frontiers in Microbiology

METHODS
published: 07 January 2022
doi: 10.3389/fmicb.2021.764058

OPEN ACCESS

Edited by:
Nicole Buan,
University of Nebraska-Lincoln,
United States

Reviewed by:
Lauren M. Lui,
Lawrence Berkeley National Laboratory, United States
Yanni Sun,
City University of Hong Kong,
Hong Kong SAR, China

***Correspondence:**
Srijak Bhatnagar
sbhatnagar@athabascau.ca

† These authors have contributed equally to this work and share first authorship

‡ **Present address:**
Anirban Chakraborty
Department of Biological Sciences, Idaho State University, Pocatello, ID, United States
Alyse K. Hawley,
School of Engineering, University of British Columbia Okanagan, Kelowna, BC, Canada
Srijak Bhatnagar,
Faculty of Science and Technology, Athabasca University, Athabasca, AB, Canada

Specialty section:
This article was submitted to Microbial Physiology and Metabolism, a section of the journal Frontiers in Microbiology

Received: 24 August 2021
Accepted: 08 December 2021
Published: 07 January 2022

Citation:
Khot V, Zorz J, Gittins DA, Chakraborty A, Bell E, Bautista MA, Paquette AJ, Hawley AK, Novotnik B, Hubert CRJ, Strous M and Bhatnagar S (2022) CANT-HYD: A Curated Database of Phylogeny-Derived Hidden Markov Models for Annotation of Marker Genes Involved in Hydrocarbon Degradation. Front. Microbiol. 12:764058. doi: 10.3389/fmicb.2021.764058

CANT-HYD: A Curated Database of Phylogeny-Derived Hidden Markov Models for Annotation of Marker Genes Involved in Hydrocarbon Degradation

Varada Khot[1†], Jackie Zorz[1†], Daniel A. Gittins[2], Anirban Chakraborty[2‡], Emma Bell[2], María A. Bautista[2], Alexandre J. Paquette[1], Alyse K. Hawley[1‡], Breda Novotnik[1], Casey R. J. Hubert[2], Marc Strous[1] and Srijak Bhatnagar[2‡]*

[1] Energy Bioengineering and Geomicrobiology Group, Department of Geoscience, University of Calgary, Calgary, AB, Canada, [2] Energy Bioengineering and Geomicrobiology Group, Department of Biological Sciences, University of Calgary, Calgary, AB, Canada

Many pathways for hydrocarbon degradation have been discovered, yet there are no dedicated tools to identify and predict the hydrocarbon degradation potential of microbial genomes and metagenomes. Here we present the Calgary approach to ANnoTating HYDrocarbon degradation genes (CANT-HYD), a database of 37 HMMs of marker genes involved in anaerobic and aerobic degradation pathways of aliphatic and aromatic hydrocarbons. Using this database, we identify understudied or overlooked hydrocarbon degradation potential in many phyla. We also demonstrate its application in analyzing high-throughput sequence data by predicting hydrocarbon utilization in large metagenomic datasets from diverse environments. CANT-HYD is available at https://github.com/dgittins/CANT-HYD-HydrocarbonBiodegradation.

Keywords: hydrocarbon degradation, Marker genes, Hidden Markov Models, gene annotation, hydrocarbon cycling

INTRODUCTION

Hydrocarbons are diverse compounds consisting of carbon and hydrogen atoms that differ in size, structure, and reactivity. They can be the product of geological processes as well as produced biogenically by organisms in all domains of life (Tornabene et al., 1969; Lerdau et al., 1997; Lea-Smith et al., 2015). Assessing hydrocarbon use by microorganisms, as a source of carbon and/or energy, is important for evaluating the consequences of hydrocarbon presence or contamination (Atlas and Hazen, 2011), understanding the global carbon cycle (González-Gaya et al., 2019), and for industrial applications, such as the synthesis of biocatalysts (Prier and Kosjek, 2019). Degradation of hydrocarbon molecules is kinetically challenging due to the chemical inertness of the organic C–H bond, and when present, the stability of aromatic ring structures (Rabus et al., 2016). Microorganisms employ a range of enzymes to use hydrocarbons (Rabus et al., 2016; Xu et al., 2018) in oxic and anoxic conditions. Catabolism of these hydrocarbons is coupled with

Abbreviations: CANT-HYD, Calgary approach to ANnoTating HYDrocarbon degradation genes; HMM, Hidden Markov Models; GTDB, Genome Taxonomy Database; BTEX, Benzene, Toluene, Ethyl Benzene, and Xylene.

reduction of terminal electron acceptors such as oxygen, nitrate, sulfate, and iron or *via* syntrophy with methanogens (Zhang et al., 2019).

The discovery of hydrocarbon degrading microorganisms has traditionally relied on cultivation in the laboratory using hydrocarbon substrates (Rueter et al., 1994; Kniemeyer et al., 2007). Successful cultivation preceded the identification of genes involved in hydrocarbon metabolism with techniques such as gene knockouts, protein expression analyses, and gene sequencing (Schneiker et al., 2006; Wang and Shao, 2014; Gregson et al., 2018; Wang et al., 2018; Liu et al., 2019; Li et al., 2020). These studies are crucial for providing fundamental knowledge on the ever-growing diversity of hydrocarbon degrading microorganisms as well as uncovering new degradation pathways. The recent exponential rise in sequence data and the consequential increase in known microbial diversity have provided new opportunities to explore hydrocarbon degradation potential in diverse environments and uncultured microorganisms. One approach for exploring sequence data is to annotate genes using Hidden Markov Model (HMM). HMMs are trained on the multiple sequence alignments of amino acid sequences and produce position-specific scores and penalties when searching query sequences. HMMs have better sensitivity and recall for identifying homologs of conserved protein domains, compared to conventional pairwise alignment tools such as blastp (McGinnis and Madden, 2004), which use a position-independent scoring matrix (Eddy, 2004). Detection of metabolic potential in whole genomes or metagenomic datasets is generally accomplished using functional annotation tools aided by HMM databases such as KEGG (Kanehisa and Goto, 2000) and Pfam (Finn et al., 2014). While these large databases can confidently identify central metabolic and other well studied pathways, specific HMMs and tools for accurate annotation of catalytic genes in hydrocarbon degradation pathways are currently lacking. Genes involved in hydrocarbon degradation can share sequence similarity to genes from other metabolic pathways and consequently, are often misannotated (Callaghan et al., 2008; Khelifi et al., 2014). Hence, there is a need for a purpose-built tool for the accurate detection of hydrocarbon degradation pathways in sequence data.

Here we present the Calgary approach to ANnoTating HYDrocarbon degradation genes (CANT-HYD), a database of 37 HMMs designed for the identification and annotation of marker genes that are critical for the aerobic and anaerobic degradation of alkane and aromatic hydrocarbons. CANT-HYD is tested and validated against 72 genomes of known hydrocarbon degrading bacteria, representing a broad spectrum of hydrocarbon metabolism. Using these validated HMMs, over 30,000 representative genomes covering the entire bacterial and archaeal tree of life are analyzed to identify hydrocarbon degrading microorganisms. Forty-one publicly available metagenomes from diverse environments are also analyzed using CANT-HYD to explore hydrocarbon degradation potential in diverse environments. Lastly, we compare the performance of CANT-HYD HMMs to their counterparts from eggNOG, Pfam, and KEGG Orthology (KO) databases.

METHODS

Selection and Clustering of Archetype Reference Sequences

Enzymes involved in the activation of hydrocarbon substrates in aerobic and anaerobic hydrocarbon degradation pathways of aliphatic and aromatic compounds were identified through a literature search (**Figure 1**). Amino acid sequences encoding the catalytic subunits of these enzymes were obtained from Genbank and were classified as either "experimentally verified" or "putative." The "experimentally verified" sequences refer to amino acid sequences from published studies with experimental proof of the intended function. Experimental proof consisted of gene cloning or protein purification and corresponding enzyme assays, or gene knockout studies. Gene sequences labeled "putative" refer to sequences with strong evidence of function but lacking these definitive prerequisite analyses. Putative sequences often originated from isolates or enrichment cultures where there is evidence of hydrocarbon degradation or genomic and/or proteomic evidence for the enzyme responsible. The resulting curated 105 amino acid sequences from 53 different species (**Supplementary Table 1**) are referred henceforth as "archetype" as they represent the gene sequence pattern for a verifiable hydrocarbon degradation function. Amino acid sequences were clustered into homologous groups based on $\geq$20% amino acid identity as determined by blastp v2.9.0 (McGinnis and Madden, 2004) to place related archetype sequences on the same phylogenetic tree and reduce the downstream computational resource requirements. A loose grouping as carried out here is unlikely to affect the final HMM, as manual curation and pruning of phylogenetic trees in downstream processing took this clustering into account. Archetype sequences that did not cluster at 20% were either manually added into homologous groups of similar function or left as singletons.

Homology Search to Obtain Sequences Similar to Archetypes

Amino acid sequences sharing sequence homology to the 105 archetype genes were recruited from the NCBI non-redundant (nr) protein database using a Diamond homology search (Buchfink et al., 2014). All hits with query coverage $\geq$70% and e-value $\leq 10^{-4}$ were retained as putatively phylogenetically related sequences and added to the query's homologous group. The sequences of homologous groups were dereplicated, followed by clustering at 98% amino acid identity using the USEARCH v9.0.2 *derep_fulllength* and *cluster_fast* commands (Edgar, 2010). The resulting sequences from the Diamond homology search are referred to as the "DIAMOND sequences" database and were used in downstream steps of HMM construction, and to generate cutoff scores for the HMMs.

Grouping Genes With Similar Functions Using Phylogenetic Analysis

A multiple sequence alignment was generated for each clustered (98%) homologous group using MUSCLE v3.8.31 (Edgar, 2004). The alignments were used to create maximum-likelihood trees

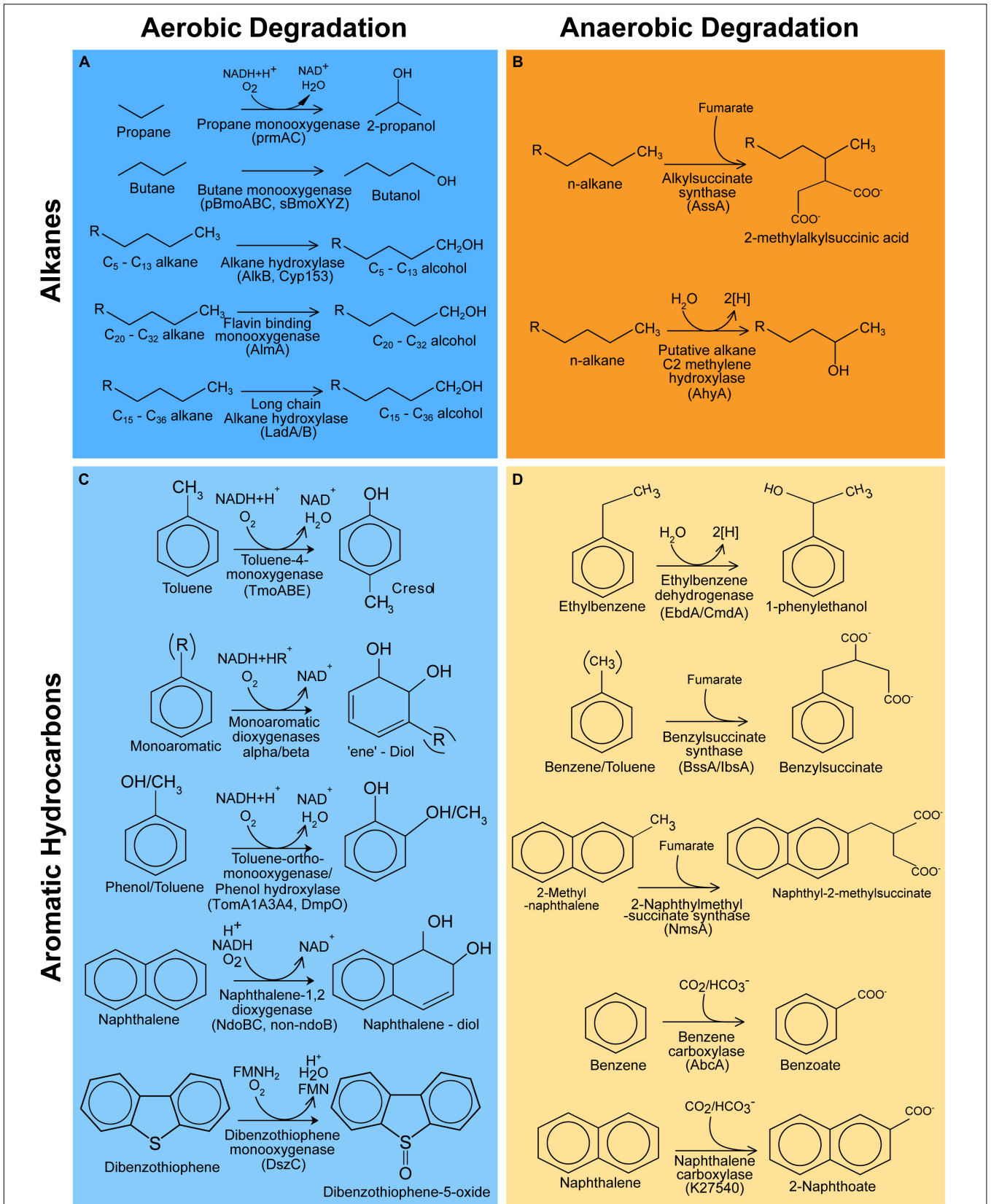

FIGURE 1 | Hydrocarbon degradation reactions covered by CANT-HYD. Reactions for the degradation of alkanes through aerobic **(A)** and anaerobic pathways **(B)** and degradation of aromatic hydrocarbons through aerobic **(C)** and anaerobic **(D)** pathways.

using FastTreeMP v2.1 (Price et al., 2010) with the parameters –*pseudo* and *-spr 4*. Trees were manually inspected using iTOL v5.6.3 (Letunic and Bork, 2019) or Dendroscope v3.6.3 (Huson et al., 2007) to identify monophyletic clades of genes containing experimentally verified archetype sequences (**Supplementary Figure 1**). These monophyletic clades were easily identifiable as archetype clustering was carried out at a low threshold (on ≥20% amino acid identity) and thus a substrate-specific set of seed sequences was extracted from each clade. In some instances, where a clear monophyletic distinction was lacking, a broader function-specific set of seed sequences were extracted (e.g., MAH_alpha group includes TcbA, IpbA, BnzA, and BphA).

Processing Homologous Groups With >5,000 Sequences

As some group sizes were in the order of 10^5 sequences and the computational requirement for aligning sequences grows exponentially with every added sequence, a nested clustering and phylogenetic pruning approach was implemented to overcome computational challenges for groups >5,000 sequences. The homologous group was first clustered at a lower identity (e.g., 50%) to reduce the size of the group, followed by alignment, phylogenetic reconstruction, and phylogenetic neighborhood pruning as described above. This process reduced the sequence search space around the archetype sequences by pruning the phylogenetic trees at a higher identity threshold. Because each sequence in a clustered group represents a group of sequences, for the selected pruned neighborhood, the clustered sequences were placed back in. Then the process of pruning the phylogenetic neighborhood of the reference sequence(s) was iterated using a higher clustering identity (e.g., 70%, 90%, etc.) until the prune group was ≤5,000 sequences or the clustering identity was raised to 98% (**Figure 2**), at which point the seed sequences for HMM creation were selected as described above.

HMM Creation and Determination of Cutoff Scores

Because HMMs are sensitive to the alignment and position of each amino acid, the seed sequences that are now orders of magnitude smaller than the homologous groups they were derived from were realigned using a sensitive aligner, Clustal-Omega v1.2.4 (Sievers et al., 2011) followed by manual inspection of the alignment using Jalview v2.11.1.4 (Waterhouse et al., 2009), and generation of HMMs using the *hmmbuild* command of HMMER v3.2.1 (Eddy, 2011). The HMMs were used to search the archetype reference sequences and 'DIAMOND Sequences' database (**Figure 2**) using *hmmsearch* of HMMER v3.2.1 (Eddy, 2011). The domain scores of the hits to each HMM were plotted to visualize the frequency distribution pattern of the scores (**Supplementary Figure 2**). A "trusted" and a "noise" cutoff was chosen for each HMM using these score distributions. The trusted cutoff is the domain score above which a sequence can be confidently annotated for the function, as all experimentally verified genes used for the HMM scored above this cutoff. The noise cutoff was chosen to exclude genes that were predicted to have a different function. Thus, any hits scoring below the noise cutoffs are expected to *not* carry out the function represented by the HMM. **Supplementary Table 2** includes information on genes that are the closest phylogenetic relatives to the archetype sequences of CANT-HYD HMMs.

Validation of CANT-HYD HMMs Using Genomes of Known Hydrocarbon Degraders

Seventy-two genomes of microorganisms with published experimental evidence of an ability to degrade hydrocarbons were downloaded from GenBank and RefSeq (O'Leary et al., 2016) and categorized by the type of substrate and respiration (**Supplementary Table 3**). If the exact strain was not available, its closest relative from the Genome Taxonomy Database (GTDB) was chosen. For example, *Aromatoleum aromaticum* EbN1 anaerobically degrades aromatic compounds (Wöhlbrand et al., 2007). The genomes were searched using CANT-HYD HMMs and the resulting gene annotations, scoring above the trusted cutoff, were compared to the established degradation capability of the organism. If a gene hit multiple HMMs above the confidence threshold, it was assigned to the highest scoring HMM.

Analysis of GTDB Genomes to Identify Potentially Novel Hydrocarbon Degrading Bacteria

The GTDB database (05-RS95 17th July 2020) (Parks et al., 2020) of representative bacterial and archaeal genomes was downloaded and searched using the CANT-HYD HMMs. For further investigation, gene sequences from cyanobacterial genomes with hits to LadA beta (above the noise cutoff) were combined with archetype reference sequences of long-chain alkane monooxygenases (LadA-alpha, LadA-beta, and LadB). The combined sequences were then clustered at 70% amino acid identity using USEARCH v9.0.2132_i86linux64 (Edgar, 2010) *cluster_fast*. Representative sequences of each cluster were then aligned using Muscle v3.8.31 (Edgar, 2004), followed by a maximum-likelihood phylogenetic reconstruction using FastTreeMP v2.1 (Price et al., 2010) (**Supplementary Data Sheet 1**).

Analysis of Diverse Metagenomes Using CANT-HYD

Metagenomes representing diverse environments such as petroleum reservoirs (Hu et al., 2016; Nie et al., 2016; Liu et al., 2018; Christman et al., 2020), oil spill experimental microcosms (Tan et al., 2015; Dombrowski et al., 2016), marine systems (Orellana et al., 2017; Tully et al., 2018; Dong et al., 2019, 2020), host-associated microbiomes (Feigelman et al., 2017; Herman et al., 2020; Avila-Magaña et al., 2021), and other environments (Yao et al., 2017; Zorz et al., 2019), were downloaded either as unassembled raw data from the NCBI SRA or as predicted gene sequences from the JGI Genome Portal (**Supplementary Table 4**). Raw reads from unassembled metagenomes were filtered using BBDuk[1] for a minimum quality of 15 and a minimum read length of 150 bp. Reads passing quality control

[1]Bushnell, B. *BBTools*. Available online at: https://jgi.doe.gov/data-andtools/bbtools/

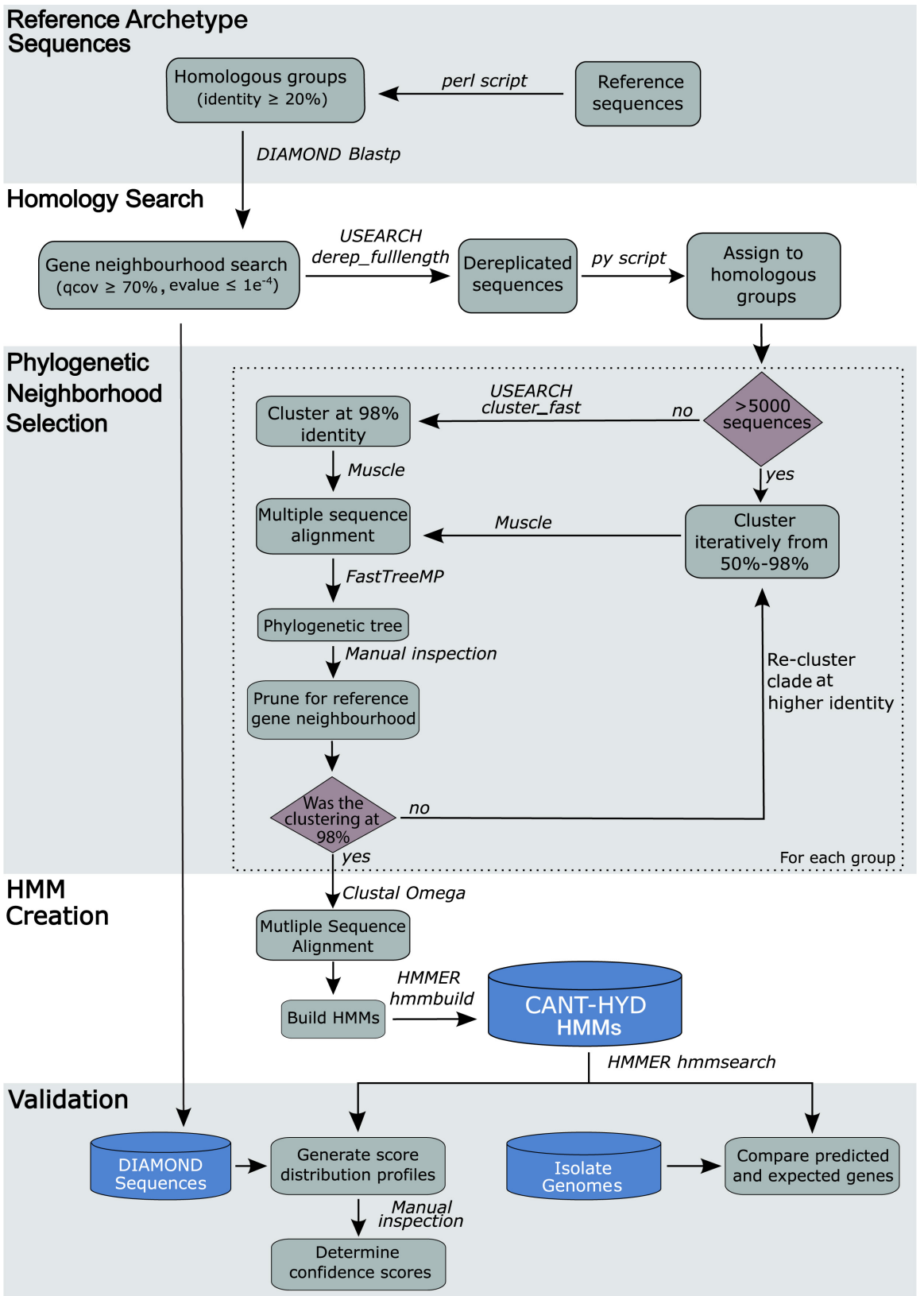

FIGURE 2 | Workflow of the process underlying the creation of the CANT-HYD HMM Database.

were assembled using MEGAHIT (Li et al., 2015) with default parameters, followed by gene calling by Prodigal v2.6.3 (Hyatt et al., 2010) with the metagenomic option (*-p meta*). The amino acid sequences of predicted genes were searched against the CANT-HYD database using the *hmmsearch* command of HMMER v3.2.1 (Eddy, 2011) and only hits scoring above the noise cutoff for each HMM were visualized. Hit count for each metagenome was normalized by the total number of predicted genes.

Comparison of CANT-HYD HMMs to Existing HMM Databases

The CANT-HYD HMMs were compared to equivalent HMMs from Pfam (Bateman et al., 2000), eggNOG (Huerta-Cepas et al., 2019) and KO (Kanehisa et al., 2016). Archetype sequences used to build CANT-HYD HMMs were annotated using eggNOG mapper (Cantalapiedra et al., 2021) with default parameters to identify and retrieve the closest eggNOG, Pfam and Kofam HMMs (**Supplementary Table 5**). These equivalent HMMs were used to annotate the isolate genomes, and hydrocarbon enrichment and host-associated metagenomes using *hmmsearch* (Eddy, 2011). Because suggested cutoffs were not included with eggNOG, Pfam, or KO database, an e-value cutoff of 10^{-50} was used to filter results.

RESULTS AND DISCUSSION

Validation of CANT-HYD HMMs

Genomes of 72 microorganisms with experimental evidence of hydrocarbon degradation were analyzed with the CANT-HYD HMMs for validation. For 62 out of 72 organisms, gene predictions using CANT-HYD were consistent with experimental data (**Figure 3**). Of the remaining 10 genomes, two genomes had hits with a score between the noise and trusted cutoffs, and eight genomes lacked hits above the noise cutoff. In a few instances, genomes were not available for the exact strain and a GTDB representative genome was used in their place. Although GTDB representatives share 95% average nucleotide identity with the cluster they represent, hydrocarbon degradation genes may be missing or different. Genomes of three organisms isolated on phenanthrene, chlorophenol and benzoate did not yield hits to any CANT-HYD HMMs. Although these substrates can be degraded by dioxygenases which share homology with mono- and polyaromatic ring hydroxylating dioxygenases, the lack of hits, even below the noise cutoff, indicates that the three organisms potentially use alternative metabolic pathways which were not covered by CANT-HYD. CANT-HYD predicted additional or unreported hydrocarbon substrate degradation capabilities for 16 genomes. For example, genes for toluene-2-monooxygenase (Tom) and toluene-4-monooxygenase (Tmo), and monoaromatic dioxygenase (MAH_alpha and MAH_beta) were found in the genome of *Pseudoxanthomonas spadix* BD-a59, a well-known benzene, toluene, ethylbenzene, and xylene (BTEX) degrader (Chun et al., 2010). Anaerobic hydrocarbon degradation genes were only detected in the genomes of anaerobes, further showing the prediction accuracy of CANT-HYD HMMs. Every HMM had at least one hit, except for the bacterial benzene carboxylase (AbcA_1) and toluene-benzene monooxygenase beta subunit (TmoB_BmoB). Anaerobic benzene degradation *via* benzene carboxylase (AbcA_1) has been identified in a single uncultured organism belonging to *Clostridia*, for which a genome is currently unavailable (Abu Laban et al., 2010). Toluene-benzene monooxygenase beta subunit (TmoB_BmoB) was found adjacent to TmoA_BmoA gene on the *Pseudoxanthomonas spadix* BD-a59 genome with a score above the noise cutoff, which indicates that it likely is a TmoB_BmoB gene divergent from the seed sequences that were used to make the HMM. Overall, these results show that CANT-HYD reliably identifies hydrocarbon degradation marker genes and can thus be used to predict the hydrocarbon degradation potential of genomes and in metagenomes.

Diversity of Hydrocarbon Degrading Bacteria and Archaea

A large number of bacterial (30,238) and archaeal (1,672) genomes from the Genome Taxonomy Database (GTDB) were searched against the CANT-HYD HMMs (Parks et al., 2020). In total, 4,601 representative genomes from 18 bacterial phyla, had at least one hit to an HMM that scored higher than the trusted cutoff (**Supplementary Table 2**), and in total, 5,845 genomes from 24 bacterial phyla had hits to at least one CANT-HYD HMM above the noise cutoff (**Figure 4**). HMM hits from diverse bacterial phyla demonstrate the widespread potential for hydrocarbon degradation across bacteria (**Figures 4A,B**).

Many of these phyla contain no cultured representatives, and therefore annotation tools like CANT-HYD become important for offering clues about their metabolic potential. For instance, the potential for hydrocarbon degradation was found in genomes of poorly represented phyla including Abyssubacteria, Tectomicrobia, and Eremiobacterota (**Figure 4B** and **Supplementary Figure 4**). The phylum Abyssubacteria, often found in association with subsurface and hydrocarbon contaminated environments (Momper et al., 2018), had hits to anaerobic alkane degradation (AhyA). Eremiobacterota (formerly WPS-2), previously found in hydrocarbon enrichment cultures (Ramadass et al., 2018), had three members with aerobic aromatic hydrocarbon degradation potential (NdoB, MAH_alpha, MAH_beta). Two genomes from Tectomicrobia had high HMM scores to enzymes responsible for the aerobic degradation of monoaromatics (MAH_Beta), polyaromatics (NdoB), and long-chain alkanes (LadA_alpha and CYP153). There is currently no literature associating Tectomicrobia with hydrocarbon containing environments, however, high confidence matches to CANT-HYD HMMs suggest that they may have a previously unidentified role in the aerobic metabolism of a range of hydrocarbons.

Hydrocarbon Degradation in Archaea

Archaea contained fewer hydrocarbon degradation genes compared to bacteria. Only four genomes, all from the phylum Halobacteriota, had HMM hits above the trusted cutoff. Another 102 genomes, also from Halobacteriota, had at least one HMM hit

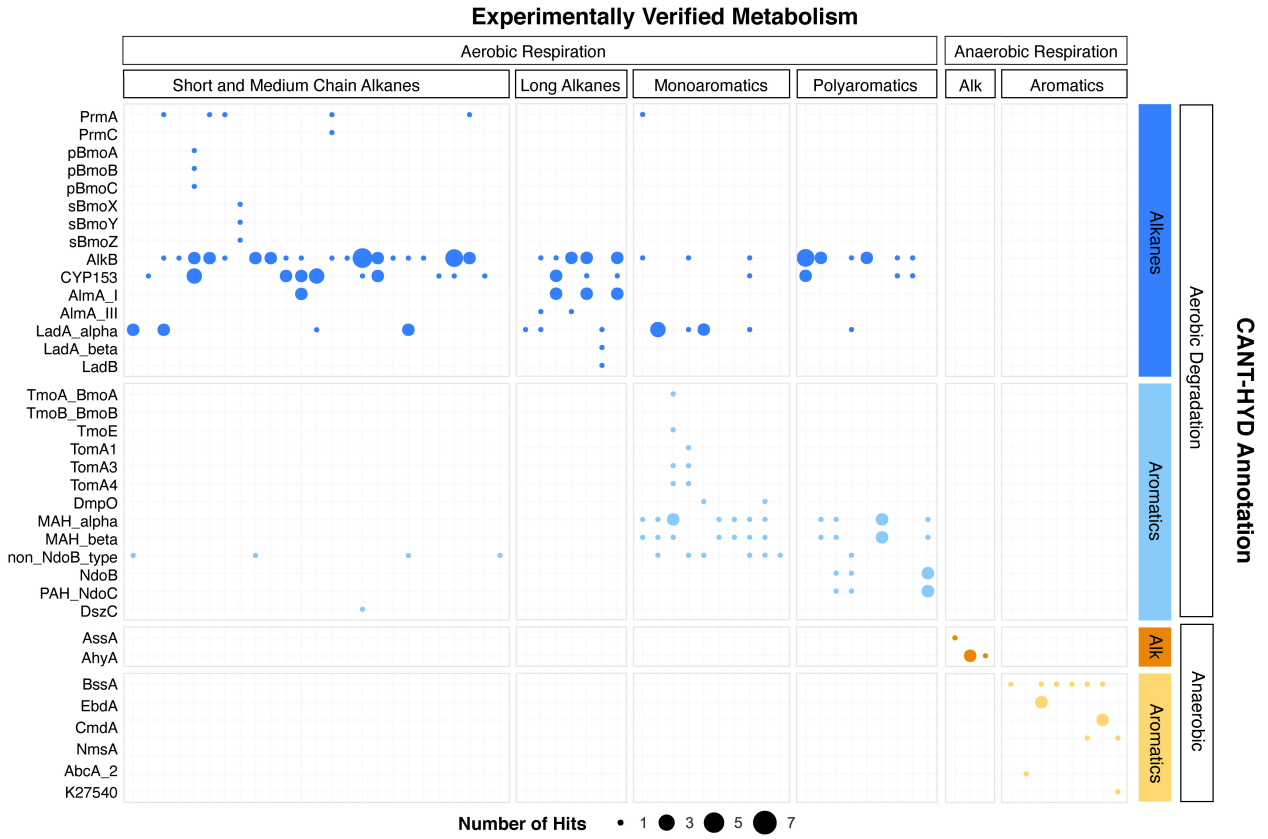

FIGURE 3 | Concordance of CANT-HYD annotations of genomes and their experimentally verified hydrocarbon degradation activity. Each bubble represents hits above the confidence threshold from experientially verified hydrocarbon degrading genomes (x-axis) plotted against the HMMs of the CANT-HYD database (y-axis). The size of the bubble represents number of unique hits in the genome. The genomes (x-axis) and the CANT-HYD HMMs (y-axis) are organized by the hydrocarbon substrate and respiration. A complete list of isolate genomes, their Genbank accession, and their published hydrocarbon degradation capability can be found in **Supplementary Table 3**.

above the noise cutoff (**Supplementary Figure 3**). The phylum Halobacteriota (formerly a member of phylum Euryarchaeota) is known to contain halophilic hydrocarbon degrading species (Al-Mailem et al., 2010; Oren, 2019). The identification of only a few archaeal hydrocarbon degraders may be due to either a lower representation of sequenced archaeal genomes, or an increased phylogenetic distance of archaeal hydrocarbon degradation genes to the primarily bacterial sequences that have been experimentally validated. Additionally, methanotrophy, the most well studied archaeal hydrocarbon degradation, is not covered by CANT-HYD. As more experimental evidence of archaeal genes emerge, the annotation of archaeal hydrocarbon degradation will improve.

Cyanobacteria as Alkane Degraders

Thirty-six genes from 29 cyanobacterial GTDB representative genomes, mostly from the family Nostocaceae and the genera *Nostoc* and *Aulosira*, were predicted to contain LadA beta, a long-chain alkane monooxygenase (**Figure 5**). LadA beta is one of the three LadA-type long-chain alkane monooxygenase enzymes with experimental evidence of long-chain alkane degradation (Boonmak et al., 2014). Phylogenetically, these cyanobacterial genes were related to the experimentally verified LadA beta sequence from *Geobacillus thermoleovorans* (BAM76372.1), suggesting that the genes perform a similar role in their photosynthetic hosts (**Figure 5A**). Many cyanobacterial species produce long-chain alkanes, potentially at globally relevant levels (Lea-Smith et al., 2015; Love et al., 2021), and alkane degradation has been observed in microbial communities with abundant cyanobacteria (Abed, 2010). Thus far however, it has been inconclusive whether the alkane degradation is performed by cyanobacteria or other heterotrophic community members, and if the cyanobacteria are responsible, which degradation pathways they utilize (Al-Hasan et al., 1998; Qiao et al., 2020). Strong hits to LadA beta suggest that some cyanobacterial species have the metabolic potential for long-chain alkane degradation *via* LadA, although experiments are needed to confirm if this genetic potential is realized.

Hydrocarbon Degradation in Diverse Environments

The CANT-HYD HMMs were used to search for hydrocarbon degradation potential in 41 metagenomes, representing diverse environments including hydrocarbon degrading enrichment

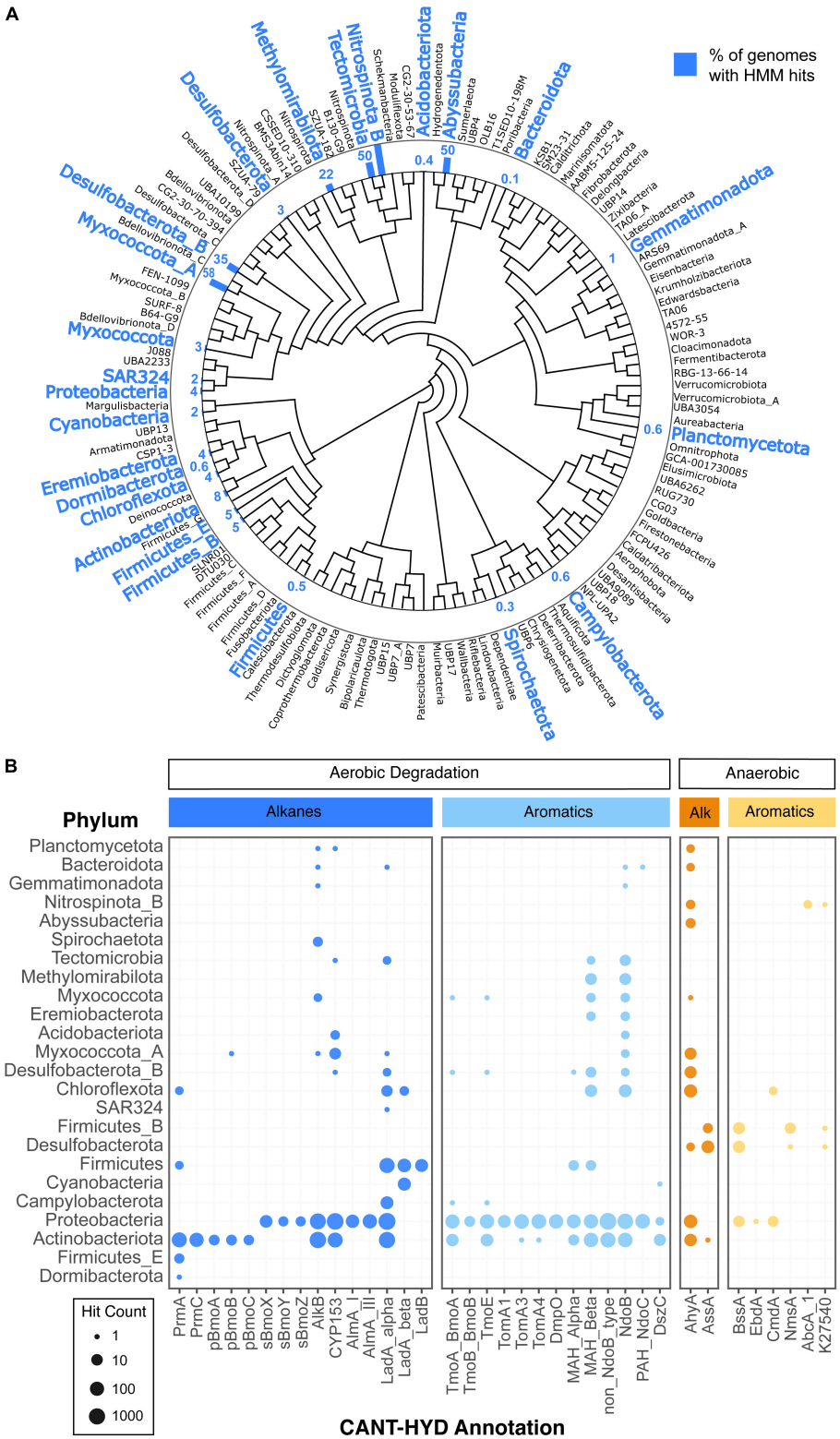

FIGURE 4 | Diversity of hydrocarbon-degrading bacteria and archaea. GTDB representative genomes with CANT-HYD HMM hits shown on panel **(A)** the GTDB phylogenomic tree collapsed at phylum level. Phyla with genomes containing genes with a CANT-HYD HMM score greater than the noise cutoff are shown in blue. The corresponding bar and number indicates the percentage of the representative genomes in that phylum containing at least one hit to a CANT-HYD HMM. For example, Abyssubacteria contained two genome representatives in GTDB, and one of the genomes contained a high confidence match to a CANT-HYD HMM; thus, the bar shows 50%. **(B)** Distribution and number of hits for each phylum in GTDB across the CANT-HYD HMMs.

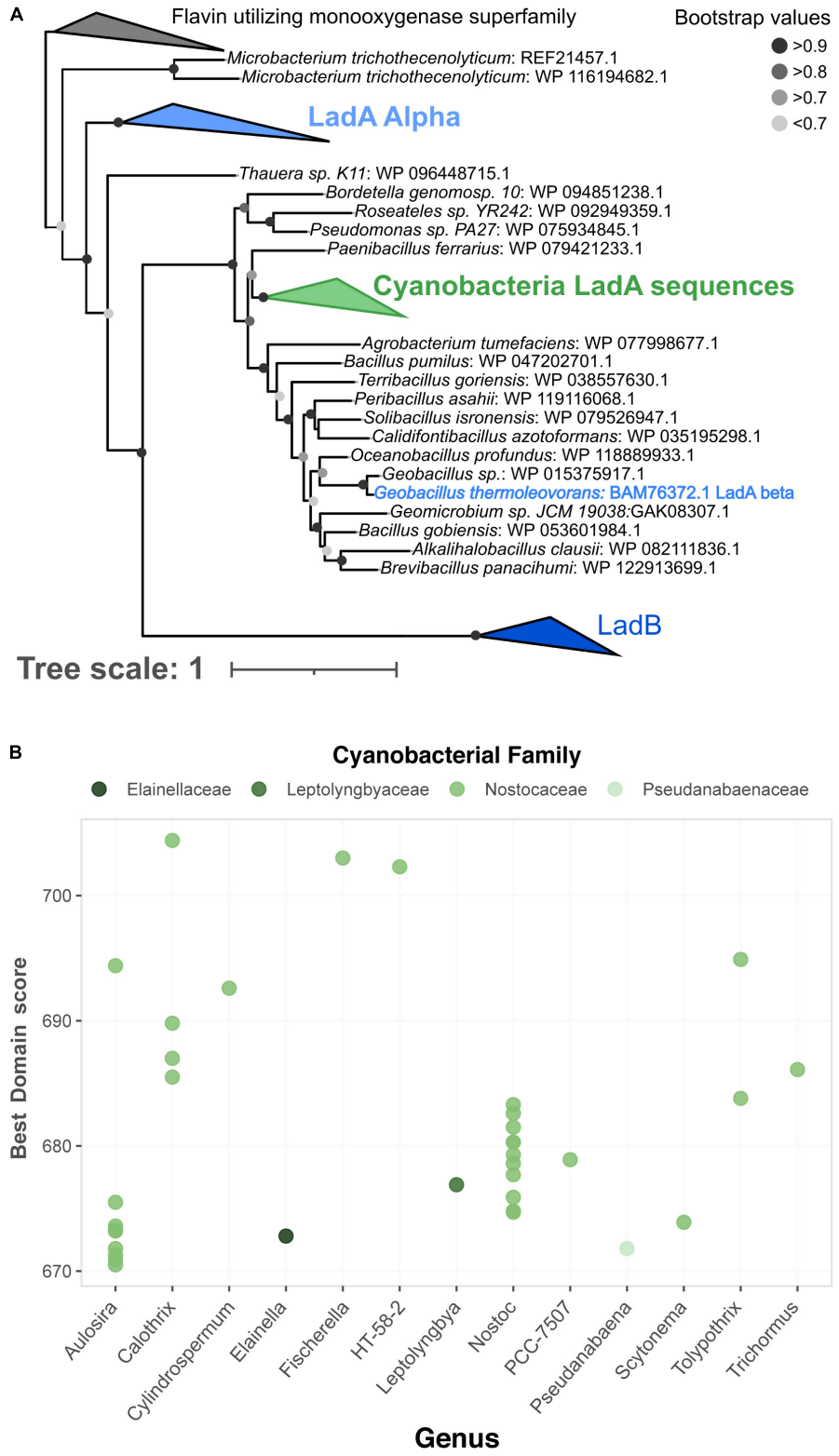

FIGURE 5 | Detection of long-chain alkane monooxygenase (Lad) family of genes in Cyanobacteria. **(A)** Phylogenetic tree of Lad sequences including a cluster of 102 sequences detected in Cyanobacteria genomes by CANT-HYD (green). **(B)** Distribution of scores for LadA beta HMM hits in Cyanobacteria genomes. Original tree file for is available as **Supplementary Data Sheet 1**.

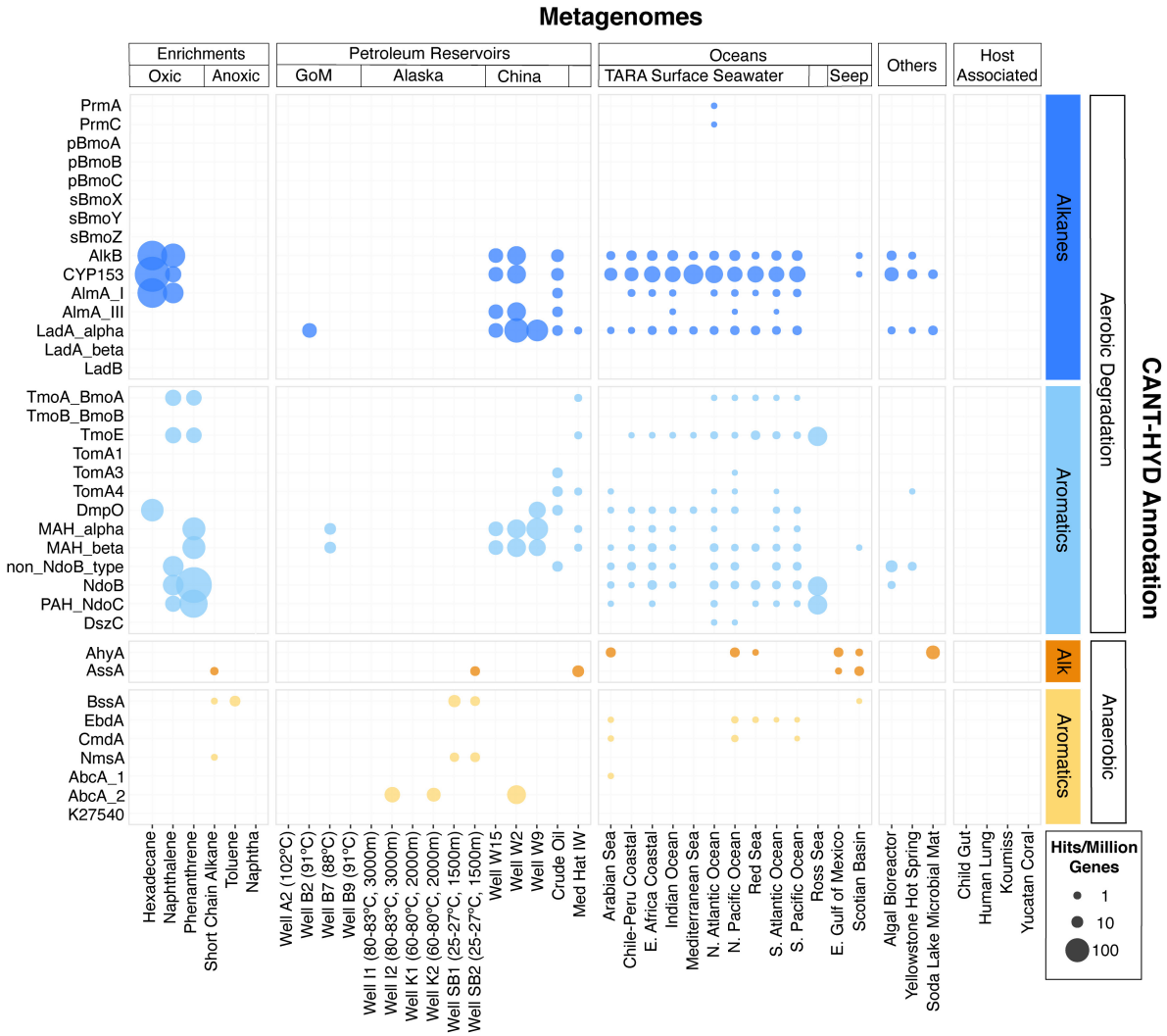

FIGURE 6 | Metagenomes from diverse environments analyzed with the CANT-HYD HMMs. CANT-HYD HMM hits (scores ≥ noise cutoff) (y-axis) to each metagenome (x-axis) normalized to per million protein coding genes. The size of the bubble depicts proportion of hydrocarbon degrading genes to total genic content of the metagenome. The metagenomes (x-axis) are grouped under Enrichments (hydrocarbon-degrading enrichment cultures), Petroleum Reservoirs (produced well water from petroleum reservoirs from Alaska, Gulf of Mexico, Jiangsu and Qinghai, China, and Medicine Hat, Alberta), Oceans (cold seeps, TARA surface seawater), Host-Assoc (host-associated microbiomes), Other: (other environments). The CAN-HYD HMMs are grouped by hydrocarbon substrate (alkane and aromatic) and respiration (aerobic and anaerobic). See **Supplementary Table 4** for details regarding these metagenomes.

cultures, petroleum reservoirs, oceans, host-associated microbiomes, alkaline lakes, and hot springs. Hydrocarbon degradation genes were detected in all these environments, except for the host-associated microbiomes, which are presumed to have a limited presence of hydrocarbons (**Figure 6**).

When normalized for total gene content, the highest proportion of hydrocarbon degradation genes were detected in the metagenomes of hydrocarbon-degrading enrichments. Genes for aerobic alkane degradation, such as AlkB, CYP153, and AlmA, were the most widely detected in this dataset. Many marker genes for aerobic hydrocarbon degradation were found ubiquitously in metagenomes from ocean surface waters, while genes for anaerobic hydrocarbon degradation were largely detected in metagenomes sequenced from anoxic habitats such as petroleum reservoirs, subseafloor sediments, and anoxic hydrocarbon degrading microcosms. Further, some degradation enzymes covered by the CANT-HYD HMMs yielded no hits, namely butane monooxygenases (pBmoA, pBmoB, and pBmoC, and sBmoX, sBmoY, and sBmoZ), and benzene and naphthalene carboxylases (AbcA_1 and K27540). These HMMs were made with less than five seed sequences as they had only a few close relatives (≥50% sequence identity) in the public nucleotide database at the time of this work.

Enrichment Cultures of Hydrocarbon Degrading Microorganisms

Genes from the glycyl radical enzyme family (Ass/Bss/Nms) for anaerobic hydrocarbon degradation were identified in two

out of the three anoxic cultures, namely the toluene and short-chain-alkane enrichments (Tan et al., 2015; **Figure 6**). The number of genes identified in this study were half of those reported in the original, which were also annotated using custom HMMs (**Table 1**). Observed discrepancies include the detection of partial gene sequences in the metagenomic assemblies (**Table 1**: denoted with an *) (Tan et al., 2015). Partial gene sequence matches to the CANT-HYD Ass/Bss/Nms HMMs scored below the noise cutoffs and did not pass the threshold. Partial hits to HMMs scoring below the noise cutoffs cannot be reliably annotated as they may be a partial sequence of a related gene that encodes an enzyme with a different function. In this case, the partial hits for genes encoding for glycyl radical enzymes (Ass/Bss/Nms) can be functionally similar (**Supplementary Figure 1**) to and share high sequence homology with pyruvate formate lyase. While CANT-HYD will not perform optimally with unassembled, partially or poorly assembled data, the noise and trusted cutoffs are recommended for a low false-positive rate, by filtering out partial genes that risk being misannotated.

Several genes for aerobic hydrocarbon degradation were identified in oxic cultures inoculated with the Deepwater Horizon oil plume and enriched on hexadecane, naphthalene, and phenanthrene as hydrocarbon substrates (**Figure 6**). Genes for the aerobic degradation of medium and long-chain alkanes (CYP153, AlkB, and AlmA_GroupI) were detected in the hexadecane and naphthalene enrichments, while a variety of genes for aromatic hydrocarbon degradation were detected in the oxic naphthalene and phenanthrene enrichment cultures.

Petroleum Reservoirs and Hydrocarbon Biodegradation

Multiple marker genes were identified in all metagenomes from petroleum reservoirs, except for four reservoirs that were either at a high temperature (>80°C) or were deep subsurface (**Figure 6**). Produced water metagenomes from petroleum reservoirs in Alaska (wells SB1, SB2, K1, K2, I1, and I2) were found to contain only anaerobic hydrocarbon degradation genes in agreement with the associated study (Hu et al., 2016). CANT-HYD further detected the presence of a putative benzene carboxylase gene (AbcA_2) in the metagenomes from two of the oil wells (K2 and I2), which was not reported in the original study. These oil reservoirs were reported to contain complete and partial genomes of a sulfide-producing archaeon, *Archaeoglobus,* and our results indicate that it can potentially metabolize benzene anaerobically.

TABLE 1 | Number of AssA, BssA, and NmsA genes detected in this analysis compared to the original study (Tan et al., 2015).

Substrate	Short-chain alkanes		Toluene		Naphtha	
Genes	**Original study**	**CANT-HYD**	**Original study**	**CANT-HYD**	**Original study**	**CANT-HYD**
AssA	4 + 1*	4	1*	–	1*	–
BssA	3	1	1	1	1 + 2*	–
NmsA	1	1	–	–	1*	–

Asterisk (*) indicates partial gene homologs.

Similar observation of an *Archaeoglobus* metagenome-assembled-genome (MAG) and its association with a benzene carboxylase (AbcA_2) gene comes from the metagenome of well W2 from the Jiangsu Oil Reservoir, China (Liu et al., 2018). Although the original study found alkyl succinate synthase genes (Ass) in an *Archaeoglobus* MAG, in this study, genes for the glycyl radical family of enzymes from these metagenomes and the *Archaeoblogus* genome scored below the noise cutoff. Any hits below the noise cutoff cannot be reliably annotated automatically and therefore require manual curation, such as using gene phylogeny to differentiate between glycyl radical enzyme and pyruvate formate lyase. A study by Khelifi et al. (2014) also shows evidence for anaerobic long-chain alkane degradation by *Archaeoglobus*, however, the genes identified as responsible for this metabolism share low sequence homology with bacterial alkyl succinate synthase alpha subunit and will not be annotated by the CANT-HYD AssA HMM. Therefore, a separate HMM for archaeal alkyl succinate synthase alpha subunit would be required when strong experimental evidence for these genes becomes available. Several genes for aerobic alkane and monoaromatic hydrocarbon degradation (**Figure 6**) were also identified in the Well W2 metagenome, which were not originally reported. These findings highlight the utility of CANT-HYD, which can search for a comprehensive suite of hydrocarbon degradation markers, independent of *a priori* knowledge of the system.

Widespread Hydrocarbon Degradation Potential in Global Surface Seawaters

Widespread potential for aerobic hydrocarbon degradation was detected in the surface seawater metagenomes collected by the TARA Oceans survey (Karsenti et al., 2011; Tully et al., 2018) and other studies (Orellana et al., 2017). Predicted hydrocarbon metabolism was largely driven by medium- and long-chain alkane hydroxylases, and ring-hydroxylating dioxygenases. This pervasive metabolic capability in the global ocean surface could be the result of biogenic alkanes synthesized by cyanobacteria (Lea-Smith et al., 2015) or the accumulation of polyaromatic hydrocarbons by other ocean phytoplankton, resulting in a "cryptic hydrocarbon cycle" (Binark et al., 2000; Love et al., 2021). An ocean metagenome from the Ross Sea, Antarctica had an exceptionally high abundance of marker genes involved in polyaromatic hydrocarbon degradation. The Ross Sea is well-known for its seasonal algal blooms and rapid carbon turnover (Smith et al., 2000; Peloquin and Smith, 2007), which have been associated with biogenic alkanes, polyaromatic hydrocarbons, and PAH degraders (Gutierrez et al., 2011; Lea-Smith et al., 2015; Love et al., 2021). Overall, our findings support the recent experimental evidence of a marine hydrocarbon cycle (Love et al., 2021).

Comparison of CANT-HYD HMMs to Existing HMM Databases

The annotation performance of CANT-HYD HMMs was compared to HMMs from Pfam, KO, and eggNOG databases. As the purpose-built CANT-HYD database is curated for hydrocarbon degradation genes, it assigned more specific annotations than Pfam, eggNOG and KO databases. Pfam and

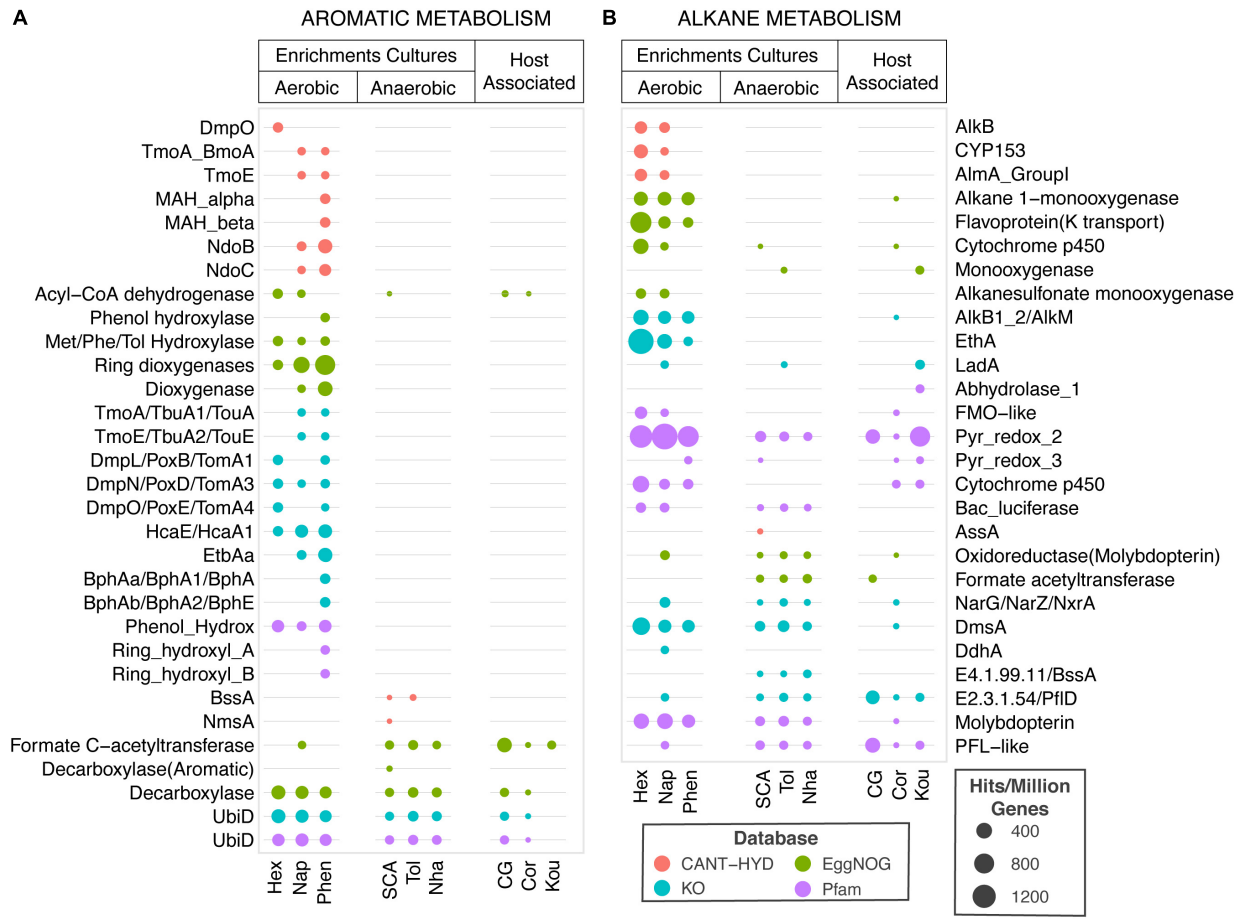

FIGURE 7 | Comparison of hydrocarbon metabolism prediction in metagenomes. Hits to aromatic hydrocarbon **(A)** and alkane **(B)** metabolism HMMs (y-axis) from CANT-HYD (red), eggNOG (green), KO (teal), and Pfam (purple) to each metagenome (x-axis) normalized to per million coding genes. The metagenomes (x-axis) are grouped under Enrichment Cultures or Host associated microbiome. The enrichment culture metagenomes representing microbial community grown on under aerobic conditions on hexadecane (Hex), naphthalene (Nap), or phenanthrene (Phen) or under anaerobic conditions on short-chain alkanes (SCA), Toluene (Tol), or Naphtha (Nha). The host-associated microbiomes are from child gut (CG), Coral (Cor), or Koumiss (Kou). The size of the bubble depicts proportion of hydrocarbon degrading genes to total genic content of the metagenome. See **Supplementary Table 5** for details regarding the HMMs and the underlying data.

eggNOG often annotated genes as the broad protein family to which the hydrocarbon degradation genes belong. Examples of these generic descriptions included "monooxygenase" for long-chain alkane monooxygenase (LadA), and "pyruvate-formate-lyase like protein" for the AssA, BssA, and NmsA genes (**Supplementary Table 5**). While these descriptions are broadly accurate, they do not give sufficient information about the gene for strong functional inference. The KO database assigned more specific annotations of the enzyme function compared to eggNOG and Pfam, but not necessarily to the same level of substrate specificity as the CANT-HYD HMMs. The KO database also missed annotations to more recently discovered hydrocarbon degradation genes such as naphthalene carboxylase (K27540), and long-chain alkane monooxygenase beta (LadA beta) (**Supplementary Table 5**).

The equivalent HMMs from the three databases were compared with CANT-HYD HMMs using genomes of 62 experimentally verified hydrocarbon degrading isolates and 10 metagenomes from hydrocarbon enrichments and host associated microbiomes. This search resulted in the 1000s of hits to from the Pfam, eggNOG and KO database and hence to remove spurious matches and increase confidence in annotations, only hits with an e-value below 10^{-50} were retained for analysis (**Figure 7** and **Supplementary Table 5**). After filtering, Pfam, eggNOG and KO still reported >1,000 hits for 62 genomes, while the total number of genes identified by CANT-HYD above the trusted cutoff was only 190. Furthermore, as previously mentioned that not all HMMs of expected specific metabolisms were found in these public databases, a large proportion of hits were to HMMs describing broad protein families such as "cytochrome p450," "Pyr_redox_2," and "monooxygenase." A search of the metagenomes resulted in a total of 925, 525, and 463 genes identified by Pfam, KO, and eggNOG HMMs, respectively. In contrast, only 50 genes were identified in the same dataset using the noise cutoff of the CANT-HYD HMMs. These discrepancies in identified genes in genomes

and metagenomes originates from the differences in the HMM targets of the databases. The lack of specificity of Pfam, eggNOG, and some KO is likely to detect genes that are related to hydrocarbon degradation genes, but not involved in hydrocarbon degradation. For the same reasons, while all other databases produced hits in host-associated metagenomes, CANT-HYD did not (**Figure 7**). Together, these comparisons highlight the usefulness and accuracy of CANT-HYD to identify and annotate specific hydrocarbon metabolic potential by using curated cutoffs for HMMs designed specifically for hydrocarbon degradation marker genes.

CONCLUSION

Here, we describe CANT-HYD, an HMM database of marker genes for hydrocarbon degradation. These phylogenetically informed HMMs accurately identify over 37 genes relevant to aerobic and anaerobic metabolisms of aliphatic and aromatic hydrocarbons in genomes and metagenomes. Each CANT-HYD HMM includes a manually curated trusted and noise cutoff score for automated reliable detection of these hydrocarbon degradation marker genes. To the best of our knowledge, CANT-HYD is the first dedicated tool for annotation of hydrocarbon degradation genes in genomes and metagenomes. We demonstrate the use of CANT-HYD as an exploratory tool by surveying all genomes in GTDB (30,238 bacterial and 1,672 archaeal), as well as several large metagenomic datasets. We uncovered the potential for long-chain alkane degradation in some cyanobacterial genomes and identified widespread potential for aerobic hydrocarbon degradation in global ocean surface waters, supporting a recently discovered marine hydrocarbon cycle. The comparison to other publicly available HMMs highlights the need for a curated HMM database for specific and precise annotation of hydrocarbon degradation genes and large-scale detection of hydrocarbon degrading capabilities in genomes and metagenomes.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

VK, JZ, DAG, AC, EB, MAB, AJP, AKH, BN, and SB carried out the literature review and sequence data searching. VK, JZ, DAG, AC, EB, MAB, AJP, and SB made and validated the HMMs. VK, JZ, DAG, AC, EB, and SB wrote the manuscript. VK, JZ, DAG, AC, MAB, and SB performed the genomic and metagenomic data analyses. VK, JZ, AC, and SB made the figures. BN and MS conceived the study. DAG created and maintained the GitHub archive. All authors contributed toward methodology development and editing and reviewing the manuscript.

FUNDING

This work was supported by funds from Genome Canada to *GENICE—The microbial genomics for oil spill preparedness in the Canadian Arctic*, to CRJH and MS. We acknowledge support from the Canada First Research Excellence Fund to BN and AJP and from the Government of Alberta to VK, JZ, and MS. Additional support for JZ, AJP, and AKH was provided by Natural Sciences and Engineering Research Council of Canada (NSERC).

ACKNOWLEDGMENTS

We would like to thank Lisa Gieg, Gerrit Voordouw, and Muhe Diao for providing valuable insights into hydrocarbon metabolism. We would also like to thank Dongshan An, Gerrit Voordouw, Daniel Colman, Eric Boyd, Monica Orellana, Viridiana Avila-Magaña, and Monica Medina for some of the metagenomic data analyzed here. We would like to extend our thanks to Xiaoli Dong for help with access and use of the computational servers. We thank the members of Energy Bioengineering and Geomicrobiology group and countless other people for their moral support during this "hackathon" that lasted more than a year.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.764058/full#supplementary-material

REFERENCES

Abed, R. M. M. (2010). Interaction between cyanobacteria and aerobic heterotrophic bacteria in the degradation of hydrocarbons. *Int. Biodeterior. Biodegradation* 64, 58–64. doi: 10.1016/j.ibiod.2009.10.008

Abu Laban, N., Selesi, D., Rattei, T., Tischler, P., and Meckenstock, R. U. (2010). Identification of enzymes involved in anaerobic benzene degradation by a strictly anaerobic iron-reducing enrichment culture. *Environ. Microbiol.* 12, 2783–2796. doi: 10.1111/j.1462-2920.2010.02248.x

Al-Hasan, R. H., Al-Bader, D. A., Sorkhoh, N. A., and Radwan, S. S. (1998). Evidence for n-alkane consumption and oxidation by filamentous cyanobacteria from oil-contaminated coasts of the Arabian Gulf. *Mar. Biol.* 130, 521–527. doi: 10.1007/s002270050272

Al-Mailem, D. M., Sorkhoh, N. A., Al-Awadhi, H., Eliyas, M., and Radwan, S. S. (2010). Biodegradation of crude oil and pure hydrocarbons by extreme halophilic archaea from hypersaline coasts of the Arabian Gulf. *Extremophiles* 14, 321–328. doi: 10.1007/s00792-010-0312-9

Atlas, R. M., and Hazen, T. C. (2011). Oil biodegradation and bioremediation: a tale of the two worst spills in U.S. history. *Environ. Sci. Technol.* 45, 6709–6715. doi: 10.1021/es2013227

Avila-Magaña, V., Kamel, B., DeSalvo, M., Gómez-Campo, K., Enríquez, S., Kitano, H., et al. (2021). Elucidating gene expression adaptation of phylogenetically divergent coral holobionts under heat stress. *Nat. Commun.* 12:5731. doi: 10.1038/S41467-021-25950-4

Bateman, A., Birney, E., Durbin, R., Eddy, S., Howe, K., and Sonnhammer, E. (2000). The Pfam protein families database. *Nucleic Acids Res.* 28, 263–266. doi: 10.1093/nar/28.1.263

Binark, N., Güven, K. C., Gezgin, T., and Ünlü, S. (2000). Oil pollution of marine algae. *Bull. Environ. Contam. Toxicol.* 64, 866–872. doi: 10.1007/s0012800083

Boonmak, C., Takahashi, Y., and Morikawa, M. (2014). Cloning and expression of three ladA-type alkane monooxygenase genes from an extremely thermophilic alkane-degrading bacterium Geobacillus thermoleovorans B23. *Extremophiles* 18, 515–523. doi: 10.1007/s00792-014-0636-y

Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Callaghan, A. V., Wawrik, B., Ní Chadhain, S. M., Young, L. Y., and Zylstra, G. J. (2008). Anaerobic alkane-degrading strain AK-01 contains two alkylsuccinate synthase genes. *Biochem. Biophys. Res. Commun.* 366, 142–148. doi: 10.1016/j.bbrc.2007.11.094

Cantalapiedra, C. P., Hern Andez-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* 1:msab293. doi: 10.1093/MOLBEV/MSAB293

Christman, G. D., León-Zayas, R. I., Zhao, R., Summers, Z. M., and Biddle, J. F. (2020). Novel clostridial lineages recovered from metagenomes of a hot oil reservoir. *Sci. Rep*. 10:8048. doi: 10.1038/s41598-020-64904-6

Chun, J., Kim, K., Lee, J. H., and Choi, Y. (2010). The analysis of oral microbial communities of wild-type and toll-like receptor 2-deficient mice using a 454 GS FLX Titanium pyrosequencer. *BMC Microbiol.* 10:101. doi: 10.1186/1471-2180-10-101

Dombrowski, N., Donaho, J. A., Gutierrez, T., Seitz, K. W., Teske, A. P., and Baker, B. J. (2016). Reconstructing metabolic pathways of hydrocarbon-degrading bacteria from the Deepwater Horizon oil spill. *Nat. Microbiol.* 1, 1–7. doi: 10.1038/nmicrobiol.2016.57

Dong, X., Greening, C., Rattray, J. E., Chakraborty, A., Chuvochina, M., Mayumi, D., et al. (2019). Metabolic potential of uncultured bacteria and archaea associated with petroleum seepage in deep-sea sediments. *Nat. Commun.* 10, 1–12. doi: 10.1038/s41467-019-09747-0

Dong, X., Rattray, J. E., Campbell, D. C., Webb, J., Chakraborty, A., Adebayo, O., et al. (2020). Thermogenic hydrocarbon biodegradation by diverse depth-stratified microbial populations at a Scotian Basin cold seep. *Nat. Commun.* 11:5825. doi: 10.1038/s41467-020-19648-2

Eddy, S. R. (2004). What is a hidden markov model? *Nat. Biotechnol.* 22, 1315–1316. doi: 10.1038/nbt1004-1315

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.* 7:1002195. doi: 10.1371/journal.pcbi.1002195

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Feigelman, R., Kahlert, C. R., Baty, F., Rassouli, F., Kleiner, R. L., Kohler, P., et al. (2017). Sputum DNA sequencing in cystic fibrosis: non-invasive access to the lung microbiome and to pathogen details. *Microbiome* 5:20. doi: 10.1186/s40168-017-0234-1

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

González-Gaya, B., Martínez-Varela, A., Vila-Costa, M., Casal, P., Cerro-Gálvez, E., Berrojalbiz, N., et al. (2019). Biodegradation as an important sink of aromatic hydrocarbons in the oceans. *Nat. Geosci.* 12, 119–125. doi: 10.1038/s41561-018-0285-3

Gregson, B. H., Metodieva, G., Metodiev, M. V., Golyshin, P. N., and McKew, B. A. (2018). Differential protein expression during growth on medium versus long-chain alkanes in the obligate marine hydrocarbon-degrading bacterium *Thalassolituus oleivorans* MIL-1. *Front. Microbiol.* 9:3130. doi: 10.3389/fmicb.2018.03130

Gutierrez, T., Singleton, D. R., Aitken, M. D., and Semple, K. T. (2011). Stable isotope probing of an algal bloom to identify uncultivated members of the *Rhodobacteraceae* associated with low-molecular-weight polycyclic aromatic hydrocarbon degradation. *Appl. Environ. Microbiol.* 77, 7856–7860. doi: 10.1128/AEM.06200-11

Herman, D. R., Rhoades, N., Mercado, J., Argueta, P., Lopez, U., and Flores, G. E. (2020). Dietary habits of 2- to 9-year-old american children are associated with gut microbiome composition. *J. Acad. Nutr. Diet.* 120, 517–534. doi: 10.1016/j.jand.2019.07.024

Hu, P., Tom, L., Singh, A., Thomas, B. C., Baker, B. J., Piceno, Y. M., et al. (2016). Genome-resolved metagenomic analysis reveals roles for candidate phyla and other microbial community members in biogeochemical transformations in oil reservoirs. *mBio* 7:e01669-15. doi: 10.1128/mBio.01669-15

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi: 10.1093/NAR/GKY1085

Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M., and Rupp, R. (2007). Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinform.* 8:460. doi: 10.1186/1471-2105-8-460

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* 11:119. doi: 10.1186/1471-2105-11-119

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi: 10.1093/NAR/GKV1070

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., et al. (2011). A holistic approach to marine eco-systems biology. *PLoS Biol.* 9:e1001177. doi: 10.1371/journal.pbio.1001177

Khelifi, N., Amin Ali, O., Roche, P., Grossi, V., Brochier-Armanet, C., Valette, O., et al. (2014). Anaerobic oxidation of long-chain n-alkanes by the hyperthermophilic sulfate-reducing archaeon, *Archaeoglobus fulgidus*. *ISME J.* 8, 2153–2166. doi: 10.1038/ismej.2014.58

Kniemeyer, O., Musat, F., Sievert, S. M., Knittel, K., Wilkes, H., Blumenberg, M., et al. (2007). Anaerobic oxidation of short-chain hydrocarbons by marine sulphate-reducing bacteria. *Nature* 449, 898–901. doi: 10.1038/nature06200

Lea-Smith, D. J., Biller, S. J., Davey, M. P., Cotton, C. A. R., Sepulveda, B. M. P., Turchyn, A. V., et al. (2015). Contribution of cyanobacterial alkane production to the ocean hydrocarbon cycle. *Proc. Natl. Acad. Sci. U.S.A.* 112, 13591–13596. doi: 10.1073/pnas.1507274112

Lerdau, M., Guenther, A., and Monson, R. (1997). Plant production and emission of volatile organic compounds: plant-produced hydrocarbons influence not only the plant itself but the atmosphere a well. *BioScience* 47, 373–383. doi: 10.2307/1313152

Letunic, I., and Bork, P. (2019). Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259. doi: 10.1093/nar/gkz239

Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly *via* succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033

Li, Y. P., Pan, J. C., and Ma, Y. L. (2020). Elucidation of multiple alkane hydroxylase systems in biodegradation of crude oil n-alkane pollution by *Pseudomonas aeruginosa* DN1. *J. Appl. Microbiol.* 128, 151–160. doi: 10.1111/jam.14470

Liu, Y. F., Galzerani, D. D., Mbadinga, S. M., Zaramela, L. S., Gu, J. D., Mu, B. Z., et al. (2018). Metabolic capability and in situ activity of microorganisms in an oil reservoir. *Microbiome* 6:5. doi: 10.1186/s40168-017-0392-1

Liu, Y. F., Qi, Z. Z., Shou, L. B., Liu, J. F., Yang, S. Z., Gu, J. D., et al. (2019). Anaerobic hydrocarbon degradation in candidate phylum 'Atribacteria' (JS1) inferred from genomics. *ISME J.* 13, 2377–2390. doi: 10.1038/s41396-019-0448-2

Love, C. R., Arrington, E. C., Gosselin, K. M., Reddy, C. M., Van Mooy, B. A. S., Nelson, R. K., et al. (2021). Microbial production and consumption of hydrocarbons in the global ocean. *Nat. Microbiol.* 6, 489–498. doi: 10.1038/s41564-020-00859-8

McGinnis, S., and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 32, 20–25. doi: 10.1093/nar/gkh435

Momper, L., Aronson, H. S., and Amend, J. P. (2018). Genomic description of 'candidatus abyssubacteria,' a novel subsurface lineage within the candidate

phylum hydrogenedentes. *Front. Microbiol.* 9:1993. doi: 10.3389/fmicb.2018.01993

Nie, Y., Zhao, J.-Y., Tang, Y.-Q., Guo, P., Yang, Y., Wu, X.-L., et al. (2016). Species divergence vs. functional convergence characterizes crude oil microbial community assembly. *Front. Microbiol.* 7:1254. doi: 10.3389/fmicb.2016.01254

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189

Orellana, M. V., López-García de Lomana, A., Jennings, M. K., Lee, A., Hansman, R. L., Thompson, A. W., et al. (2017). *On the Influence of Diatom Programmed Cell Death on Carbon Export in the Ross Sea*. Honolulu: American Fisheries Society.

Oren, A. (2019). "Aerobic hydrocarbon-degrading archaea," in *Taxonomy, Genomics and Ecophysiology of Hydrocarbon-Degrading Microbes*, ed. T. J. McGenity (Berlin: Springer International Publishing), 41–51.

Parks, D. H., Chuvochina, M., Chaumeil, P. A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* 38, 1079–1086. doi: 10.1038/s41587-020-0501-8

Peloquin, J. A., and Smith, W. O. (2007). Phytoplankton blooms in the Ross Sea, Antarctica: interannual variability in magnitude, temporal patterns, and composition. *J. Geophys. Res. Oceans* 112:C08013. doi: 10.1029/2006JC003816

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e0009490. doi: 10.1371/journal.pone.0009490

Prier, C. K., and Kosjek, B. (2019). Recent preparative applications of redox enzymes. *Curr. Opin. Chem. Biol.* 49, 105–112. doi: 10.1016/j.cbpa.2018.11.011

Qiao, Y., Wang, W., and Lu, X. (2020). High light induced alka(e)ne biodegradation for lipid and redox homeostasis in cyanobacteria. *Front. Microbiol.* 11:1659. doi: 10.3389/fmicb.2020.01659

Rabus, R., Boll, M., Heider, J., Meckenstock, R. U., Buckel, W., Einsle, O., et al. (2016). Anaerobic microbial degradation of hydrocarbons: from enzymatic reactions to the environment. *J. Mol. Microbiol. Biotechnol.* 26, 5–28. doi: 10.1159/000443997

Ramadass, K., Megharaj, M., Venkateswarlu, K., and Naidu, R. (2018). Bioavailability of weathered hydrocarbons in engine oil-contaminated soil: impact of bioaugmentation mediated by *Pseudomonas* spp. on bioremediation. *Sci. Total Environ.* 636, 968–974. doi: 10.1016/j.scitotenv.2018.04.379

Rueter, P., Rabus, R., Wilkest, H., Aeckersberg, F., Rainey, F. A., Jannasch, H. W., et al. (1994). Anaerobic oxidation of hydrocarbons in crude oil by new types of sulphate-reducing bacteria. *Nature* 372, 455–458. doi: 10.1038/372455a0

Schneiker, S., Dos Santos, V. A. P. M., Bartels, D., Bekel, T., Brecht, M., Buhrmester, J., et al. (2006). Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium *Alcanivorax borkumensis*. *Nat. Biotechnol.* 24, 997–1004. doi: 10.1038/nbt1232

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539. doi: 10.1038/msb.2011.75

Smith, W. O., Marra, J., Hiscock, M. R., and Barber, R. T. (2000). The seasonal cycle of phytoplankton biomass and primary productivity in the Ross Sea, Antarctica. *Deep Sea Res. II Top. Stud. Oceanogr.* 47, 3119–3140. doi: 10.1016/S0967-0645(00)00061-8

Tan, B., Jane Fowler, S., Laban, N. A., Dong, X., Sensen, C. W., Foght, J., et al. (2015). Comparative analysis of metagenomes from three methanogenic hydrocarbon-degrading enrichment cultures with 41 environmental samples. *ISME J.* 9, 2028–2045. doi: 10.1038/ismej.2015.22

Tornabene, T. G., Kates, M., Gelpi, E., and Oro, J. (1969). Occurrence of squalene, di- and tetrahydrosqualenes, and vitamin MK8 in an extremely halophilic bacterium, *Halobacterium cutirubrun*. *J. Lipid Res.* 10, 294–303. doi: 10.1016/s0022-2275(20)43087-1

Tully, B. J., Graham, E. D., and Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* 5:203. doi: 10.1038/sdata.2017.203

Wang, W., and Shao, Z. (2014). The long-chain alkane metabolism network of *Alcanivorax dieselolei*. *Nat. Commun.* 5, 1–11. doi: 10.1038/ncomms6755

Wang, W., Wang, L., and Shao, Z. (2018). Polycyclic aromatic hydrocarbon (PAH) degradation pathways of the obligate marine PAH degrader *Cycloclasticus* sp. strain P1. *Appl. Environ. Microbiol.* 84:e01261-18. doi: 10.1128/AEM.01261-18

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033

Wöhlbrand, L., Kallerhoff, B., Lange, D., Hufnagel, P., Thiermann, J., Reinhardt, R., et al. (2007). Functional proteomic view of metabolic regulation in "*Aromatoleum aromaticum*" strain EbN1. *Proteomics* 7, 2222–2239. doi: 10.1002/pmic.200600987

Xu, X., Liu, W., Tian, S., Wang, W., Qi, Q., Jiang, P., et al. (2018). Petroleum hydrocarbon-degrading bacteria for the remediation of oil pollution under aerobic conditions: a perspective analysis. *Front. Microbiol.* 9:2885. doi: 10.3389/fmicb.2018.02885

Yao, G., Yu, J., Hou, Q., Hui, W., Liu, W., Kwok, L. Y., et al. (2017). A perspective study of koumiss microbiome by metagenomics analysis based on single-cell amplification technique. *Front. Microbiol.* 8:165. doi: 10.3389/fmicb.2017.00165

Zhang, K., Hu, Z., Zeng, F., Yang, X., Wang, J., Jing, R., et al. (2019). Biodegradation of petroleum hydrocarbons and changes in microbial community structure in sediment under nitrate-, ferric-, sulfate-reducing and methanogenic conditions. *J. Environ. Manage.* 249:109425. doi: 10.1016/j.jenvman.2019.109425

Zorz, J. K., Sharp, C., Kleiner, M., Gordon, P. M. K., Pon, R. T., Dong, X., et al. (2019). A shared core microbiome in soda lakes separated by large distances. *Nat. Commun.* 10:4230. doi: 10.1038/s41467-019-12195-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.