marine microorganisms are broadly exposed to aliphatic compounds with different chain lengths, even in environments without oil spills or industrial influence. This can explain why marine ecosystems host a plethora of hydrocarbonoclastic bacteria [20][21].

Enzymes xylX, ndoB, bphA1, and todC1 are composed of two pfam domains, PF00355 (Rieske center) and PF00848 (Ring_hydroxyl_A). These common domains impact the branching in the phylogenetic tree and lead to the neighboring branching of these enzymes (**Figure 5**).

RHO enzymes are predominantly present in *Burkholderiales*, *Pseudomonadales, Sphingomonadales*, *Caulobacterales,* and *Nevskiales* orders of the phylum *Proteobacteria* (35 different Proteobacterial orders) (**Figure 5**, B part). However, a significant number of pheA and, to a lesser degree, xylX and tmoA enzymes were also present in *Actinobacteriota* phylum (9 different Actinobacteriotal orders) (**Figure 5**, B part).

Sphingomonadales are prominent bacteria in the rhizosphere and are also abundant in littoral zones of inland waters. Accordingly, we suggest that these bacteria may have evolved a capacity to degrade different aromatic compounds in response to the high concentrations of aromatic secondary metabolites typically seen in the plant rhizosphere. Additionally, *Sphingomonadales* are known for their large plasmids with intraspecies transmission [22].

Among all investigated RHO genes, the highest phylogenetic diversity was observed in tmoA (208 genomes in 12 phyla and 38 orders) and xylX (1486 genomes in 9 phyla and 38 orders) genes (**Figure 5**, B part). In the case of tmoA gene, it might be due to the wide range of HC compounds susceptible to this enzyme (e.g., benzenes, some PAHs, and alkenes)[23][24]. Therefore, more diverse genera harbor tmoA gene and can degrade different types of HCs.

Underrepresented microbial groups with a limited number of RHO genes also featured tmoA, xylX, and pheA genes. *Myxococcota*, *Acidobacteriota*, *Chloroflexota*, *Firmicutes*_I,E,K, and *Cyanobacteria* with tmoA gene were clustered separately, reflecting their distinct protein sequence and the lower possibility of HGT among these groups. For xylX, *Eremiobacterota* affiliated genes were placed together with genes from *Gammaproteobacteria,* and *Tectomicrobia*, *Binatota*, *Chloroflexota,* and *Firmicutes*_I were placed in separate branches near *Actinobacteriota*. In addition, *Acidobacteriota*, *Eremiobacterota,* and *Campylobacteria* with pheA gene were nested within *Alphaproteobacteria* members. The phylogeny of RHO genes was also more consistent with taxonomy than the phylogeny of alkB/M and ladA.

Bionatota, a recently described phylum shown to be efficient in HC degradation, harbored todC1, bphA1 (in Binatales order), and xylX (Bin18, *Binatales*) genes from RHOs and ladA (in Bin18) from alkane hydroxylases. Representatives of this phyla have been reported to play a role in methane and alkane metabolism [25]. However, we also noted the further potential of *Binatales* and Bin18 orders of this phylum in aromatic HC degradation.

RHOs can be located either on the chromosome or plasmid, depending on the organism. For instance, todC1, bphA1, and tmoA genes were reported to be on the chromosome [26], while in another study, they were detected on a plasmid [27]. Other RHOs, including xylX, xylM, pheA, and ndoB have mainly been reported to be hosted by plasmids [26][24].

Multiple copies of RHO genes in one genome were detected for xylX and pheA. *Immundisolibacter cernigliae* surprisingly contained 21 variants of xylX. This genome also had six copies of alkB/M and was isolated from a PAH-contaminated site [28]. The high HC degradation potential of other members of this genus has also been reported in the marine ecosystem [29][30]. *Rugosibacter aromaticivorans* (containing 5, 2 and 2 copies of xylX, ndoB, and tmoA genes, respectively), *Pseudoxanthomonas*_A spadix_B (with 4, 2 and 2 copies of xylX, todC1 and bphA1 genes, respectively), *Thauera* sp002354895 (4), *Pigmentiphaga* sp002188635 (4) are other examples of genomes that have multiple copies of the xylX gene. Although xylX gene was detected in *Actinobacteriota*, multiple copies in a genome were seen only among the *Proteobacteria* phylum.

The BLAST identity among variants of the xylX gene in *Immundisolibacter cernigliae* ranged between 35 to 81 percent. Three sequences of these 21 xylX copies (xylX 18, 19, and 22, in **Supplementary Figure S7**) showed higher BLAST identity with the xylX gene of the *Rugosibacter* genus than other copies in the *Immundisolibacter cernigliae* genome itself (**Supplementary Table S7** and **Supplementary Figure S7**). Several xylX copies of *I. cernigliae* (10, 11, 13, and 15) had more edges than others in the network, and their interactions (**Supplementary Figure S7, highlighted in red**) represent their similarity with xylX copies of *Caballeronia*, *Sphingobium*, and *Pseudoxanthomonas*, *Pseudomonas,* and *Thauera* genera. In addition, xylX 5 and 7 of *Immundisolibacter* had almost similar blast identity with *Pigmentiphaga* genus and other xylX copies in *I. cernigliae*. This suggests that multiple copies of the xylX gene in *I. cernigliae* potentially originated from horizontal transfer.

On the other hand, *Glutamicibacter mysorens* (4), *Enteractinococcus helveticum* (4), and many other genomes from the *Castellaniella*, *Kocuria*, and *Halomonas* genera, had several pheA copies in their individual genomes. To a lesser degree, tmoA gene was present in multiple copies in *Pseudonocardia dioxanivorans* (4), *Rhodococcus sp003130705* (3), *Amycolatopsis rubida* (3) and *Zavarzinia compransoris_A* (3) genera.

While bphA1 and todC1 have different KO identifiers (**Supplementary Table S1**), our manual checks showed that they had the same conserved domain based on NCBI CD-Search [31]. We kept both annotations for cases where one gene was annotated with both KO identifiers. Previous studies also report similar homology and substrate specificity between todC1 and bphA1 [27].

xylM, as one of the enzymes mediating the initial steps in toluene/xylene degradation, showed the lowest abundance and phylogenetic diversity (27 genomes in 1 phylum and 6 orders). Toluene/benzene can generally be degraded through different routes and three of the most prevalent approaches were studied here. xylX, todC1, and tmoA are the initial oxygenase enzymes of these three pathways. They

255 are diverse in starting the degradation and composed of different domains, while downstream
256 degradation converges to catechol derivatives as intermediates. xylM can also initiate toluene
257 degradation in addition to xylene. xylX then converts produced benzoate to catechol. Therefore, while
258 we report a lower diversity of genomes harboring xylM, there are alternative degradation pathways in
259 different microorganisms that can degrade the same compound.
260 As the number of rings in aromatic compounds increases, the number and diversity of microbial groups
261 capable of degrading them decreases, and microbial groups with ndoB (naphthalene 1,2-dioxygenase)
262 accordingly showed the lowest abundance after xylM gene. The genomes hosting ndoB had limited
263 phylogenetic diversity (35 genomes in 1 phylum and 6 orders) and were found mainly in representatives
264 of *Alphaproteobacteria* (*Sphingomonadales* (17) and *Caulobacterales* (2)) and *Gammaproteobacteria*
265 (*Pseudomonadales* (5), *Burkholderiales* (1), *Nevskiales* (1)).
266 **Ecological strategy of HC degrading bacteria.** Microorganisms are broadly divided into two main
267 functional growth categories, i.e., oligotrophic/slow-growing/k-strategist or copiotrophic/fast-
268 growing/r-strategist. These ecological strategies are associated with the genome size that, in turn,
269 directly correlates with the GC content [32]. To get further insights into the ecological strategies of
270 organisms that feature HC degrading genes, we compared the distribution of GC content and estimated
271 genome size. This analysis revealed that HC degrading genes were present in genomes with a broad
272 genome size range (1.34 to 16.9 Mb) and GC content (26.9 to 76.6 %) (**Supplementary Figure S8**, data
273 available in **Supplementary Table S8**). Genomes with GC percent equal to or lower than 30% mainly had
274 alkB gene and belonged to representatives of the Flavobacteriales order (genome sizes in the range of
275 1.4 to 4.2 Mb). The largest genome studied here, *Minicystis rosea* from the phylum Myxococcota
276 (genomes size of 16.9 Mb), also contained alkB. The alkB gene of *Minicystis rosea* phylogenetically
277 clustered together with homologs from Gammaproteobacteria representatives (*Immundisolibacter* and
278 *Cycloclasticus* genera) (**Figure 3**). The large genome size of *Minicystis rosea* and its alkB gene placement
279 together with the Gammaproteobacteria in the reconstructed phylogeny suggests horizontal transfer for
280 this gene to *Minicystis rosea*. These analyses suggest that HC degradation ability is present in both k-
281 strategist and r-strategists microorganisms. Earlier studies have shown that r-strategist serves as the
282 principal HC degraders after oil spills and at other point sources of pollution in marine environments
283 [33–36]. Indeed, most obligate hydrocarbonoclastic bacteria are r-strategists (Proteobacteria domain)
284 and are mainly reported to be isolated from marine samples [37]. This group is adapted to live in
285 oligotrophic environments with transient nutrient inputs and rapid consumption of substrates upon
286 episodic inputs by means of fast growth and population expansion [38]. In contrast, reports of oil-

polluted soil samples suggest a predominance of k-strategists, especially in the harsh conditions (High concentration of HC, soil dryness, etc.) commonly seen in many such soil environments [39–41]. Hosting multiple copies of genes coding for HC degrading enzymes seems to be a shared feature in both r- and k-strategists with small and large genome sizes alike and appears to be a universal evolutionary strategy for HC degradation.

Genome-level analysis of HC degradation. Microorganisms are known to use division of labor or mutualistic interactions to perform HC degradation in the environment [42][43]. However, 92 genomes (less than 0.5%) of 23446 investigated bacterial genomes do in fact contain all the enzymes required to degrade at least one HC compound completely. These 92 genomes all belong to *Actinobacteriota* (n=25) and *Proteobacteria* (n=67)(**Figure 6**).

Microorganisms have evolved two pathways for naphthalene degradation that involve the production of either catechol or gentisate as aromatic degradation intermediate (**Supplementary Figure S1**). Catechol can in turn, be further degraded via meta- or -ortho cleavage. Several microorganisms, including *Novosphingobium naphthalenivorans*, *Pseudomonas*_E *fluorescens*_AQ, *Pseudomonas*_E *frederiksbergensis*_E, and *Herbaspirillum* sp000577615, feature both of the mentioned pathways and even have the ability to perform ortho and meta cleavage simultaneously (**Figure 6**).

Moreover, *Cupriavidus pauculus*_A (long-chain alkanes and also biphenyl), *Cycloclasticus* sp002700385 and *Paraburkholderia*_B *oxyphila* (*Cycloalkane* and xylene/benzene), *Pigmentiphaga* sp002188465 (Cycloalkane and phenol), *Rhodococcus* sp003130705, *Burkholderia puraquae*, and *Paraburkholderia*_B *mimosarum* (Toluene and biphenyl) can degrade more than one HC compound autonomously (**Figure 6**). Members of Burkholderiales were able to degrade even more diverse compounds individually, while *Actinobacteriota* representatives mainly contribute to the degradation of aliphatic compounds. This ability was also apparent in **Figures 1, 3,** and **4**. The potential for autonomous HC degradation wasn't detected in genomes of more rare bacterial phyla. Moreover, none of the archaeal genomes investigated in this study contained all genes for the complete degradation of HCs.

HC degradation across domain archaea. Generally, HC degradation ability seems to be less prevalent among archaea as compared to bacteria. The phylum *Halobacterota* had the highest proportion of enzymes involved in the degradation of both aliphatic (n=14 enzymes of aliphatic degradation pathway) and aromatic (n=25 enzymes of aromatic degradation pathway) compounds among the studied archaea (**Supplementary Figure S9**). The alkB enzyme, responsible for short-chain alkane degradation, was detected in two copies in a single member of the phylum Nanoarchaeota (ARS21 sp002686215). This gene was clustered together with alkB identified in Gammaproteobacteria

representatives (GCA-002705445 order) (**Figure 3**). Genes needed to initiate degradation of long-chain alkanes and cyclododecane/cyclohexane as well as cyclopentane degradation via ladA and cddA/chnB genes were more prevalent among *Halobacterota* representatives (75 genomes in 7 families; *Haloferacaceae*, *Haloarculaceae*, *Natrialbaceae*, *Halococcaceae*, *Halalkalicoccaceae*, *Haloadaptaceae*, and *Halobacteriaceae*) (**Figures 4 and Supplementary Figure S9**). Among investigated RHOs, only tmoA that initiates toluene degradation was present in 5 Sufolobales and 2 Thermoproteales genomes of the phylum Crenarchaeota (**Figure 5**). Detected archaeal tmoA and ladA genes branched separately from bacteria in the phylogenetic trees (**Figures 4 and 5**). Apart from alkB, gene duplications were present in several genomes for both tmoA (*Sulfolobus* and *Acidianus* genera) and ladA (*Halopenitus persicus* and *Halopenitus malekzadehii*).

Key enzymes needed to initiate HC degradation were rarely present in archaea (**Figures 3, 4, and 5**), indicating that Archaea might not play a significant role in the typically rate-limiting initial degradation of HCs. However, several studies report the ability of halophilic archaeal isolates (e.g., *Halorubrum* sp., *Halobacterium* sp., *Haloferax* sp., *Haloarcula* sp.) to degrade both aliphatic (n-alkanes with chain lengths up to C18 and longer) and aromatic (e.g., naphthalene, phenanthrene, benzene, toluene and *p*-hydroxybenzoic acid) HCs and use them as their sole source of carbon [44–46]. This may imply that archaea carry alternative and hitherto unknown enzymes for triggering HC degradation. However, there is no complete genome information available for the mentioned isolates to screen them for the presence of alternative degrading enzymes [11]. The *Haloferax* sp., capable of using a wide range of HCs as its sole source of carbon, present in the AnnoTree database (RS_GCF_000025685.1), contained none of the key degrading genes. The AnnoTree website chooses representative genomes having completeness of higher than 90%, whcih reduces the likelihood of incompleteness of the studied genome as a reason for the absence of these genes. Therefore, alternative HC degrading genes that are present in the accessory part of the genomes might be responsible for the observed degradation.

On the other hand, the recent reconstruction of three metagenome-assembled Thermoplasmatota genomes (Poseidonia, MGIIa-L2, MGIIb-N1) from oil-exposed marine water samples (not included in the GTDB release89) contained enzymes involved in alkane (alkB) and xylene (xylM) degradation [30]. Hence as these global genome depositories continue to expand, we may have to revise or update our findings.

A total number of 597 archaeal genomes contain enzymes involved in the degradation of aromatic compounds regarding the conversion of catechol to TCA intermediates. This is observed in the phyla *Halobacterota* (176 genomes in *Haloferacaceae*, *Haloarculaceae*, *Natrialbaceae*, *Halococcaceae*,

Halobacteriaceae, *Methanocullaceae*, *Methanoregulaceae*, *Methanosarcinaceae*, *Archaeoglobaceae*, and some other methano-prefixed families), *Thermoplasmatota* (175 genomes in *Poseidoniales*, Marine Group III, *Methanomassiliicoccales*, UBA10834, *Acidiprofundales*, DHVEG-1, UBA9212), and *Crenarchaeota* (110 genomes in *Nitrospherales*, *Desulfurococcales*, *Sufolobales*, *Thermoproteales*). This widespread capacity for degrading downstream intermediates in aromatic HC degradation implies that archaea interact closely with bacteria in HC degradation.

Conclusions

HCs are ubiquitously distributed in the biosphere and do not exclusively originate from oil. In this study, the distribution of key HC degrading enzymes involved in the degradation of certain HCs (aliphatic and aromatic types) is provided at genome resolution for both the archaeal and bacterial domains. Extensive environmental genome and metagenome sequencing over the last decades has significantly increased the number of available microbial genomes and enriched contemporary genomic databases. The genome-based taxonomy using average nucleotide identity (ANI) or relative evolutionary divergence adopted by the Genome Taxonomy Database; GTDB [47,48] as a reproducible method has in parallel revised and updated some taxonomic ranks. The order Oceanospirillales, as an example, is a well-known taxon in the marine oil degradation context, and its representatives have been frequently reported as one of the main HC degrading members in response to oil pollution [49,50,37]. Nonetheless, this taxonomic rank has been removed from the genome-based taxonomy, and its members have been mainly placed in the order Pseudomonadales [51]. This could potentially cause a communication gap between the existing literature and new research. An updated comprehensive metabolic survey of Bacteria and Archaea for HC degradation potential at genome resolution could thus help bridge this gap. Our extensive survey shows that a greater diversity of bacteria is involved in aliphatic HC degradation compared to aromatic HCs. Few genomes were detected to contain all necessary enzymes to carry out complete degradation pathways. This reiterates previous findings that microbes generally cooperate for HC degradation by “division of labor” and a community perspective would therefore be crucial to predicting the fate of oil HCs in the ecosystem. We detected HC degrading ability among both r and k strategists and found signals of gene duplication and horizontal transfer of key HC degrading genes. This could be an efficient way to increase degradation capability among microbial members and potentially help them adapt to the available pool of HCs in their ecosystem.

Materials and methods

Data collection of HC Degrading enzymes. Representative compounds from each category of HCs, including saturated aliphatic (short/long-chain alkanes) and alicyclic (cyclohexane/cyclododecane), compounds with mono-aromatic (toluene, phenol, xylene, and benzene), and poly-aromatic (PAHs) (naphthalene, phenanthrene, and biphenyl as representatives) hydrocarbons were selected to survey the distribution of Bacteria and Archaea capable of their degradation under aerobic conditions.

A complete list of enzymes involved in the degradation pathway of mentioned HCs was compiled from previous reports [52–57]. We explored these enzymes in Kyoto Encyclopedia of Genes and Genomes (KEGG)[58], Pfam [59], TIGRFAMs [60], InterPro [61], and UniProt [62] databases. The accession number of enzymes in each mentioned database, their function, name, reaction (if available), EC number, and additional information are provided in **Supplementary Table S1**.

Distribution of HC degrading enzymes among bacterial and archaeal representative genomes. The distribution of the compiled HC degrading enzymes described in **Supplementary Table S1** was assessed across domains Bacteria and Archaea using AnnoTree (http://annotree.uwaterloo.ca) [12]. AnnoTree database is providing functional annotations for 24692 genome representatives in the genome taxonomy database (GTDB) release 89. The phylogenetic classification of genomes is derived from the GTDB database (release R89). In total, the annotation information for 18, 10, and 90 enzymes involved in the degradation process of alkane, cycloalkane, and aromatic HCs, respectively, were analyzed. Genome hits were collected at the thresholds of percent identity ≥ 50, e-value cut off $\leq 1e^{-5}$, subject/query percent alignment ≥ 70 for KEGG annotations, and e-value cut off ≤ $1e^{-5}$ for Pfam and TIGRFAMs annotations. For each HC degrading enzyme, we first checked KEGG annotations. If there were no KEGG accession numbers for the enzyme, the second priority was TIGRFAMs; otherwise, the Pfam annotation was considered. The table contains information for the distribution of HC degrading enzymes of each pathway present in representative genomes from bacteria and archaea domains, as is shown in **Supplementary Tables S2** and **S3**, respectively.

Phylogeny of bacteria and archaea augmented with the abundance of HC degrading enzymes. Evolview, a web-based tool for the phylogenetic tree visualization, management, and annotation, was used to present the distribution view of HC degrading enzymes in representative genomes across bacterial/archaeal phylogenomic trees [63][64].

The phylogenomic tree of bacteria and archaea in the Newick format, at the phylum level (123 and 14 leaves, respectively), was adopted from the AnnoTree website (November 21st, 2020). Trees were uploaded as the reference tree in Evolview. According to the abundance tables of HC degrading enzymes

prepared for each degradation pathway, four heatmaps were plotted for bacteria and archaea domains (separately for aliphatic and aromatic compounds).

Single gene phylogeny. To provide the evolutionary history of key enzymes in each HC degradation pathway, the protein sequence of that enzyme was manually confirmed by inspecting their conserved domains using the NCBI web CD-Search tool (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) [31]. Validated amino acid sequences were then aligned using Kalign3 software [65], and their phylogenetic tree was reconstructed using FastTree2 [66].

Acknowledgments

The computational analysis was performed at the Center for High-Performance Computing, School of Mathematics, Statistics, and Computer Science, University of Tehran.

Author contributions

M.M. devised the study. M.R.S., L.G.M., and M.M., performed the bioinformatics analysis M.R.S. and M.M. interpreted the data with input from S.M.M.D., S.B., M.A.A., and M.S.. M.R.S. and M.M drafted the manuscript. All authors read and approved the manuscript.

Conflict of interests

Author declare no conflict of interest.

References

1. Abdel-Aal HK, Aggour M, Fahim MA. petroleum and gas field processing. 2003.
2. Speight JG. Origin and occurance. Edition, F. Speight JG, HEINEMANN H, editors. Chem. Technol. Pet. Technol. Pet. CRC Press, Taylor & Francis Group; 2014.
3. Liu Y-F, Qi Z-Z, Shou L-B, Liu J-F, Yang S-Z, Gu J-D, et al. Anaerobic hydrocarbon degradation in candidate phylum "Atribacteria" (JS1) inferred from genomics. ISME J. Nature Publishing Group; 2019;13:2377–90.
4. Liu Q, Tang J, Bai Z, Hecker M, Giesy JP. Distribution of petroleum degrading genes and factor analysis of petroleum contaminated soil from the Dagang Oilfield, China. Sci Rep. Nature Publishing Group; 2015;5:1–12.
5. Fuchs G, Boll M, Heider J. Microbial degradation of aromatic compounds—from one strategy to four. Nat Rev Microbiol. Nature Publishing Group; 2011;9:803–16.
6. Peixoto RS, Vermelho AB, Rosado AS. Petroleum-degrading enzymes: bioremediation and new prospects. Enzyme Res. Hindawi; 2011;2011.
7. Sierra-Garcia IN, de Oliveira VM. Microbial hydrocarbon degradation: efforts to understand

444 biodegradation in petroleum reservoirs. Biodegrad Technol. InTech. doi; 2013;10:55920.
445 8. Xu X, Liu W, Tian S, Wang W, Qi Q, Jiang P, et al. Petroleum hydrocarbon-degrading bacteria for the
446 remediation of oil pollution under aerobic conditions: a perspective analysis. Front Microbiol. Frontiers;
447 2018;9:2885.
448 9. Varjani SJ. Microbial degradation of petroleum hydrocarbons. Bioresour Technol. Elsevier;
449 2017;223:277–86.
450 10. Xue J, Yu Y, Bai Y, Wang L, Wu Y. Marine oil-degrading microorganisms and biodegradation process
451 of petroleum hydrocarbon in marine environments: a review. Curr Microbiol. Springer; 2015;71:220–8.
452 11. McGenity TJ. Taxonomy, Genomics and Ecophysiology of Hydrocarbon-Degrading Microbes.
453 Springer; 2019.
454 12. Mendler K, Chen H, Parks DH, Lobb B, Hug LA, Doxey AC. AnnoTree: visualization and exploration of
455 a functionally annotated microbial tree of life. Nucleic Acids Res. Oxford University Press; 2019;47:4442–
456 8.
457 13. Cohen Y. Bioremediation of oil by marine microbial mats. Int Microbiol. Springer; 2002;5:189–93.
458 14. Ibraheem IBM. BIODEGRADABILITY OF HYDROCARBONS BY CYANOBACTERIA 1. J Phycol. Wiley
459 Online Library; 2010;46:818–24.
460 15. Raghukumar C, Vipparty V, David J, Chandramohan D. Degradation of crude oil by marine
461 cyanobacteria. Appl Microbiol Biotechnol. Springer; 2001;57:433–6.
462 16. Rodrigues EM, Gerais M, Gerais M, Gerais M. Detection of horizontal transfer of housekeeping and
463 hydrocarbons catabolism genes in bacterial genus with potential to application in bioremediation
464 process. Open Access Libr J. Scientific Research Publishing; 2018;5:1.
465 17. Phale PS, Shah BA, Malhotra H. Variability in assembly of degradation operons for naphthalene and
466 its derivative, carbaryl, suggests mobilization through horizontal gene transfer. Genes (Basel).
467 Multidisciplinary Digital Publishing Institute; 2019;10:569.
468 18. Korshunova A V, Tourova TP, Shestakova NM, Mikhailova EM, Poltaraus AB, Nazina TN. Detection
469 and transcription of n-alkane biodegradation genes (alk B) in the genome of a hydrocarbon-oxidizing
470 bacterium Geobacillus subterraneus K. Microbiology. Springer; 2011;80:682–91.
471 19. Hashmat AJ, Afzal M, Fatima K, Anwar-ul-Haq M, Khan QM, Arias CA, et al. Characterization of
472 hydrocarbon-degrading bacteria in constructed wetland microcosms used to treat crude oil polluted
473 water. Bull Environ Contam Toxicol. Springer; 2019;102:358–64.
474 20. Nie Y, Chi C-Q, Fang H, Liang J-L, Lu S-L, Lai G-L, et al. Diverse alkane hydroxylase genes in
475 microorganisms and environments. Sci Rep. Nature Publishing Group; 2014;4:1–11.

476 21. Love CR, Arrington EC, Gosselin KM, Reddy CM, Van Mooy BAS, Nelson RK, et al. Microbial
477 production and consumption of hydrocarbons in the global ocean. Nat Microbiol. Nature Publishing
478 Group; 2021;6:489–98.
479 22. Kertesz MA, Kawasaki A, Stolz A. Aerobic hydrocarbon-degrading alphaproteobacteria:
480 Sphingomonadales. Taxon genomics Ecophysiol Hydrocarb microbes. Springer; 2019;105–24.
481 23. Tao Y, Fishman A, Bentley WE, Wood TK. Altering toluene 4-monooxygenase by active-site
482 engineering for the synthesis of 3-methoxycatechol, methoxyhydroquinone, and methylhydroquinone. J
483 Bacteriol. Am Soc Microbiol; 2004;186:4705–13.
484 24. Parales RE, Parales J V, Pelletier DA, Ditty JL. Diversity of microbial toluene degradation pathways.
485 Adv Appl Microbiol. Elsevier; 2008;64:1–73.
486 25. Murphy CL, Sheremet A, Dunfield PF, Spear JR, Stepanauskas R, Woyke T, et al. Genomic Analysis of
487 the Yet-Uncultured Binatota Reveals Broad Methylotrophic, Alkane-Degradation, and Pigment
488 Production Capacities. MBio. Am Soc Microbiol; 2021;12:e00985-21.
489 26. Khomenkov VG, Shevelev AB, Zhukov VG, Zagustina NA, Bezborodov AM, Popov VO. Organization of
490 metabolic pathways and molecular-genetic mechanisms of xenobiotic degradation in microorganisms: A
491 review. Appl Biochem Microbiol. Springer; 2008;44:117–35.
492 27. Furukawa K, Suenaga H, Goto M. Biphenyl dioxygenases: functional versatilities and directed
493 evolution. J Bacteriol. Am Soc Microbiol; 2004;186:5189–96.
494 28. Corteselli EM, Aitken MD, Singleton DR. Description of Immundisolibacter cernigliae gen. nov., sp.
495 nov., a high-molecular-weight polycyclic aromatic hydrocarbon-degrading bacterium within the class
496 Gammaproteobacteria, and proposal of Immundisolibacterales ord. nov. and Immundisolibacteraceae f.
497 Int J Syst Evol Microbiol. Microbiology Society; 2017;67:925.
498 29. Schreiber L, Fortin N, Tremblay J, Wasserscheid J, Sanschagrin S, Mason J, et al. In situ microcosms
499 deployed at the coast of British Columbia (Canada) to study dilbit weathering and associated microbial
500 communities under marine conditions. FEMS Microbiol Ecol. Oxford University Press; 2021;97:fiab082.
501 30. Somee MR, Dastgheib SMM, Shavandi M, Maman LG, Kavousi K, Amoozegar MA, et al. Distinct
502 microbial communities along the chronic oil pollution continuum of the Persian Gulf converge with oil
503 spill accidents. bioRxiv. Cold Spring Harbor Laboratory; 2020;
504 31. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved
505 domain database in 2020. Nucleic Acids Res. Oxford University Press; 2020;48:D265–8.
506 32. Okie JG, Poret-Peterson AT, Lee ZMP, Richter A, Alcaraz LD, Eguiarte LE, et al. Genomic adaptations
507 in information processing underpin trophic strategy in a whole-ecosystem nutrient enrichment

508 experiment. Elife. eLife Sciences Publications Limited; 2020;9:e49816.
509 33. Somee MR, Dastgheib SMM, Shavandi M, Maman LG, Kavousi K, Amoozegar MA, et al. Distinct
510 microbial community along the chronic oil pollution continuum of the Persian Gulf converge with oil spill
511 accidents. Sci Rep. Nature Publishing Group; 2021;11:1–15.
512 34. Barbato M, Mapelli F, Crotti E, Daffonchio D, Borin S. Cultivable hydrocarbon degrading bacteria
513 have low phylogenetic diversity but highly versatile functional potential. Int Biodeterior Biodegradation.
514 Elsevier; 2019;142:43–51.
515 35. Kleindienst S, Grim S, Sogin M, Bracco A, Crespo-Medina M, Joye SB. Diverse, rare microbial taxa
516 responded to the Deepwater Horizon deep-sea hydrocarbon plume. ISME J. Nature Publishing Group;
517 2016;10:400–15.
518 36. Bacosa HP, Liu Z, Erdner DL. Natural sunlight shapes crude oil-degrading bacterial communities in
519 Northern Gulf of Mexico surface waters. Front Microbiol. Frontiers; 2015;6:1325.
520 37. Gutierrez T. Marine, aerobic hydrocarbon-degrading gammaproteobacteria: overview. Taxon
521 Genomics Ecophysiol Hydrocarb Microbes. Springer; 2017;1–10.
522 38. Sun X, Kostka JE. Hydrocarbon-degrading microbial communities are site specific, and their activity is
523 limited by synergies in temperature and nutrient availability in surface ocean waters. Appl Environ
524 Microbiol. Am Soc Microbiol; 2019;85:e00443-19.
525 39. Margesin R, Labbe D, Schinner F, Greer CW, Whyte LG. Characterization of hydrocarbon-degrading
526 microbial populations in contaminated and pristine alpine soils. Appl Environ Microbiol. Am Soc
527 Microbiol; 2003;69:3085–92.
528 40. Brzeszcz J, Steliga T, Kapusta P, Turkiewicz A, Kaszycki P. r-strategist versus K-strategist for the
529 application in bioremediation of hydrocarbon-contaminated soils. Int Biodeterior Biodegradation.
530 Elsevier; 2016;106:41–52.
531 41. Guo Q, Yin Q, Du J, Zuo J, Wu G. New insights into the r/K selection theory achieved in methanogenic
532 systems through continuous-flow and sequencing batch operational modes. Sci Total Environ. Elsevier;
533 2022;807:150732.
534 42. Wang M, Chen X, Liu X, Fang Y, Zheng X, Huang T, et al. Even allocation of benefits stabilizes
535 microbial community engaged in metabolic division of labor. bioRxiv. Cold Spring Harbor Laboratory;
536 2021;
537 43. Tsoi R, Wu F, Zhang C, Bewick S, Karig D, You L. Metabolic division of labor in microbial systems. Proc
538 Natl Acad Sci. National Acad Sciences; 2018;115:2526–31.
539 44. Krzmarzick MJ, Taylor DK, Fu X, McCutchan AL. Diversity and niche of archaea in bioremediation.

Archaea. Hindawi; 2018;2018.

45. Somee MR, Dastgheib SMM, Shavandi M, Zolfaghar M, Zamani N, Ventosa A, et al. Halophiles in bioremediation of petroleum contaminants: challenges and prospects. Bioremediation Environ Sustain. Elsevier; 2021. p. 251–91.

46. Fathepure BZ. Recent studies in microbial degradation of petroleum hydrocarbons in hypersaline environments. Front Microbiol. Frontiers; 2014;5:173.

47. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. Nature Publishing Group; 2018;

48. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. Nature Publishing Group; 2020;1–8.

49. Mason OU, Scott NM, Gonzalez A, Robbins-Pianka A, Bælum J, Kimbrel J, et al. Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. ISME J. Nature Publishing Group; 2014;8:1464.

50. King GM, Kostka JE, Hazen TC, Sobecky PA. Microbial responses to the Deepwater Horizon oil spill: from coastal wetlands to the deep sea. Ann Rev Mar Sci. Annual Reviews; 2015;7:377–401.

51. Liao H, Lin X, Li Y, Qu M, Tian Y. Reclassification of the taxonomic framework of orders cellvibrionales, oceanospirillales, pseudomonadales, and alteromonadales in class gammaproteobacteria through phylogenomic tree analysis. Msystems. Am Soc Microbiol; 2020;5:e00543-20.

52. Pérez-Pantoja D, González B, Pieper DH. Aerobic degradation of aromatic hydrocarbons. Handb Hydrocarb lipid Microbiol. Springer; 2010;799–837.

53. Abbasian F, Lockington R, Mallavarapu M, Naidu R. A comprehensive review of aliphatic hydrocarbon biodegradation by bacteria. Appl Biochem Biotechnol. Springer; 2015;176:670–99.

54. Abbasian F, Lockington R, Megharaj M, Naidu R. A review on the genetics of aliphatic and aromatic hydrocarbon degradation. Appl Biochem Biotechnol. Springer; 2016;178:224–50.

55. Meckenstock RU, Boll M, Mouttaki H, Koelschbach JS, Tarouco PC, Weyrauch P, et al. Anaerobic degradation of benzene and polycyclic aromatic hydrocarbons. J Mol Microbiol Biotechnol. Karger Publishers; 2016;26:92–118.

56. Rabus R, Boll M, Heider J, Meckenstock RU, Buckel W, Einsle O, et al. Anaerobic microbial degradation of hydrocarbons: from enzymatic reactions to the environment. J Mol Microbiol Biotechnol. Karger Publishers; 2016;26:5–28.

57. Espínola F, Dionisi HM, Borglin S, Brislawn CJ, Jansson JK, Mac Cormack WP, et al. Metagenomic

572 analysis of subtidal sediments from polar and subpolar coastal environments highlights the relevance of
573 anaerobic hydrocarbon degradation processes. Microb Ecol. Springer; 2018;75:123–39.
574 58. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. Oxford
575 University Press; 2000;28:27–30.
576 59. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families
577 database: towards a more sustainable future. Nucleic Acids Res. Oxford University Press; 2016;44:D279–
578 85.
579 60. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res. Oxford
580 University Press; 2003;31:371–3.
581 61. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, et al. The InterPro database, an
582 integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res.
583 Oxford University Press; 2001;29:37–40.
584 62. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The universal protein
585 resource (UniProt). Nucleic Acids Res. Oxford University Press; 2005;33:D154–9.
586 63. Subramanian B, Gao S, Lercher MJ, Hu S, Chen W-H. Evolview v3: a webserver for visualization,
587 annotation, and management of phylogenetic trees. Nucleic Acids Res. Oxford University Press;
588 2019;47:W270–5.
589 64. He Z, Zhang H, Gao S, Lercher MJ, Chen W-H, Hu S. Evolview v2: an online visualization and
590 management tool for customized and annotated phylogenetic trees. Nucleic Acids Res. Oxford
591 University Press; 2016;44:W236–41.
592 65. Lassmann T. Kalign 3: multiple sequence alignment of large datasets. Oxford University Press; 2020.
593 66. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large
594 alignments. PLoS One. Public Library of Science; 2010;5:e9490.

Figures

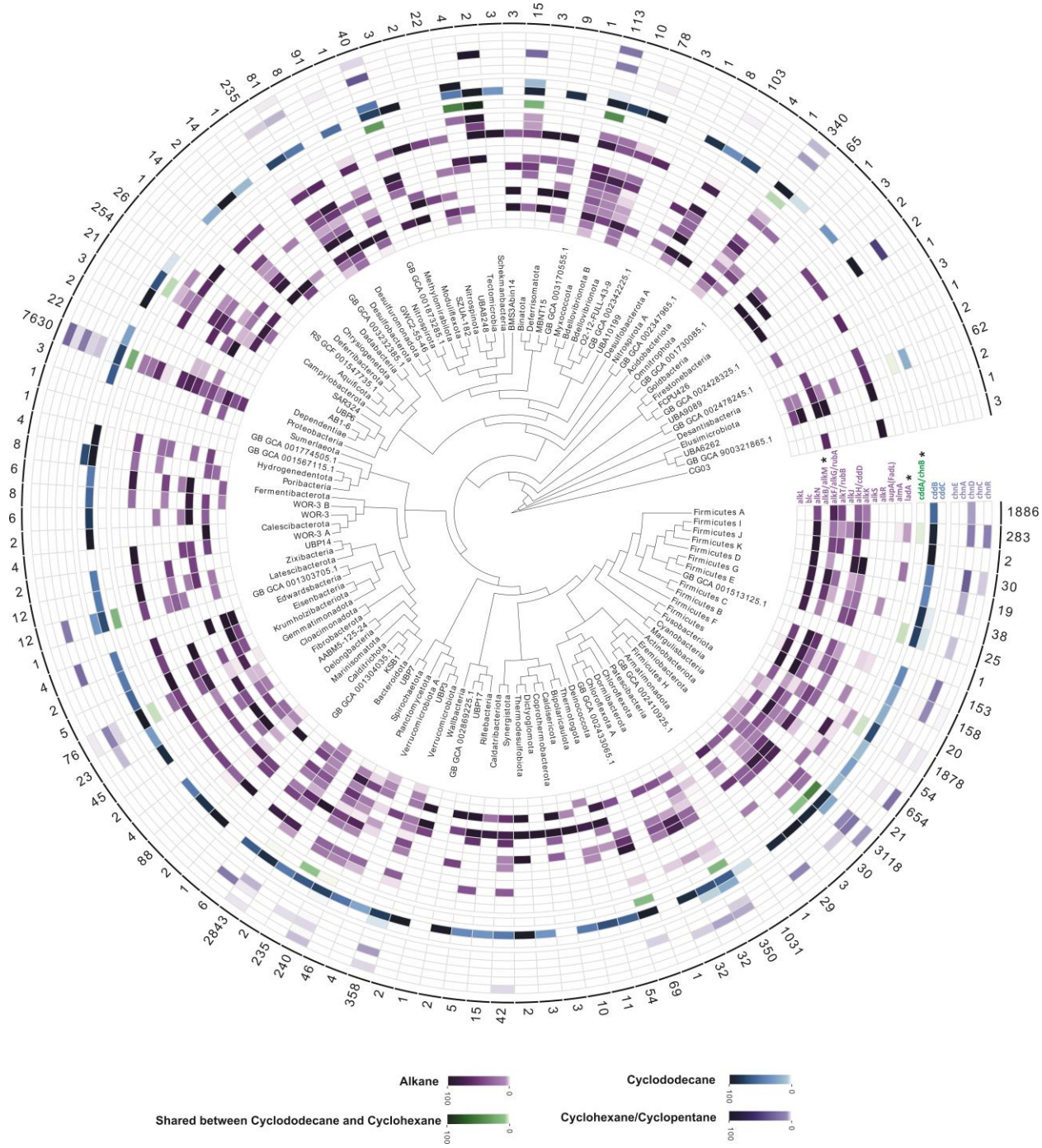


Figure 1- Distribution of aliphatic hydrocarbon-degrading genes across domain bacteria at the phylum level. Each circle of the heatmap represents a gene involved in HC degradation. Various compounds are shown in different colors, as represented in the color legend at the bottom of the figure. Genes marked with an asterisk represent key enzymes of the degradation pathway. Numbers written on each row's edge indicate the number of screened genomes in that phylum in the AnnoTree website (adopted from GTDB R89). The color gradient for genes of each compound indicates the percentage of HC degrading members of each phylum.

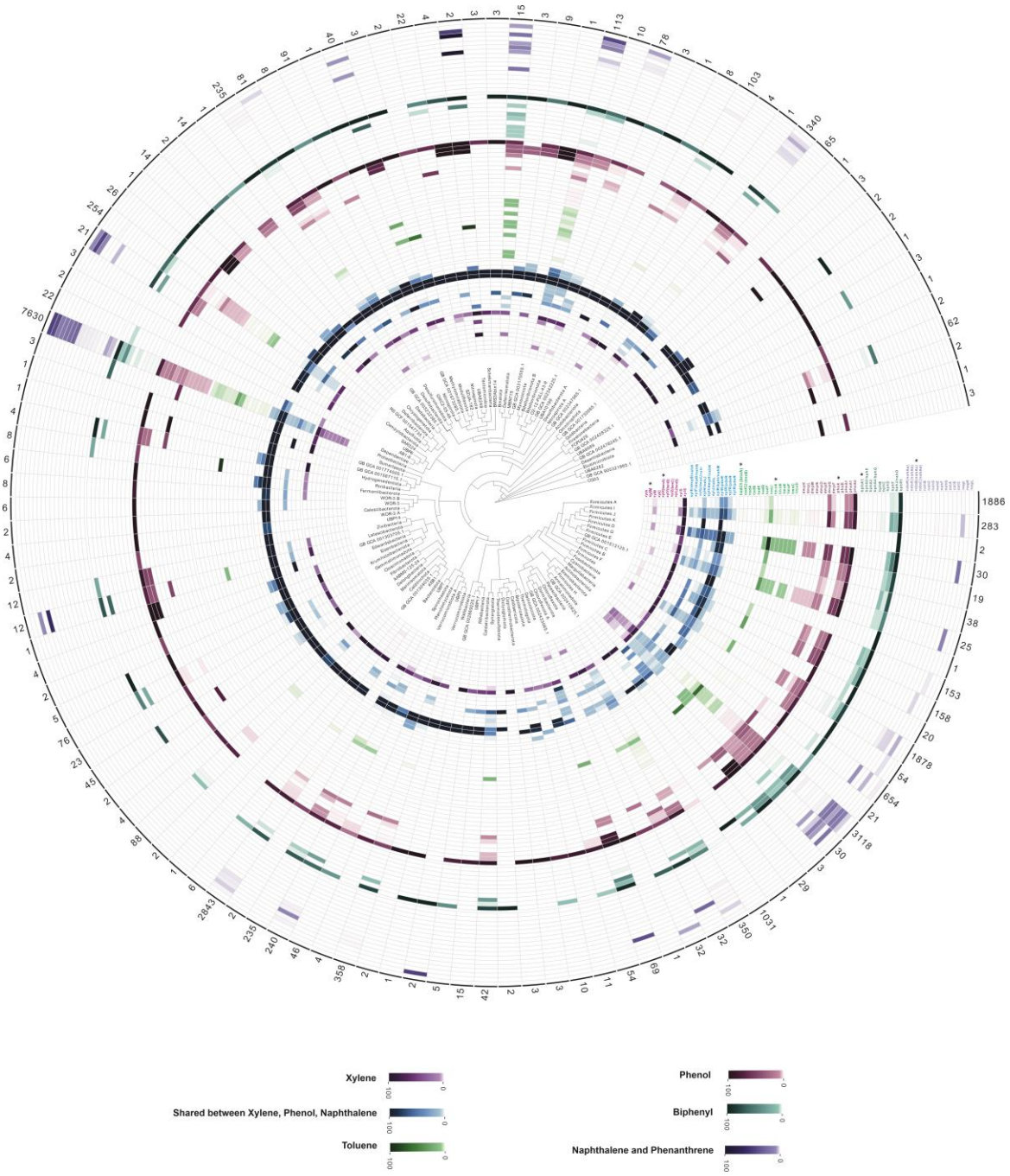


Figure 2- Distribution of aromatic hydrocarbon-degrading genes across domain bacteria at the phylum level. Each circle of the heatmap represents a gene involved in HC degradation. Various compounds are shown in different colors, as represented in the color legend at the bottom of the figure. Genes marked with an asterisk represent key enzymes of the degradation pathway. Numbers written on each row's edge indicate the number of screened genomes in that phylum in the AnnoTree website (adopted from GTDB R89). The color gradient for genes of each compound indicates the percentage of HC degrading members of each phylum.

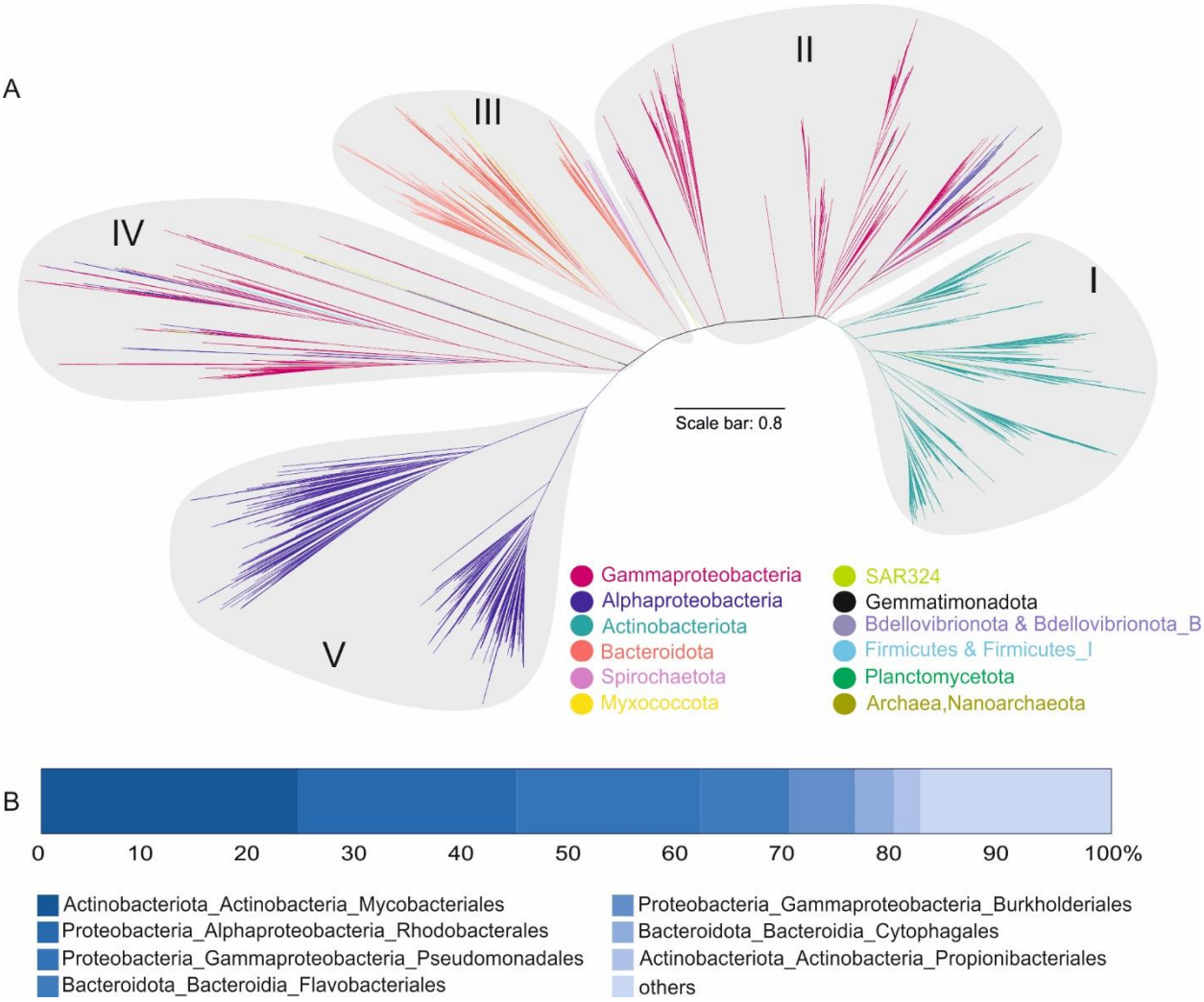


Figure 3- Maximum-likelihood phylogenetic reconstruction of amino acid sequences of alkB/M protein recovered from genomes (short-chain length alkane monooxygenase). A: Major clusters of alkB/M genes according to the reconstructed phylogeny. The scale bar indicates 0.1 branch distance. B: Bar plot representations of the distribution of recovered genes at the order level. The detailed information of the fraction "others" is provided in Supplementary Table S6.

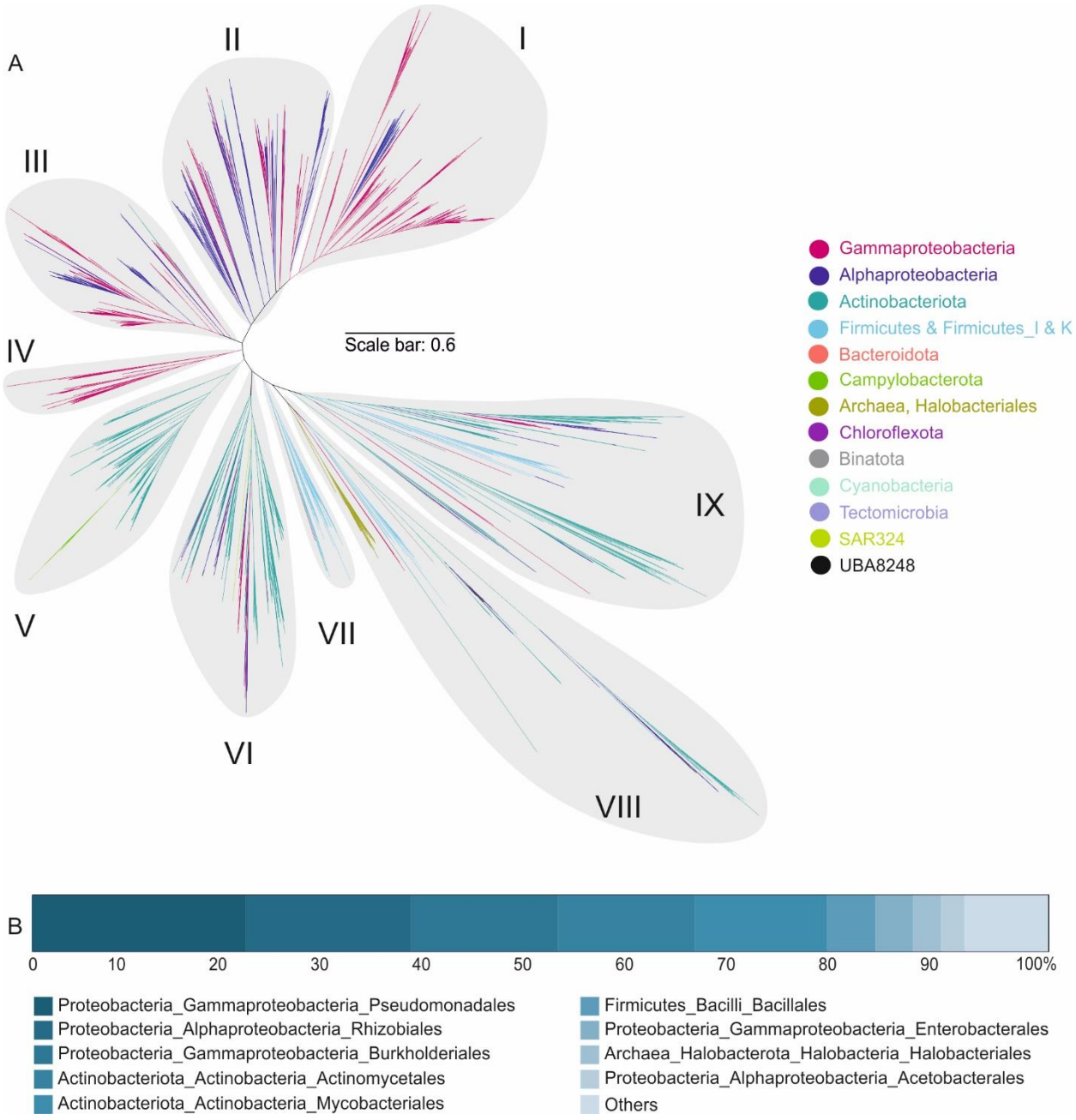


Figure 4- Maximum-likelihood phylogenetic reconstruction of amino acid sequences of ladA protein recovered from genomes (long-chain length alkane monooxygenase). A: Major clusters of ladA genes. The scale bar indicates 0.6 branch distance. B: Bar plot representations of the distribution of recovered genes at the order level. The detailed information of the fraction "others" is provided in Supplementary Table S6.

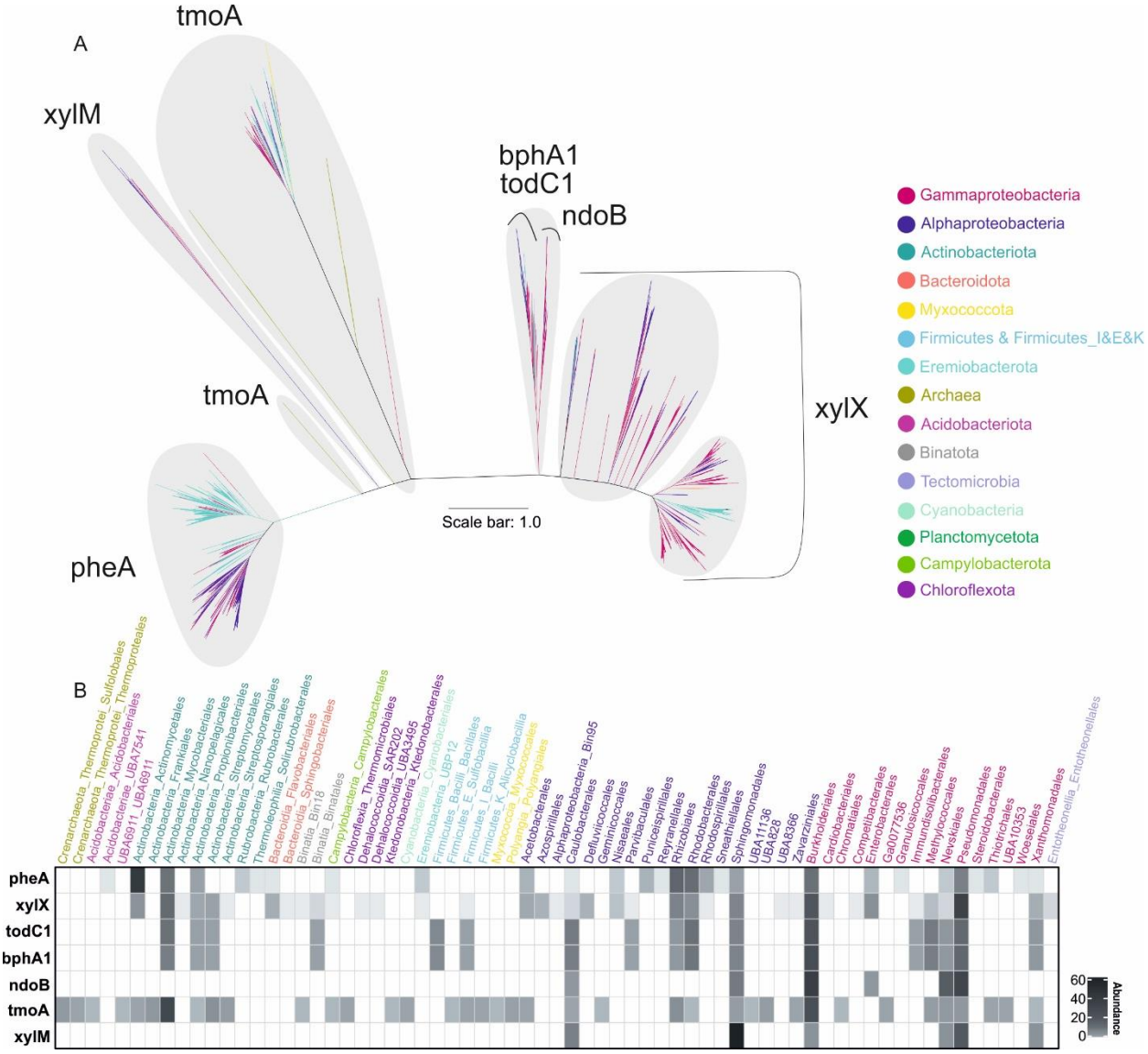


Figure 5- Maximum-likelihood phylogenetic reconstruction of amino acid sequences of ring-hydroxylating oxygenase (RHO) protein recovered from genomes. A: Major clusters of RHO genes. The scale bar indicates 1.0 branch distance. B: Heatmap representations of the distribution of recovered genes at the order level.

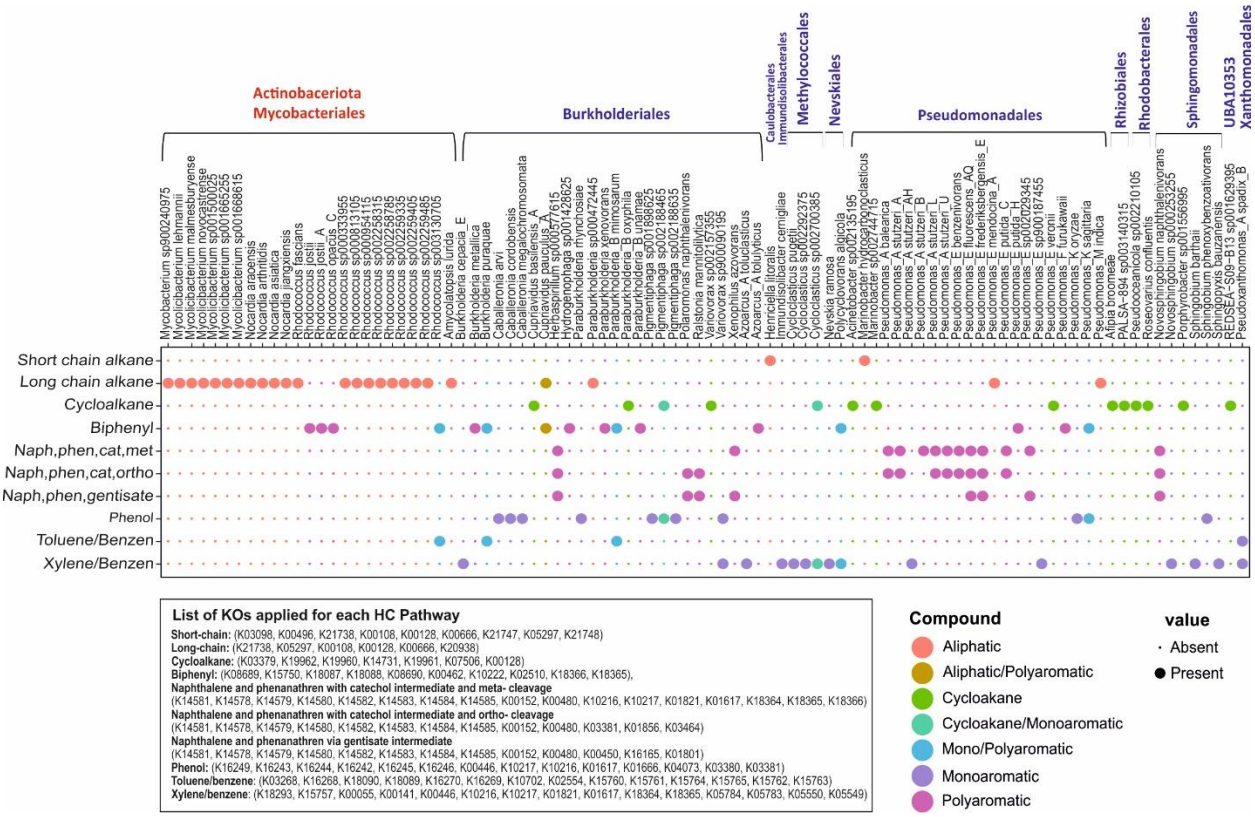
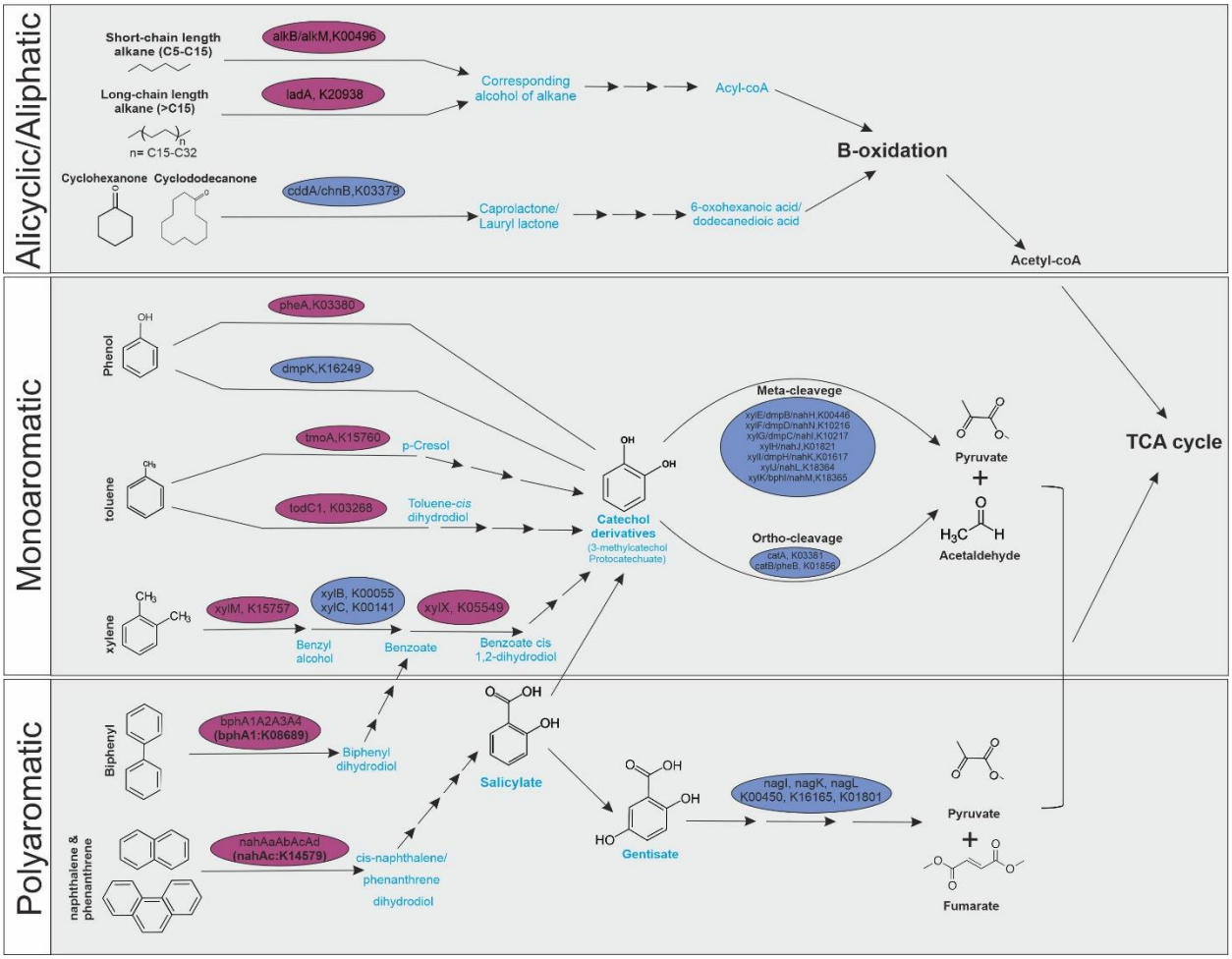
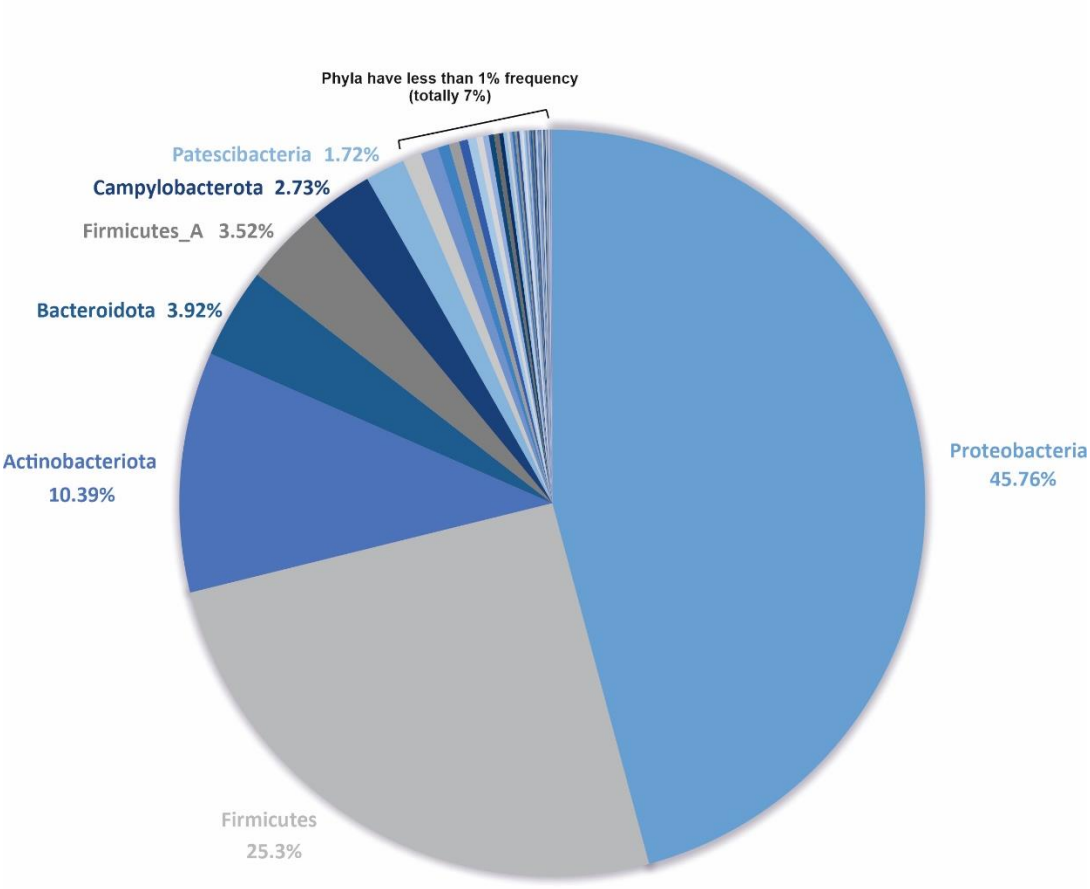


Figure 6- Genomes with complete/near complete degradation pathways of different HCs. Colors represent the type of HC that microbial genomes could degrade. Rows represent the type of HCs and columns show the name of genomes. Orders belonging to Proteobacteria and Actinobacteriota phyla are written in blue and red, respectively. KEGG orthologous accession number of enzymes for the complete degradation process of each compound is written at the figure's bottom.

Supplementary Figures

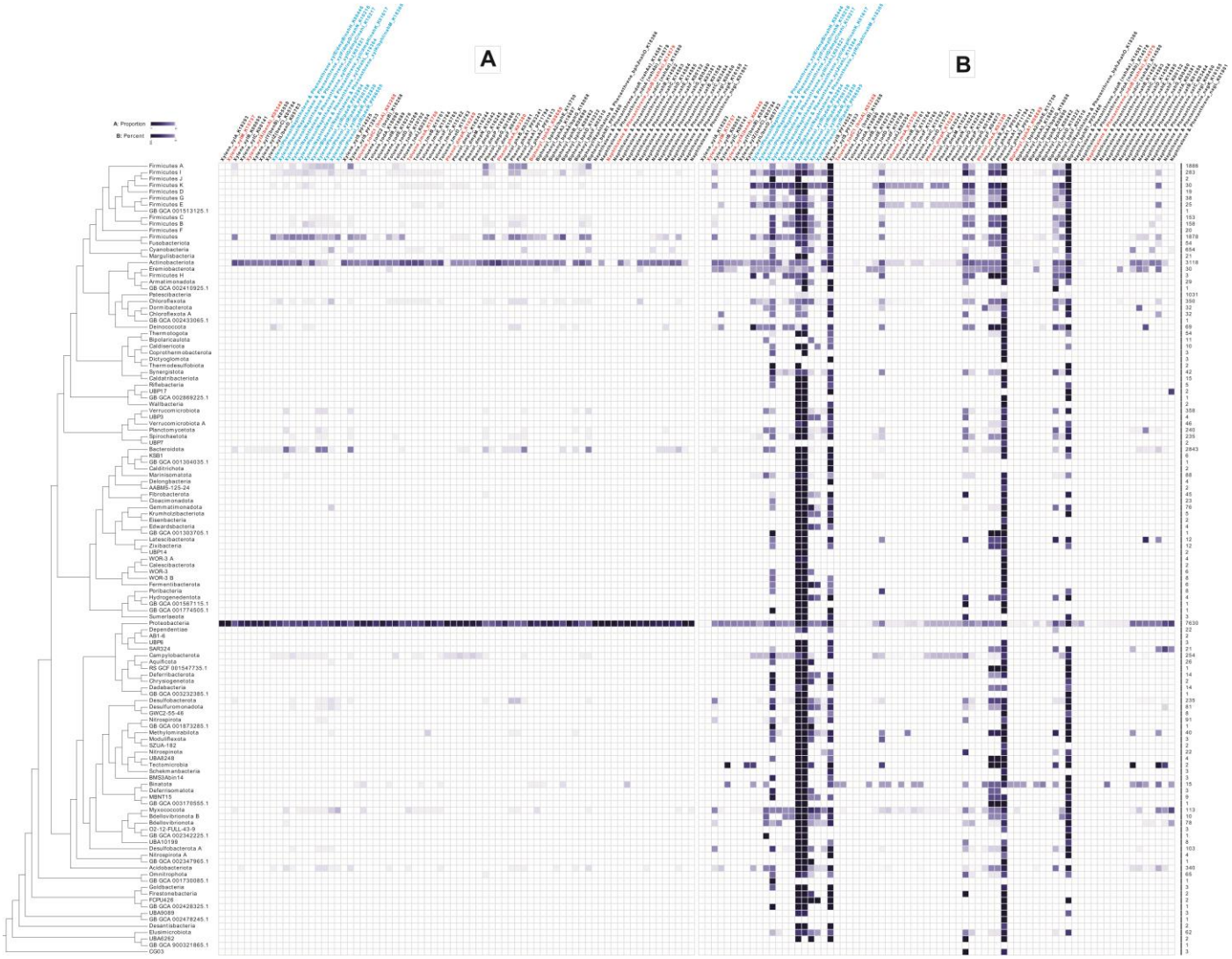


Supplementary Figure S1- Schematic representation of HC degradation pathways studied in this work. Purple circles show key HC degrading enzymes trigerring the degradation. Blue circles are other crucial enzymes. Important intermediate compounds are written in blue.

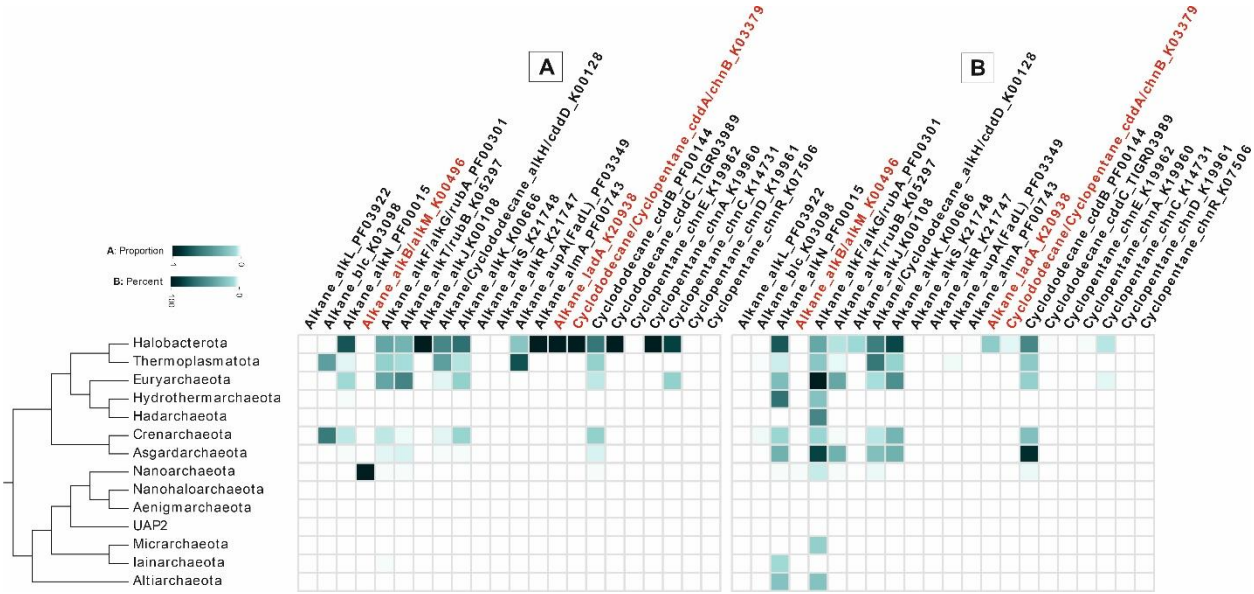


Supplementary Figure S2- Distribution of 143512 genomes of the GTDB database release 89 in different phyla.

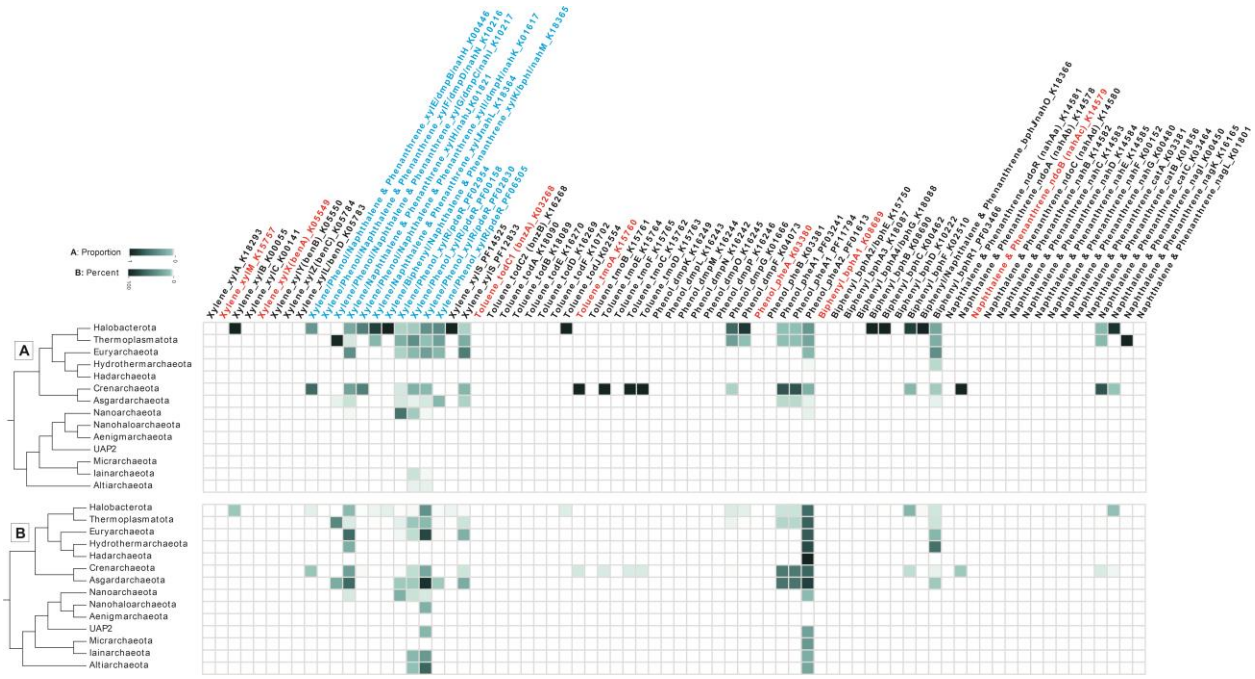
Supplementary Figure S3- Distribution of aliphatic hydrocarbon-degrading genes across domain bacteria at the phylum level. In plot A, the color gradient indicates the proportion of degrading members of each phylum to the entire HC degrading community. In plot B, the color gradient shows the percentage of HC degrading members of each phylum. Columns are the name of genes involved in HC degradation, which key ones are represented in red.



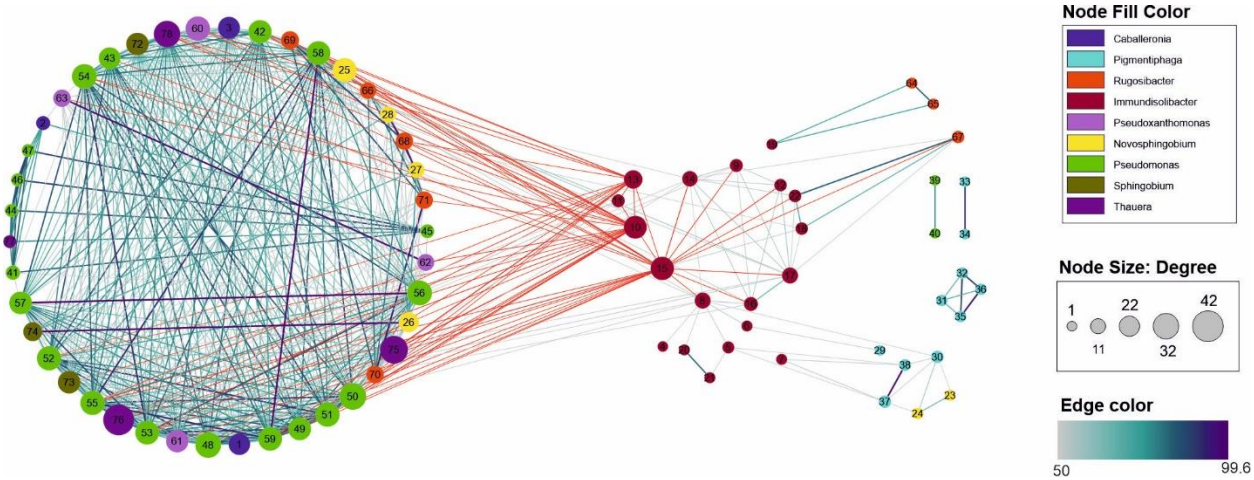
Supplementary Figure S4- Distribution of aromatic hydrocarbon-degrading genes across domain bacteria at the phylum level. In plot A, the color gradient indicates the proportion of degrading members of each phylum to the entire HC degrading community. In plot B, the color gradient shows the percentage of HC degrading members of each phylum. Columns are the name of genes involved in HC degradation, which key ones are represented in red. Enzymes written in blue are shared among the degradation processes of different aromatic compounds (xylene, phenol and naphthalene).



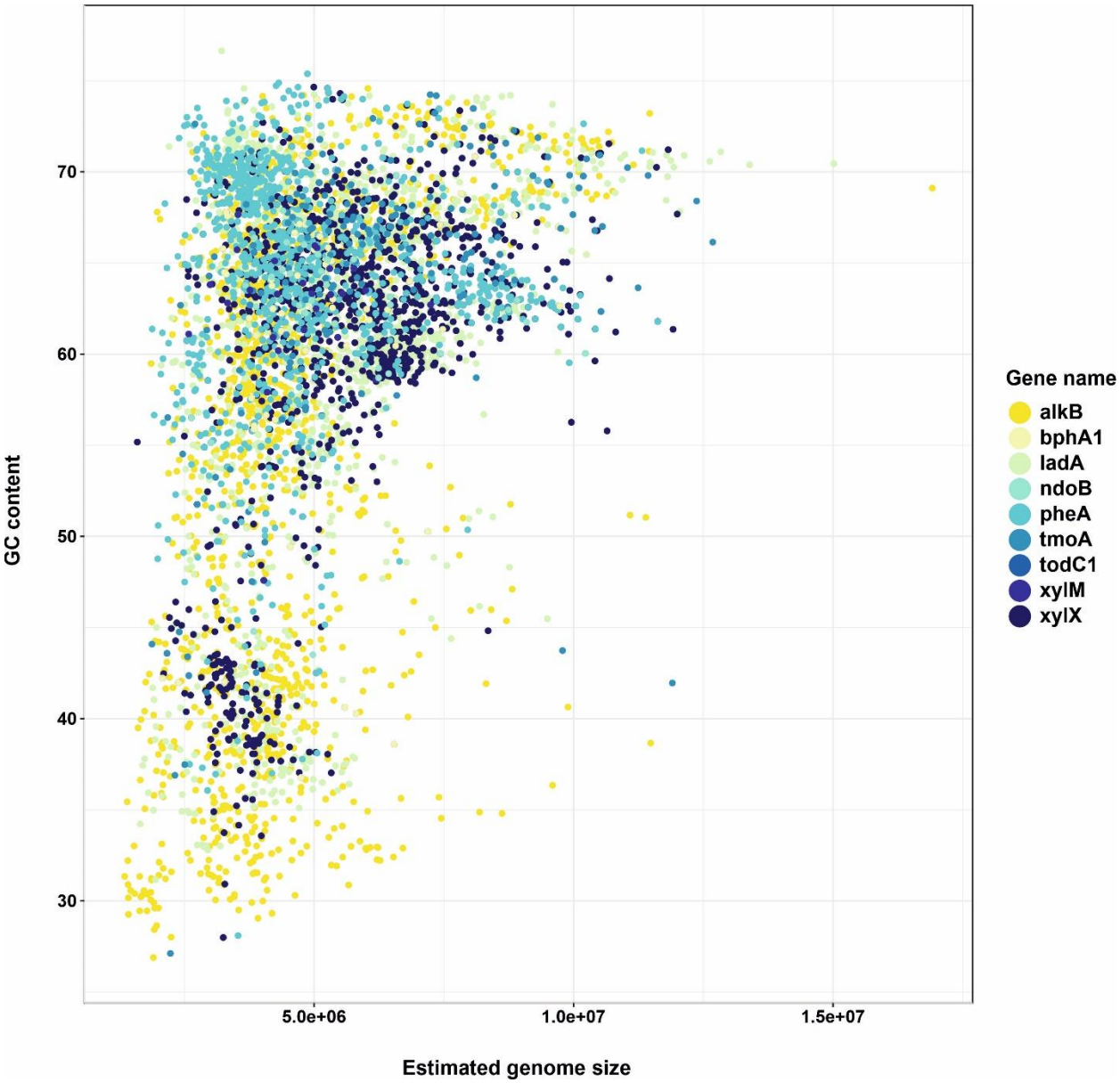
Supplementary Figure S5- Distribution of aliphatic hydrocarbon-degrading genes across domain archaea at the phylum level. In plot A, the color gradient indicates the proportion of degrading members of each phylum to the entire HC degrading community. In plot B, the color gradient shows the percentage of HC degrading members of each phylum. Columns are the name of genes involved in HC degradation, which key ones are represented in red.



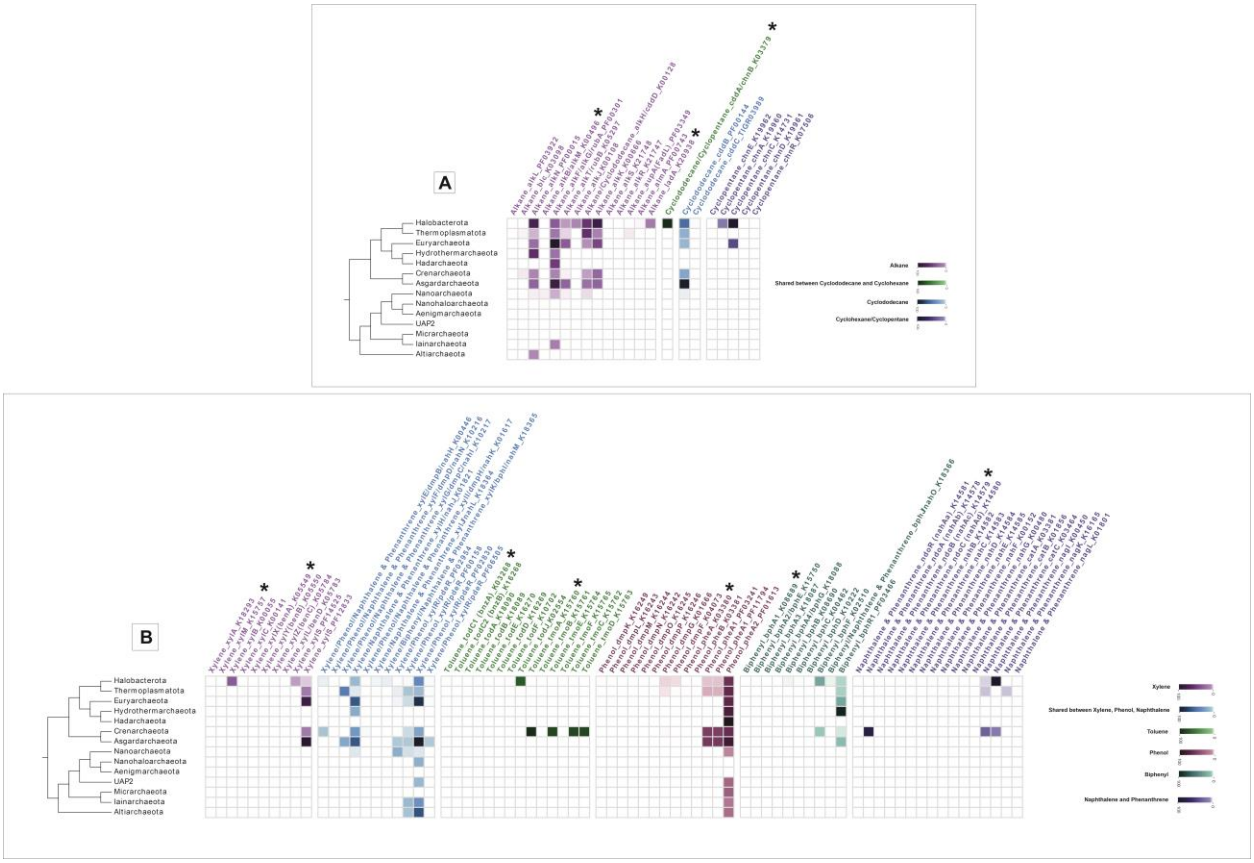
Supplementary Figure S6- Distribution of aromatic hydrocarbon-degrading genes across domain archaea at the phylum level. In plot A, the color gradient indicates the proportion of degrading members of each phylum to the entire HC degrading community. In plot B, the color gradient shows the percentage of HC degrading members of each phylum. Columns are the name of genes involved in HC degradation, which key ones are represented in red. Enzymes with blue color are shared among the degradation processes of different aromatic compounds (xylene, phenol and naphthalene).



Supplementary Figure S7- Network interaction between 18 copies of xylX gene in *Immundisolibacter cernigliae* and other genomes with more than two copies of this gene. Only the blast identity values between 50 to 100 percent are shown. Edges are color-coded based on their blast identity. The size of nodes is based on the "Degree," which is determined by the number of edges of each node. Edges in red are versions of xylX in *Immundisolibacter cernigliae* that had a higher degree than others. The gene ID of the assigned number of each node is represented in Supplementary Table S7.



Supplementary Figure S8- Distribution of genome size versus GC content of the studied genomes with key HC degrading genes.



Supplementary Figure S9- Distribution of aliphatic (A) and aromatic (B) hydrocarbon-degrading genes across domain archaea at the phylum level. Columns show the name of genes involved in HC degradation and are represented in different colors for various compounds. The color gradient for genes of each compound indicates the percentage of HC degrading members of each phylum.