Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Petroleum-Associated Genome Database - PaGeD: An Interactive Web Platform of genes and genomes related to the oil supply chain

Kelly Hidalgo-Martinez

kellyjhm@unicamp.br

State University of Campinas

Renato Augusto Corrêa dos Santos

São Paulo State University

João Vitor Leite Novoletti

São Paulo State University

Tulio de Lucca Capelini

State University of Campinas

Ana Letícia Malheiros Oliveira

State University of Campinas

Diego Mauricio Riaño-Pachón

State University of Campinas

German Zafra

Industrial University of Santander

Maira Paula de Sousa

Petrobras (Brazil)

Luiz Fernando Martins

Petrobras (Brazil)

Valéria Maia de Oliveira

State University of Campinas

Research Article

Keywords: microbiomes, genomics, petroleum, hydrocarbon, bioremediation

Posted Date: March 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9012371/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Microbial communities play key roles across the oil supply chain, mediating both detrimental processes, such as microbial influenced corrosion and souring, and beneficial applications like bioremediation. However, genomic and functional data of these microorganisms remain fragmented across repositories, hindering integrative analyses and biotechnological use.

Results

Here, we present the Petroleum-associated Genome Database (PaGeD), an open-source interactive platform designed to organize, explore, and analyze genomic and functional information from oil-associated microbiomes. PaGeD compiles 3,334 curated prokaryotic genomes retrieved from oil reservoirs, produced waters, hydrocarbon-polluted sites, and related environments worldwide. These genomes were clustered into 2,522 species-level units, more than half representing previously undescribed taxa, underscoring the hidden microbial diversity in petroleum systems. Functional profiling based on 144 KEGG ortholog groups related to hydrocarbon degradation revealed a modular organization in which central catabolic pathways are broadly conserved across habitats and phylogenetic groups, whereas peripheral, substrate-specific pathways are restricted to a limited number of taxa. A bipartite gene-genome network analysis highlighted this architecture, revealing a gradient of functional richness across genomes: taxa encoding broad repertoires of degradation genes occupy central positions, whereas genomes encoding few such functions are located at the periphery. Despite geographic and ecological heterogeneity, hydrocarbon degradation profiles were largely conserved, indicating convergent functional adaptation to petroleum-associated conditions.

Conclusions

PaGeD provides a scalable framework for genome-resolved exploration of oil-associated microbiomes, enabling the identification of candidate keystone taxa and functional markers and supporting future studies on the ecological and biotechnological potential of microorganisms in petroleum environments.

Introduction

The Oil & Gas industry is a crucial sector in global economy, mainly due to fossil fuels still serving as the primary energy source [1]. Microorganisms play critical roles across this supply chain, mediating processes that can be both detrimental, such as microbial influenced corrosion, souring, and biofouling, and beneficial, like bioremediation and wastewater treatment. For decades, extreme conditions such as high pressures, temperatures, salinity, and concentrations of heavy metals and organic solvents, led to the belief that microbial life could not persist in oil reservoirs, pipelines, and related environments [2, 3].

However, microorganisms have demonstrated remarkable adaptability [4, 5], thriving in these hostile settings due to the presence of water in many stages of production systems as well as in reservoirs [2].

In this context, several fundamental questions arise: i) Which microorganisms can grow under these extreme conditions? ii) What genomic traits and metabolic strategies enable their persistence and ecological success? iii) which microbial functions of industrial relevance can be inferred from their genomes, particularly those associated with hydrocarbon transformation? and iv) can these microorganisms be systematically identified and prioritized for future biotechnological applications within the oil supply chain?

Advances in large-scale DNA sequencing and bioinformatics have revolutionized our understanding of oil-associated microbiomes [2], overcoming many of the limitations of culture-based methods. Shotgun metagenomic now enables the reconstruction of near-complete genomes (Metagenome-Assembled Genomes - MAGs) from complex microbial communities inhabiting produced and injection waters, oil-polluted soils, storage tanks, pipelines, and oil reservoirs. These genomic resources make it possible to explore, at unprecedent resolution, the taxonomic diversity and functional potential underlying microbial processes in petroleum-related environments.

While pioneering initiatives, such as the Genome Repository of Oil Systems (GROS) [6], have enriched our knowledge, these efforts have primarily focused on specific events, like the Deepwater Horizon spill, leaving a gap in comprehensive genomic resources that cover the entire oil supply chain.

To expand our knowledge of the taxonomy and functional potential of oil-associated microbiomes, we developed the Petroleum-associated Genome Database (https://paged.cpqba.unicamp.br/). Here, we present an interactive web platform that organizes and provides access to curate genome-associated data, including taxonomic, ecological, and functional information from petroleum-related microorganisms. PaGeD offers an easy-to-navigate interface with search capabilities, filters, and robust ways to interpret the genomic content as a function of, e.g., isolation source and metabolic traits. In addition to providing deeper insights into microbial diversity in extreme oil-associated habitats, PaGeD also represents a resource that can be leveraged in future investigations to help cope with the negative roles of microorganisms and/or develop products and processes of strategic interest for the security and sustainability of the Oil & Gas industry.

Materials and methods

Genome collection and metadata retrieval

To explore the diversity of prokaryotic genomes in the Oil & Gas industry, we searched for publicly available genomes from environments such as oil reservoirs, produced or formation water, fracturing fluids, deep seeps, and contaminated soils or waters. We retrieved genome data from major databases, including Integrated Microbial Genomes & Microbiomes (IMG/M) [7], National Center for Biotechnology Information (NCBI) [8], and the Genome Repository of Oil Systems (GROS) via the Microbial Genome

Atlas [6, 9]. Additional genomes not deposited in these repositories were identified through targeted searches in scientific literature databases (e.g., Google Scholar, Scopus). For this, we used combinations of keywords such as “oil”, “oil field”, “oil reservoir”, “petroleum”, “crude oil”, “injection water”, “formation water”, “produced water”, “production water”, “oil refinery” and “microbially influenced corrosion” to retrieve relevant studies. All searches and data retrieval were completed by February 2024. All accession numbers and associated publications for the genomes included in PaGeD are provided in Table S1.

Most metagenome-assembled genomes (MAGs) used in this study were retrieved as preassembled genome bins from these public repositories and were originally reconstructed by the respective authors from their own metagenomic datasets and pipelines. In addition, a subset of MAGs originates from studies conducted by our group [10–12] and were generated using a previously published and validated reconstruction pipeline [13]. No metagenomic assembly or binning was performed specifically for the present study; our contribution consisted of curating, quality-filtering, dereplicating, and functionally annotating these genomes and integrating them into a standardized framework.

After manual curation to remove duplicate entries, 3,885 genomes were compiled (Fig. 1). Metadata for each genome, including geographic origin, geographic coordinates, and genome type (MAG or isolate), were extracted from the respective databases (Table S1). Habitat and isolation source categories were established based on metadata fields available in public repositories, particularly the “habitat” and isolation source” descriptors provided in the NCBI Biosample and JGI/IMG records. These terms manually curated and harmonizing into consistent categories which are fully documented in Table S2.

Quality assessment and taxonomic classification

Quality assessment was performed using CheckM v1.1.3 (Parks et al., 2015), retaining only genomes with ≥ 50% completeness and < 10% contamination (according to MiMAG standards; [14]. Furthermore, tRNAs were predicted with tRNAscan-SE v2.0.12 [15] and rRNAs identified using Barrnap (https://github.com/tseemann/barrnap). These steps yielded a final dataset of 3,334 medium- and high-quality genomes. GTDB-tk v2.3.2 (Release 214.1, April 2023; [16]) was used for taxonomic annotation, and genomes were clustered into 2,522 species-level groups (average nucleotide identity - ANI > 95% and > 30% alignment coverage [17]) with representative genomes selected using dRep v3.4.3 [18]. For phylogenetic reconstruction, genomes harboring at least 30% of the marker genes were filtered and then processed with GToTree v1.7 [19], and the resulting tree was visualized using iTOL v6 [20].

Functional annotation

Open reading frames (ORFs) were predicted with Prodigal v2.6.3. The ORFs were then annotated using the sequence aligner Diamond v.0.9.30 [21] based on the KEGG database (Kyoto Encyclopedia of Genes and Genomes) [22]. Hits were retained based on the stringent confidence thresholds (sequence identity ≥ 60% and alignment coverage ≥ 60% of query length).

To specifically evaluated hydrocarbon degradation potential, we curated a list of 144 KEGG Orthology identifiers my manual inspection of KEGG pathway maps. These included genes involved in the aerobic and anaerobic degradation of alkanes and aromatic hydrocarbons (mono-, di-, and polyaromatic compounds), key petroleum components. The curated set KO table (Table S3) was used to identify genomes encoding hydrocarbon degradation pathways and to quantify the completeness of metabolic modules associated with hydrocarbon transformation.

Bipartite co-occurrence network

To explore the functional organization of hydrocarbon degradation genes across genomes, we constructed a bipartite network linking KEGG Orthologs (KOs) to the genomes encoding them. Rather than inferring unknown gene-host associations, this network was used as an exploratory framework to visualize how hydrocarbon degradation functions are distributed across genomes, identify highly connected functional hubs, and examine gradients of functional specialization among degraders.

Only genomes encoding ≥15 distinct hydrocarbon-degradation KOs were retained for network construction. Bipartite edges were defined by the presence of a given KO in a genome, yielding a genome-gene network in which nodes represent either genomes or genes and edges represent gene occurrence. In parallel, a KO–KO co-occurrence matrix was computed from a binary KO-by-genome matrix and retained for downstream analyses, but visualization focused on the bipartite structure to preserve the direct mapping between functions and hosts.

Gene nodes were classified by their frequency across genomes: rare (< 5%), intermediate (5–15%) or core (≥15%). Functional richness was defined as the number of distinct hydrocarbon degradation KOs encoded per genome and treated as as a continuous measure positioning genomes along a gradient from specialist-like to generalist-like. For visualization and interpretation, genomes in the upper quartile of functional richness were referred to as generalist-like, whereas genomes in the lower quartile were referred to specialist-like; genomes with intermediate values occupied transitional positions along this gradient.

The network was built using the igraph v1.5.1 R package and visualized with ggraph v2.1.0. Gene nodes were displayed as hollow symbols, colored by pathway class and shaped by frequency category, whereas genome nodes were displayed as filled circles, colored by phylum and scaled according to functional richness. This representation highlights conserved catabolic cores, rare and peripheral pathways, and the distribution of functional potential across genomes.

Basic topological properties of the resulting network, including the number of nodes and edges, degree distribution, and modularity, were calculated using igraph v1.5.1 and used to support the interpretation organization and hub structure.

Platform implementation and interface

The Petroleum-associated Genome Database (PaGeD) was developed as an open-source web platform for interactive exploration of taxonomic, functional, and ecological data associated with oil microbiomes.

The platform comprises a Python Flask v3.0.0. and MariaDB v14.14 as the database engine, providing a lightweight and efficient means of storing and retrieving genomic data. To support fast and flexible search functionality, the platform integrates the Whooshee extension v2.7.4, enabling full-text search across genome attributes, taxonomic classification, and functional annotations.

The front end was built using HTML, CSS, and JavaScript, using Bootstrap for responsive design. Interactive visualizations were implemented with Chart.js v4.0 and Leaflet.js v2.0.0, complemented by jQuery v3.7.1, and qTip2.js v3.0.3 for enhance interactivity. The platform includes pages for database overview, taxonomy explorer, functional explorer, and species-specific pages. Species pages integrate taxonomic classification, geographic mapping, habitat information, and genome metadata, including accession numbers from public repositories (e.g., NCBI, JGI).

Results and discussion

Global environmental distribution of the PaGeD genomes

Herein, the “Petroleum-associated Genomes Database” (PaGeD), a vast collection of genomes obtained from culturable microorganisms and from metagenomes (MAGs), was created and curated. The genomes of PaGeD were originally reported as related to the Oil & Gas industry, spanning environments such as offshore and onshore oil reservoirs, natural oil seeps, processing fluids (e.g. produced, formation and/or injection water, fracking fluids), microbiologically influenced corrosion (MIC) samples, oil spills, hydrocarbon-impacted sites or even laboratory microcosms or enrichment cultures simulating hydrocarbon contamination. These isolation sources were categorized into habitats (Table S2). Figure 2 summarizes the distribution across these habitats. Engineered systems (e.g., fracturing fluids, MIC samples, machinery facilities 34.22%, followed by marine-associated habitats (28.37%).

Genomes are represented from 74 countries across six continents and major ocean basins (Fig. 3). Although coverage is global, geographic representation remains uneven, with predominance of genomes from North America, particularly the Gulf of Mexico, reflecting the sequencing efforts following the Deepwater Horizon (DWH) spill [6, 23–26]. As in other large-scale microbial datasets, such as Earth Microbiome Project [27], underrepresentation of tropical regions from the Global South remains evident, highlighting opportunities for future sampling to fill geographic and ecological gaps.

Characterization of the PaGeD Genome Collection

The final curated dataset comprises 3,334 genomes, of which 3,014 correspond to MAGs and 319 are genomes derived from isolates. All genomes meet or exceed medium quality thresholds (≥50%

completeness, < 10% contamination). Completeness average 89.59 (± 9.80) and 3.35% (± 2.43). Approximately ~ 36% of the genomes contain at least partial sequences of the 16S rRNA gene.

Assembly metrics show large variation: genome sizes ranged from 0.27 Mb to 12.25 Mb, while contig counts varied from single-contig assemblies to 2,899 contigs (mean = 238) (Fig. 4). The smallest genomes belonging to members of Nanoarchaeota and Patescibacteria (phyla known for their reduced genomes), while larger genomes were mainly affiliated with Myxococcota and Planctomycetota, phyla characterized by complex lifestyles and metabolic versatility [28–30].

Species-level clustering at 95% ANI grouped into 2,522 clusters. Among these, 2,089 were singletons (i.e., represented by a single genome), underscoring the high microbial diversity in petroleum-associated environments. Only 47 clusters contained five or more genomes. When compared to previous initiatives such as GROSS [6], PaGeD substantially expands the known microbial diversity of oil environments. Culture-derived genomes represent only ~ 10% of all species-level clusters (Fig. 4 and S1), revealing a significant gap between cultivated and uncultivated diversity. Moreover, nearly half of the representative (centroid) genomes could not be taxonomically assigned at the genus level, highlighting the potential of PaGeD for uncovering novel microbial lineages in oil-impacted environments.

From an applied perspective, this extensive and largely unexplored diversity constitutes a valuable reservoir of metabolic potential. The species-resolved organization of PaGeD enables the systematic identification of microorganisms with relevance for oil-associated systems, supporting the screening of taxa for biotechnological applications such as bioremediation and microbial-enhanced oil recovery, as well as for microbial risk assessment. By linking genomes to habitat, geographic origin, and functional potential, PaGeD provides a practical framework that bridged large-scale genome mining with applied decision-making in environmental and industrial microbiology.

Phylogenetic Diversity and Potentially Novel Taxa in Oil-Associated Microbiomes

Taxonomic classification revealed the presence of 109 different phyla (94 bacterial and 15 archaeal phyla), 248 classes, 553 orders, 999 families, and 1,799 genera. Pseudomonadota dominates PaGeD, with 26.7% of all genomes, followed by Bacteroidota 6.3%, Thermoproteota 5.6%, Desulfobacterota 4.8% and Chloroflexota 4.6% (Figure S1). Remarkably, PaGeD includes one class, eight orders, 36 families, 197 genera, and 437 species without any cultivated representative in the GTDB, further illustrating the contribution of PaGeD to uncovering novel lineages.

The most widely represented phylum, Pseudomonadota, was detected across all continents, diverse habitat types (e.g., marine, aquatic, terrestrial, microcosms, engineered and oil fields), and isolation sources such as hydrothermal vent sediments, microcosms, oil-polluted soils, and produced waters (Figure S2).

Notable taxa include *Acinetobacter radioresistens*, abundant in MIC-related samples, consistent with its biofilm-forming abilities in a corrosion context [31, 32], and *Pseudomonas aeruginosa*, detected across a wide array of environments, reflecting its ecological versatility [33] and reinforcing its role in biosurfactant production, bioremediation and Microbial-Enhanced Oil Recovery (MEOR) [34, 35]. *P. aeruginosa* exemplifies the applied relevance of specific petroleum-associated taxa.

Interestingly, nearly 3% of the genomes were affiliated with the recently proposed Patescibacteria phylum [36], with greater prevalence in activated sludge, wastewater, and tailing ponds (Figure S3). Members of this phylum are known for their ultrasmall cell size and reduced genomes, often lacking essential biosynthetic pathways, which support the hypothesis of a symbiotic lifestyle [37]. The prevalence of such potential symbionts in oil-associated environments is consistent with the recognized importance of syntrophy for the biodegradation of aliphatic and aromatic hydrocarbons [38]

Habitat specificity of microbial genera in PaGeD

PaGeD also provides a valuable framework for exploring the ecological breadth and biogeographic distribution of oil-associated microbial genera. The distribution of genera revealed a high degree of habitat specificity: approximately 84% of the genera were found restricted to a single isolation source, while 10% occurred in just two sources (Fig. 5A). The sharp decline in genera prevalence with increasing habitat diversity indicates that petroleum-associated microbiomes are dominated by endemic rather than cosmopolitan taxa, corroborating patterns described in previous studies [39, 40]. Among genera restricted to a single isolation source, hydrothermal vent sediments from the Guaymas Basin emerged as the most compositionally unique (Fig. 5B). This finding aligns with previous reports highlighting the distinctiveness of microbiomes in hydrothermal settings [41, 42], likely driven by extreme physicochemical conditions and high levels of hydrocarbon exposure.

Functional insights into hydrocarbon-degradation profiles of oil-associated microbes

In addition to the taxonomic and ecological diversity contained in PaGeD genomes, we further investigated their functional potential with emphasis on hydrocarbon degradation. From genomic perspective, hydrocarbon metabolism is structured as a modular process, in which chemically diverse substrates are initially activated by specialized enzymes and subsequently funneled into a limited set of central intermediates such as catechol and benzoyl-CoA, that feed into core catabolic pathways [43–45]. This architecture underpins both metabolic specialization and ecological cooperation, as different taxa may contribute distinct steps of the overall degradation process.

Genome-resolved approaches have revealed that enzymes responsible for the initial activation of specific hydrocarbons (e.g., alkanes, BTEX compounds, or polycyclic aromatic hydrocarbons) are often phylogenetically restricted, whereas downstream pathways involved in the processing of central intermediates are broadly distributed across bacterial and archaeal lineages [46, 47]. Such functional stratification reflects evolutionary convergence toward shared metabolic solutions in hydrocarbon-

impacted environments and supports the view that petroleum microbiomes are organized around a combination of specialist degraders and generalist metabolizers.

Within this framework, PaGeD provides a unique opportunity to examine how these functional modules are distributed across thousands of genomes originating from diverse petroleum-related habitats. By integrating taxonomic resolution with pathway-level annotation, the platform enables the identification of conserved catabolic cores, rare and substrate-specific pathways, and lineage-specific patterns of metabolic specialization. The analyses presented in the following sections explore this functional landscape at multiple levels, from global genomic capacities to species- and lineage-resolved hydrocarbon degradation profiles.

Overview of Genomic Functional Capacities

Functional annotation identified 8,485 KOs across the dataset, with individual genomes containing an average of 664 ± 507 KOs. The most abundant categories were amino acid metabolism, carbohydrate metabolism, and energy metabolism (Figure S4). Stress response and quorum sensing genes were also common. Transporter families such as ABC (ATP-binding cassette) and MFS (major facilitator superfamily) were widely represented. These transport systems likely play a key role in substrate uptake and efflux mechanisms, which are crucial for microbial adaptation to hydrocarbon-rich environments [48].

In the PaGeD platform, users can navigate to species-specific pages that consolidate taxonomic, ecological, and functional information for each genome cluster (Fig. 6). The pages integrate: i) taxonomic classification, ii) geographic distribution of genomes via interactive maps, and iii) general metadata. Each genome entry also links directly to public accession numbers (e.g., NCBI, JGI), ensuring traceability and reuse. Importantly, from this page users can access additional tabs to explore the general and hydrocarbon-specific functional potential of each species.

While the global functional overview presented above was derived from the analyses performed for this manuscript, the PaGeD platform currently provides species-level summaries of the functional potential of the deposited genomes. For each species, the average proportion of genes assigned to KEGG Level 2 categories was calculated and displayed as bar plots, based on all genomes available in the database. These plots offer a descriptive view of the metabolic traits of individual taxa, serving as an exploratory tool for users interested in the functional repertoire of particular species (Fig. 7).

Functional landscape of hydrocarbon degradation in PaGeD genomes

To assess the hydrocarbon degradation potential encoded within PaGeD genomes, we screened all annotated protein sequences against a manually curated list of 144 KEGG Orthology (KO) identifiers (Table S3). A total of 1,505 genomes (45.1%) harbored at least one hydrocarbon degradation gene. To further characterize the hydrocarbon degrading potential, we examined the representativeness in the

dataset of 22 curated metabolic modules involved in the aerobic and anaerobic breakdown of hydrocarbons. Central pathways, such as catechol, benzoyl-CoA and benzoate degradation, were widespread and represented with high completeness (Fig. 8). These highly prevalent central pathways are responsible for degrading the core intermediates that result from the initial activation of structurally diverse hydrocarbons. The wide distribution of these modules in PaGeD's genomes further reveals that assemblages of microorganisms in oil-polluted environments are functionally enriched and capable of degrading a wide range of hydrocarbon-acids products via a convergent common central catabolic pathway, conferring them adaptive potentialities and metabolic efficiency.

Conversely, peripheral pathways responsible for the activation of specific hydrocarbon classes, such as ethylbenzene, naphthalene, or phenanthrene, were identified in fewer than 5% of genomes. However, when present, these pathways often exhibit high completeness. This pattern is consistent with a modular architecture in hydrocarbon degradation, where broadly conserved central catabolic modules are widely distributed. At the same time peripheral and substrate-specific pathways are restricted to specialized ecological niches of phylogenetic groups. Somee, M.R. et al. (2022) reported similar results; they observed a broad phylogenetic distribution of key enzymes involved in the transformation of hydrocarbon aerobic degradation intermediates, in contrast to a more limited distribution of key enzymes responsible for the initial attack of specific hydrocarbons [46].

In addition to the general functional profiles, PaGeD also offers species-level summaries of hydrocarbon degradation potential. For each curated metabolic module, pathway completeness was calculated as the proportion of expected genes detected, averaged across genomes belonging to the same species. These bar plots highlight which hydrocarbon catabolic pathways are most prevalent within in each taxon, providing users with a practical tool to investigate both the general and hydrocarbon-specific functional potential of oil-associated microbes (Fig. 9).

From Phyla to Genera: Taxonomic Layers of Hydrocarbon-Degrading Capacity

To explore how hydrocarbon degradation capacity is distributed across microbial diversity represented in PaGeD, we calculated the average completeness of catabolic modules/pathways at both phylum (Figure S5a) and genus levels (Figure S5b). This lineage-resolved approach revealed a pronounced functional stratification, where broad central degradation pathways are widely conserved across multiple phyla, while peripheral and substrate-specific routes are more sporadically distributed and often restricted to specialized genera. By integrating information on pathway completeness, we uncovered clear patterns of metabolic versatility and specialization that shape the ecological roles of petroleum-associated microbes.

At the phylum level, Pseudomonadota stood out with the broadest and most complete repertoire of pathways (Figure S5a). Genomes from this phylum exhibited high average completeness for both peripheral pathways (e.g., ethylbenzene, alkane, biphenyl) and central modules (e.g., catechol and benzoyl-CoA degradation), highlighting their metabolic versatility and ecological dominance in oil-

associated environments [43, 46, 47, 49–51]. Zooming into the genus level, we observed striking heterogeneity in catabolic specialization (Figure S5b). *Pseudomonas* emerged as a central player, showing broad representativeness across nearly all modules and particularly high average completeness in catechol ortho-cleavage (36%) and benzoate pathways. These findings are in accordance with several previous reports showing the ability of this genus to degrade different kinds of hydrocarbons [46, 52–54]. Several genera showed functional specialization, with high completeness restricted to specific modules. For instance, *Cycloclasticus* showed > 35% completeness in toluene and biphenyl degradation pathways, consistent with its known adaptation to PAH degradation in marine ecosystems [55, 56].

The catechol meta-cleavage module, previously highlighted as the most prevalent across PaGeD, was confirmed to be broadly distributed across phyla and among a wide range of genera, including metabolically streamlined or niche-adapted groups (e.g., *Sulfitobacter, Enterococcus, Oleispira, Smithella*), although often with low completeness at the individual genome level. This finding suggests a strong ecological convergence towards the ability to process aromatic intermediates. Conversely, peripheral pathways such as naphthalene, styrene, and phthalate degradation were typically restricted to a few genera, often with high pathway completeness, supporting the view that specialized functions are phylogenetically restricted [46, 57].

Altogether, these findings reveal a functional landscape shaped by both *substrate-specific degraders* and *intermediate metabolizers*: while particular genera possess peripheral catabolic pathways that enable them to initiate the degradation of specific hydrocarbon compounds (e.g., alkanes, PAHs, BTEX), others lack these activation steps but play a crucial role in the downstream processing of central intermediates like catechol and benzoyl-CoA. The widespread distribution of these central pathways across diverse phyla and genera suggests evolutionary convergence and functional redundancy, which likely enhance ecosystem resilience and enable metabolic cooperation in petroleum-contaminated environments [46].

Keystone Genomes in Hydrocarbon Degradation

We identify individual genomes encoding particularly rich repertoires of degradation genes, such as *Paraburkholderia aromaticivorans, Cycloclasticus* spp., and *Pseudomonas veronii.* These taxa encode *to* complete or near-complete modules for multiple aromatic pathways, highlighting their potential ecological importance (Figure S6).

Rather than inferring unknown gene-host associations, the bipartite network was constructed as an exploratory framework to visualize the organization of hydrocarbon functions across genomes, identify highly connected functional hubs, and examine gradients of functional specialization. The network links hydrocarbon degradation genes to the genomes encoding them (Fig. 10A,B).

The resulting bipartite network comprised 143 genome nodes and 115 gene nodes connected by 2,940 edges, and exhibited a low overall density (0.089), indicating selective and non-random associations between genes and genomes. The degree distribution was strongly rich-skewed (mean degree = 22.8; maximum degree = 108), revealing a small number of highly connected hubs contrasted with a large

fraction of low-degree nodes. Community detection yielded a modularity of Q = 0.31, consistent with a moderately modular architecture in which specialized substructures are embedded within a shared functional backbone. Together, these properties support a network organization dominated by central catabolic genes (e.g., catechol cleavage, benzoate, and benzoyl-CoA) and generalist-like genomes, surrounded by peripheral, specialist functions. These core genes were recurrently encoded by the same genomes, forming a highly connected central backbone of the bipartite network and suggesting conserved metabolic blocks and ecological selection for shared intermediate processing [44, 58].

Genomes with higher functional richness occupied more central positions in the network, whereas genomes encoding fewer degradation genes were preferentially located at the periphery. In line with the operational definitions adopted here, genomes in upper quartile of functional richness exhibit a generalist-like profile, encoding broad repertoires of hydrocarbon degradation genes, while genomes in the lower quartile display specialist-like profiles, typically encoding only one or a few such functions.

Altogether, this topology illustrates redundancy in central modules and specialization in peripheral ones, revealing an ecological and functional complementarity that may contribute to the robustness and versatility of microbial communities in petroleum-associated environments. From an applied perspective, this structure provides a conceptual basis for identifying candidate keystone taxa and functional markers that may serve as targets for monitoring and biotechnological exploration.

Hydrocarbon Degradation Profiles Across Habitats

The completeness of hydrocarbon degradation modules was further examined across habitats (Figure S7). Central modules showed consistent representation across environments, while peripheral modules displayed habitat-specific enrichment. For example, phenanthrene degradation was more prevalent in genomes form oil sand tailing ponds, consistent with selective pressures in highly contaminated engineered environments. The high prevalence of phenanthrene module observed in PaGeD genomes from these systems may reflect the presence of such specialized degraders enriched by long-term adaptation to the recalcitrant contaminants.

A principal coordinate analysis (PCoA) based on the Jaccard distance between genomes (Figure S8) revealed limited separation among habitats, indicating a high degree of functional overlap in hydrocarbon degradation profiles across environments. Rather than reflecting the absence of ecological structure, this pattern suggests that genomes from distinct petroleum-related habitats share substantial fraction of catabolic traits. Ordination methods based on presence-absence metrics, such as Jaccard distance, are widely used to compare functional repertoires and emphasize compositional similarity over abundance [59, 60]. In this context, the lack of strong clustering implies that core hydrocarbon degradation modules are broadly distributed among genomes, regardless of their environmental origin.

This functional convergence is consistent with the idea that exposure to hydrocarbons imposes similar selective pressures across aquatic, terrestrial, engineered, and subsurface systems, favoring the retention of shared catabolic functions. At the same time, the partial dispersion observed in the

ordination space reflects the presence of habitat-specific and substrate-specialized pathways, as highlighted by the differential distribution of peripheral modules (Figure S7). Together, these patterns support a model in which petroleum-associated microbiomes are structured around a common catabolic core, complemented by niche-dependent specialization, rather than forming strictly habitat-segregated functional assemblages.

Biogeographic Patterns of Hydrocarbon-Degrading Genomes

Mapping of hydrocarbon-degrading genomes (Figure S9) revealed hotspots in the Gulf of Mexico, due to heavily studied in the aftermath of the Deepwater Horizon spill, showing that is still a legacy of intense genomic exploration in petroleum-impacted environments [6, 23, 24, 61–63]. Another noteworthy cluster was observed in Eastern China, a region with known petroleum basins and extensive industrial activity. The presence of functionally rich genomes in this geographic region likely reflects both environmental contamination and intense research activity [64–70]. Despite the broad geographic coverage, underrepresentation of regions in Africa, South America and Southeast Asia persists, limiting current knowledge of petroleum-associated microbial diversity.

Conclusions

This work presents the first version of the PaGeD platform and database, offering an interactive interface for further exploring genomic and associated metadata from oil-associated microbiomes. By integrating high-quality genomes with curated functional annotations, PaGeD provides an unprecedented resource to investigate the diversity, ecology, and hydrocarbon-degradation potential of oil-associated microbial communities. This framework not only supports fundamental research but also has the potential to inform applied strategies in bioremediation and microbial risk management.

The database enables users to navigate species-level clusters, explore habitat and geographic distributions, and examine both general and hydrocarbon-specific functional profiles through user-friendly visualizations. These features facilitate comparative and integrative studies, offering a scalable framework that can be applied to diverse research question in microbiology, biotechnology, and microbial risk management.

Future developments of PaGeD will include the integration of additional functional annotations of industrial and environmental relevance, such as antimicrobial and heavy metal resistance genes (resistome), biosurfactant production, mobile genetic elements, and biosynthetic gene clusters. Regular updates with newly available genomes will further expand taxonomic and functional coverage, strengthening PaGeD as an open and evolving resource for the scientific community.

Abbreviations

PaGeD	Petroleum-Associated Genome Database

DNA	Deoxyribunocleic acid
MAG	Metagenome-Assembled Genomes
GROS	Genome Repository of Oil Systems
IMG/M	Integrated Microbial Genomes & Microbiomes
NCBI	National Center for Biotechnology Information
JGI/IMG	Joint Genome Institute/ Integrated Microbial Genomes & Microbiomes
tRNA	Transfer Ribonucleic Acid
GTDB-tk	Genome Taxonomy Data Base Tool kit
ANI	Average Nucleotide Identity
ORF	Open Reading Frame
KO	KEGG Orthology
MIC	Microbiology influenced corrosion
DWH	Deepwater Horizon
rRNA	Ribossomal Ribonucleic Acid
MEOR	Microbial-Enhanced Oil Recovery
BTEX	Benzene, Toluene, Ethylbenzene and Xylenes
ABC	ATP-binding cassette
MFS	Major facilitator superfamily
PCoA	Principal Coordinate Analysis

Declarations

Author contributions

KH-M: Conceptualization, Methodology, Data Collection, Data curation, Software, Formal analysis, Visualization, Software development (back-end & front-end) Writing (original draft, review & editing). **RACS**: Conceptualization, Software development (back-end & front-end), Writing (review & editing).. **JVLN**: Software development (back-end & front-end). **TLC**: Data contribution. **ALMO**: Data contribution. **DIMR-P**: Supervision, Writing (review & editing). **GZ**: Data contribution, Writing (review & editing).. **MPS**:

Writing (review & editing).. **LFM**: Funding acquisition, Project administration, Writing (review & editing). **VMO**: Conceptualization, Supervision, Funding acquisition, Project administration, Writing (review & editing).

Funding

This research was partially funded by Petróleo Brasileiro S.A. (PETROBRAS), regulated by the Brazilian National Agency for Petroleum, Natural Gas, and Biofuels (ANP, Resolution 50/2015). K.H-M was supported by the São Paulo Research Foundation (FAPESP, Postdoctoral Fellowship, process number 2024/15102-8). RACS was supported by the Brazilian São Paulo Research Foundation (FAPESP, Postdoctoral Fellowship, process number 2023/11133-3). DMRP acknowledges a fellowship from National Council for Scientific and Technological Development – CNPq (311558/2021-6).

Acknowledgments

We want to acknowledge the Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD-SP, Brazil) for providing the high-performance computing resources essential to this work. Special thanks go to José Luis Bezerra da Silva Filho and Daniel Franco for their valuable contributions to the development of the PaGeD web platform, and to Rodrigo Alexander de Andrade Pierini for his role as Network Analyst.

Data availability

The curated dataset underlying this study is publicly available at the Research Data Repository (REDU) under the accession DOI: https://doi.org/10.25824/redu/VKTSIB. This repository includes metadata associated with 3,334 genomes compiled for the PaGeD platform, covering taxonomic classification, genome quality, assembly statistics, geographic origin, and ecological context. Table S1 provide detailed accession numbers (NCBI, JGI/IMG) and literature DOIs for genomes with publicly available references. All additional supporting information is included as Supplementary Information files accompanying this article.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interest

The authors declare that they have no competing interests.

References

1. Lu H, Guo L, Azimi M, Huang K, Oil. Gas 4.0 era: A systematic review and outlook. Comput Ind. 2019;11168–90. https://doi.org/10.1016/j.compind.2019.06.007.
2. Wunch K, Stipaničev M, Frenzel M. Microbial Bioinformatics in the Oil and Gas Industry: Applications to Reservoirs and Processes. CRC; 2021.
3. Augustinovic Z, Birketveit O, Clements K, Freeman M, Gopi S, Ishoey T, Jackson G, Kubala G, Larsen J, Marcotte B. Microbes–oilfield enemies or allies. Oilfield Rev. 2012;24(2):1–14.
4. Foght J. Microbial communities in oil shales, biodegraded and heavy oil reservoirs, and bitumen deposits. Handbook of hydrocarbon and lipid microbiology. 2010.
5. Sierra-Garcia IN. Investigation of the microbiota and processes involded in petroleum degradation in oil reservoirs. Doctoral 2016.
6. Karthikeyan S, Rodriguez -RLM, Heritier-Robbins P, Hatt JK, Huettel M, Kostka JE, Konstantinidis KT. Genome repository of oil systems: An interactive and searchable database that expands the catalogued diversity of crude oil-associated microbes. Environ Microbiol. 2020;22(6):2094–106. https://doi.org/10.1111/1462-2920.14966.
7. Chen I-MA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Huntemann M, Varghese N, White JR, Seshadri R. IMG/M v. 5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res. 2019;47(D1):D666–77. https://doi.org/10.1093/nar/gky901.
8. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45. https://doi.org/10.1093/nar/gkv1189.
9. Rodriguez -RLM, Gunturu S, Harvey WT, Rosselló-Mora R, Tiedje JM, Cole JR, Konstantinidis KT. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. Nucleic Acids Res. 2018;46(W1):W282–8. https://doi.org/10.1093/nar/gky467.
10. Hidalgo KJT, Soriano EH, Valoni AU, Baessa E, Richnow MP, Vogt HH, Chang C, Oliveira HK. V.M. Taxonomic and functional diversity of the microbiome in a jet fuel contaminated site as revealed by combined application of in situ microcosms with metagenomic analysis. Sci Total Environ. 2020:708135152.
11. Hidalgo K, Centurion V, Lemos L, Soriano A, Valoni E, Baessa M, Richnow H, Vogt C, Oliveira V. Disentangling the microbial genomic traits associated with aromatic hydrocarbon degradation in a jet fuel-contaminated aquifer. Biodegradation. 2025;36(1):7.
12. Hidalgo K, Cueva L, Giachini A, Schneider M, Soriano A, Baessa M, Martins L, Oliveira V. Long-term microbial functional responses in soil contaminated with biofuel/fossil fuel blends triggered by different bioremediation treatments. Environmental Pollution. 2025125685.
13. Hidalgo K. Binning. In. GitHub; 2024.

14. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35(8):725–31. https://doi.org/10.1038/nbt.3893.
15. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021;49(16):9077–96. https://doi.org/10.1093/nar/gkab688.
16. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019;36(6):1925–7. https://doi.org/10.1093/bioinformatics/btz848.
17. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):5114. https://doi.org/10.1038/s41467-018-07641-9.
18. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017:11(12):2864–8. https://doi.org/10.1038/ismej.2017.126
19. Lee MD. GToTree: a user-friendly workflow for phylogenomics. Bioinformatics. 2019;35(20):4162–4. https://doi.org/10.1093/bioinformatics/btz188.
20. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research. 2019:47(W1):W256-W259. https://doi.org/10.1093/nar/gkz239
21. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2014;12(1):59. https://doi.org/10.1038/nmeth.3176.
22. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. https://doi.org/10.1093/nar/28.1.27.
23. Mason OU, Hazen TC, Borglin S, Chain PS, Dubinsky EA, Fortney JL, Han J, Holman H-YN, Hultman J, Lamendella R. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. The ISME journal. 2012:6(9):1715–27. https://doi.org/10.1038/ismej.2012.59
24. Mason OU, Scott NM, Gonzalez A, Robbins-Pianka A, Bælum J, Kimbrel J, Bouskill NJ, Prestat E, Borglin S, Joyner DC, et al. Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. ISME J. 2014;8(7):1464–75. https://doi.org/10.1038/ismej.2013.254.
25. Kimes NE, Callaghan AV, Suflita JM, Morris PJ. Microbial transformation of the Deepwater Horizon oil spill—past, present, and future perspectives. Front Microbiol. 2014:5603. https://doi.org/10.3389/fmicb.2014.00603
26. Rodriguez-r LM, Overholt WA, Hagan C, Huettel M, Kostka JE, Konstantinidis KT. Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. ISME J. 2015;9(9):1928–40. https://doi.org/10.1038/ismej.2015.5.

27. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nat 2017:551(7681):457–63. https://doi.org/10.1038/nature24621
28. Zou D, Zhang C, Liu Y, Li M. Biogeographical distribution and community assembly of Myxococcota in mangrove sediments. Environ Microbiome. 2024;19(1):47. https://doi.org/10.1186/s40793-024-00593-2.
29. Saggu SK, Nath A, Kumar S. Myxobacteria: biology and bioactive secondary metabolites. Res Microbiol 2023:174(7):104079. https://doi.org/10.1016/j.resmic.2023.104079
30. Klimek D, Herold M, Calusinska M. Comparative genomic analysis of Planctomycetota potential for polysaccharide degradation identifies biotechnologically relevant microbes. BMC Genomics. 2024;25(1):523. https://doi.org/10.1186/s12864-024-10413-z.
31. Zhu XY, Lubeck J, Kilbane JJ. Characterization of Microbial Communities in Gas Industry Pipelines. Appl Environ Microbiol. 2003;69(9):5354–63. https://doi.org/10.1128/AEM.69.9.5354-5363.2003.
32. Lerm S, Westphal A, Miethling-Graff R, Alawi M, Seibt A, Wolfgramm M, Würdemann H. Thermal effects on microbial composition and microbiologically induced corrosion and mineral precipitation affecting operation of a geothermal plant in a deep saline aquifer. Extremophiles 2013:17(2):311–27. https://doi.org/10.1007/s00792-013-0518-8
33. Crone S, Vives-Flórez M, Kvich L, Saunders AM, Malone M, Nicolaisen MH, Martínez-García E, Rojas-Acosta C, Catalina Gomez-Puerto M, Calum H, et al. The environmental occurrence of Pseudomonas aeruginosa. APMIS. 2020;128(3):220–31. https://doi.org/10.1111/apm.13010.
34. Liu WJ, Duan XD, Wu LP, Masakorala K. Biosurfactant Production by Pseudomonas aeruginosa SNP0614 and its Effect on Biodegradation of Petroleum. Appl Biochem Microbiol. 2018:54(2):155–62. https://doi.org/10.1134/S0003683818020060
35. Câmara JMDA, Sousa MASB, Barros Neto EL, Oliveira MCA. Application of rhamnolipid biosurfactant produced by Pseudomonas aeruginosa in microbial-enhanced oil recovery(MEOR). Journal of Petroleum Exploration and Production Technology. 2019:9(3):2333–2341. https://doi.org/10.1007/s13202-019-0633-x.
36. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015;523(7559):208–11.
37. Nascimento Lemos L, Manoharan L, William Mendes L, Monteiro Venturini A, Satler Pylro V, Tsai SM. Metagenome assembled-genomes reveal similar functional profiles of CPR/Patescibacteria phyla in soils. Environ Microbiol Rep. 2020:12(6):651–5. https://doi.org/10.1111/1758-2229.12880
38. Morris BEL, Henneberger R, Huber H, Moissl-Eichinger C. Microbial syntrophy: interaction for the common good. FEMS Microbiol Rev. 2013;37(3):384–406. https://doi.org/10.1111/1574-6976.12019.
39. Yang S, Wen X, Shi Y, Liebner S, Jin H, Perfumo A. Hydrocarbon degraders establish at the costs of microbial richness, abundance and keystone taxa after crude oil contamination in permafrost

environments. Sci Rep. 2016;6(1):37473. https://doi.org/10.1038/srep37473.

40. Wang J, Zhang Y, Ding Y, Song H, Liu T, Zhang Y, Xu W, Shi Y. Comparing the indigenous microorganism system in typical petroleum-contaminated groundwater. Chemosphere. 2023:311137173. https://doi.org/10.1016/j.chemosphere.2022.137173
41. Cruaud P, Vigneron A, Pignet P, Caprais J-C, Lesongeur F, Toffin L, Godfroy A, Cambon-Bonavita M-A. Comparative Study of Guaymas Basin Microbiomes: Cold Seeps vs. Hydrothermal Vents Sediments. Front Mar Sci. 2017:4. https://doi.org/10.3389/fmars.2017.00417
42. Zhou Z, St. John E, Anantharaman K, Reysenbach A-L. Global patterns of diversity and metabolism of microbial communities in deep-sea hydrothermal vent deposits. Microbiome. 2022;10(1):241. https://doi.org/10.1186/s40168-022-01424-7.
43. Fuchs G, Boll M, Heider J. Microbial degradation of aromatic compounds–from one strategy to four. Nat Rev Microbiol. 2011;9(11):803. https://doi.org/10.1038/nrmicro2652.
44. Carmona M, Zamarro MT, Blázquez B, Durante-Rodríguez G, Juárez JF, Valderrama JA, Barragán MJL, García JL, Díaz E. Anaerobic Catabolism of Aromatic Compounds: a Genetic and Genomic View. Microbiol Mol Biol Rev. 2009;73(1):71–133. https://doi.org/10.1128/mmbr.00021-08.
45. von Netzer F, Kuntze K, Vogt C, Richnow HH, Boll M, Lueders T. Functional Gene Markers for Fumarate-Adding and Dearomatizing Key Enzymes in Anaerobic Aromatic Hydrocarbon Degradation in Terrestrial Environments. J Mol Microbiol Biotechnol. 2016;26(1–3):180–94. 10.1159/000441946. http://www.karger.com/DOI/.
46. Somee MR, Amoozegar MA, Dastgheib SMM, Shavandi M, Maman LG, Bertilsson S, Mehrshad M. Genome-resolved analyses show an extensive diversification in key aerobic hydrocarbon-degrading enzymes across bacteria and archaea. BMC Genomics. 2022;23(1):690. https://doi.org/10.1186/s12864-022-08906-w.
47. Hidalgo KJ, Sierra-Garcia IN, Dellagnezze BM, de Oliveira VM. Metagenomic Insights Into the Mechanisms for Biodegradation of Polycyclic Aromatic Hydrocarbons in the Oil Supply Chain. Front Microbiol. 2020:11(2250). https://doi.org/10.3389/fmicb.2020.561506
48. Wu J, Zhao R, Zhao L, Xu Q, Lv J, Ma F. Sorption of petroleum hydrocarbons before transmembrane transport and the structure, mechanisms and functional regulation of microbial membrane transport systems. J Hazard Mater. 2023;441129963. https://doi.org/10.1016/j.jhazmat.2022.129963.
49. Ramos-Mendoza IS, Embarcadero-Jiménez S, Barrios-Navarro AF, Ledezma-González DC, Lago-Lestón A, Valenzuela-Suárez BJ, Ramírez-Álvarez N, Silva-Jiménez H. Prokaryotic community structure and predicted metabolism associated with hydrocarbon degradation in marine sediments from the Northwest Coast of Baja California, Mexico. J Soils Sediments. 2024;24(8):3148–66. https://doi.org/10.1007/s11368-024-03822-w.
50. Mariano DCO, Dias GM, Castro MR, Tschoeke DA, de Oliveira FJS, Sérvulo EFC, Neves BC. Exploring the diversity and functional profile of microbial communities of Brazilian soils with high salinity and oil contamination. Heliyon. 2024:10(14). https://doi.org/10.1016/j.heliyon.2024.e34336

51. Camacho J, Mesén-Porras E, Rojas-Gätjens D, Pérez-Pantoja D, Puente-Sánchez F, Chavarría M. Draft genome sequence of three hydrocarbon-degrading < i>Pseudomonadota strains isolated from an abandoned century-old oil exploration well. Microbiol Resource Announcements. 2024;13(2):e01076–01023. https://doi.org/10.1128/mra.01076-23.
52. Medić AB, Karadžić IM. Pseudomonas in environmental bioremediation of hydrocarbons and phenolic compounds- key catabolic degradation enzymes and new analytical platforms for comprehensive investigation. World J Microbiol Biotechnol. 2022:38(10):165. https://doi.org/10.1007/s11274-022-03349-7
53. Sivasamy S, Rajangam S, Kanagasabai T, Bisht D, Prabhakaran R, Dhandayuthapani S. Biocatalytic Potential of Pseudomonas Species in the Degradation of Polycyclic Aromatic Hydrocarbons. J Basic Microbiol 2025:65(2):e2400448. https://doi.org/10.1002/jobm.202400448
54. Baig ZT, Abbasi SA, Memon AG, Naz A, Soomro AF. Assessment of degradation potential of Pseudomonas species in bioremediating soils contaminated with petroleum hydrocarbons. J Chem Technol Biotechnol. 2022;97(2):455–65. https://doi.org/10.1002/jctb.6820.
55. Arrington EC, Tarn J, Kivenson V, Nunn BL, Liu RM, Paul BG, Valentine DL. Hydrocarbon metabolism and petroleum seepage as ecological and evolutionary drivers for Cycloclasticus. ISME J. 2024. https://doi.org/10.1093/ismejo/wrae247.
56. Vogel AL, Thompson KJ, Straub D, Musat F, Gutierrez T, Kleindienst S. Genetic redundancy in the naphthalene-degradation pathway of Cycloclasticus pugetii strain PS-1 enables response to varying substrate concentrations. FEMS Microbiol Ecol. 2024;100(6). https://doi.org/10.1093/femsec/fiae060.
57. Eze MO. Metagenome Analysis of a Hydrocarbon-Degrading Bacterial Consortium Reveals the Specific Roles of BTEX Biodegraders. Genes. 2021;12(1):98. https://doi.org/10.3390/genes12010098.
58. Cravo-Laureau C, Duran R. Marine coastal sediments microbial hydrocarbon degradation processes: contribution of experimental ecology in the omics'era. Front Microbiol 2014:Volume 5–2014. https://doi.org/10.3389/fmicb.2014.00039
59. Legendre P, Legendre L. Numerical ecology. Volume 24. Elsevier; 2012.
60. Ramette A. Multivariate analyses in microbial ecology. FEMS Microbiol Ecol. 2007;62(2):142–60. 10.1111/j.1574-6941.2007.00375.x.
61. Zhang S, Hu Z, Wang H. A Retrospective Review of Microbiological Methods Applied in Studies Following the Deepwater Horizon Oil Spill. Front Microbiol 2018:Volume 9–2018. https://doi.org/10.3389/fmicb.2018.00520
62. DeLeo DM, Herrera S, Lengyel SD, Quattrini Andrea M, Kulathinal RJ, Cordes EE. Gene expression profiling reveals deep-sea coral response to the Deepwater Horizon oil spill. Mol Ecol 2018:27(20):4066–77. https://doi.org/10.1111/mec.14847
63. Lamendella R, Strutt S, Borglin S, Chakraborty R, Tas N, Mason O, Hultman J, Prestat E, Hazen T, Jansson J. Assessment of the Deepwater Horizon oil spill impact on Gulf coast microbial

communities. Front Microbiol. 2014:5(130). https://doi.org/10.3389/fmicb.2014.00130

64. Song Q, Zhou B, Song Y, Du X, Chen H, Zuo R, Zheng J, Yang T, Sang Y, Li J. Microbial community dynamics and bioremediation strategies for petroleum contamination in an in-service oil Depot, middle-lower Yellow River Basin. Front Microbiol 2025:Volume 16–2025. https://doi.org/10.3389/fmicb.2025.1544233
65. Sun W, Li J, Jiang L, Sun Z, Fu M, Peng X. Profiling microbial community structures across six large oilfields in China and the potential role of dominant microorganisms in bioremediation. Appl Microbiol Biotechnol 2015:99(20):8751–64. https://doi.org/10.1007/s00253-015-6748-1
66. Zhao M-Y, Wang S, Li F-M, Guo S-H, Gao P. Petroleum pollution and microbial community structure in the soil of Liaohe Oilfield. Ying Yong Sheng Tai Xue Bao. 2020;31(12):4215–24. https://doi.org/10.13287/j.1001-9332.202012.040.
67. Wang J, L. Y, and, Huang LX. Study on Enhancing Oil Recovery by Indigenous Microorganisms in a Hypersalinity Reservoir. Pet Sci Technol. 2011;29(14):1504–11. https://doi.org/10.1080/10916466.2010.503475.
68. Liu Q, He W, Zhang W, Wang L, Tang J. Metagenomic analysis reveals the microbial response to petroleum contamination in oilfield soils. Sci Total Environ. 2024:912168972. https://doi.org/10.1016/j.scitotenv.2023.168972
69. Wang H, Kuang S, Lang Q, Wang L. Bacterial community structure of aged oil sludge contaminated soil revealed by illumina high-throughput sequencing in East China. World J Microbiol Biotechnol 2021:37(11):183. https://doi.org/10.1007/s11274-021-03059-6
70. Hui L, Ai M, Han S, Guo Y, Li N, Shi R, Zeng X, Leng S, Zhao J, Liu X, et al. Microbial diversity and functionally distinct groups in produced water from the Daqing Oilfield, China. Pet Sci. 2012;9(4):469–84. https://doi.org/10.1007/s12182-012-0232-7.

Figures

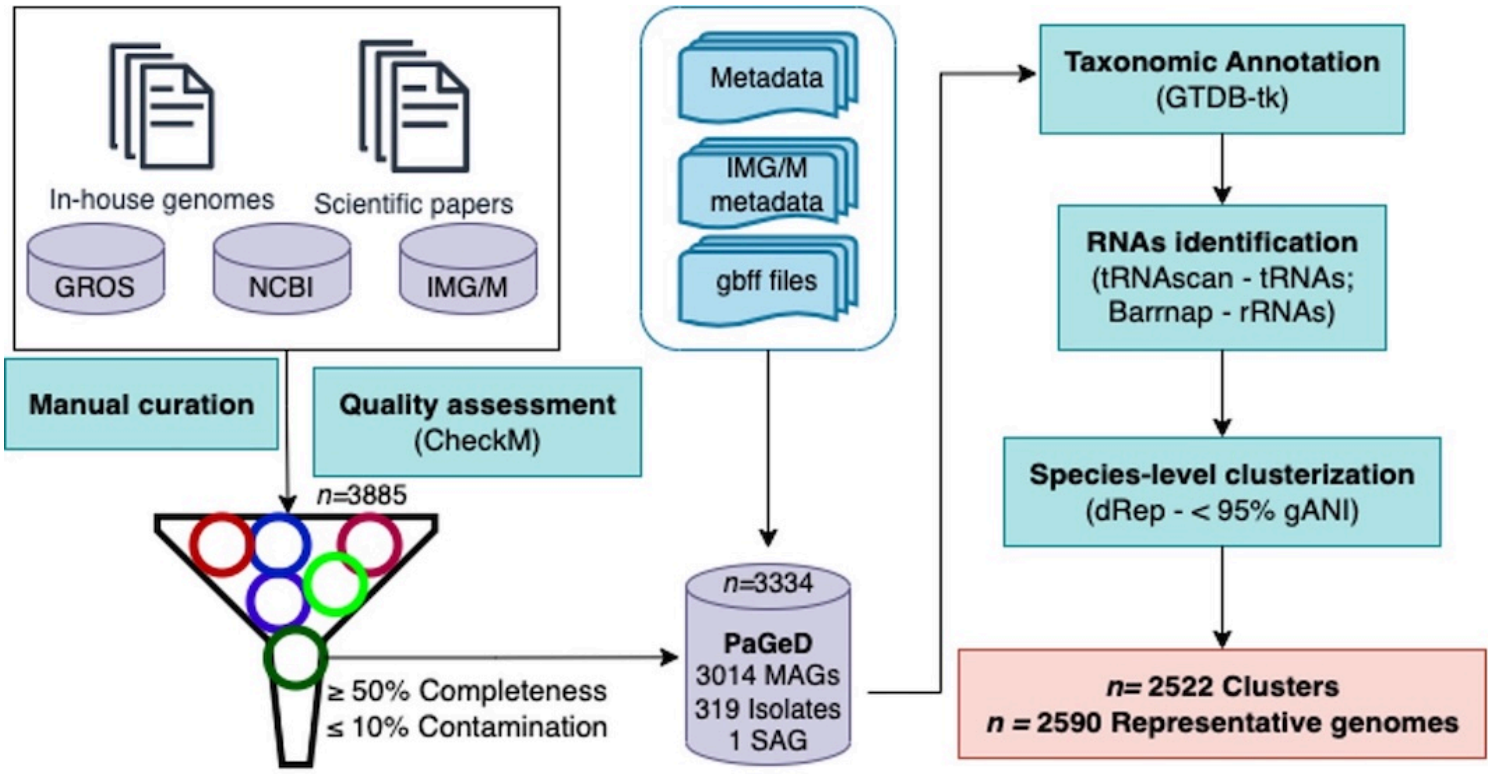


Figure 1

Workflow for genome retrieval, filtering, annotation, and dereplication to construct the PaGeD.

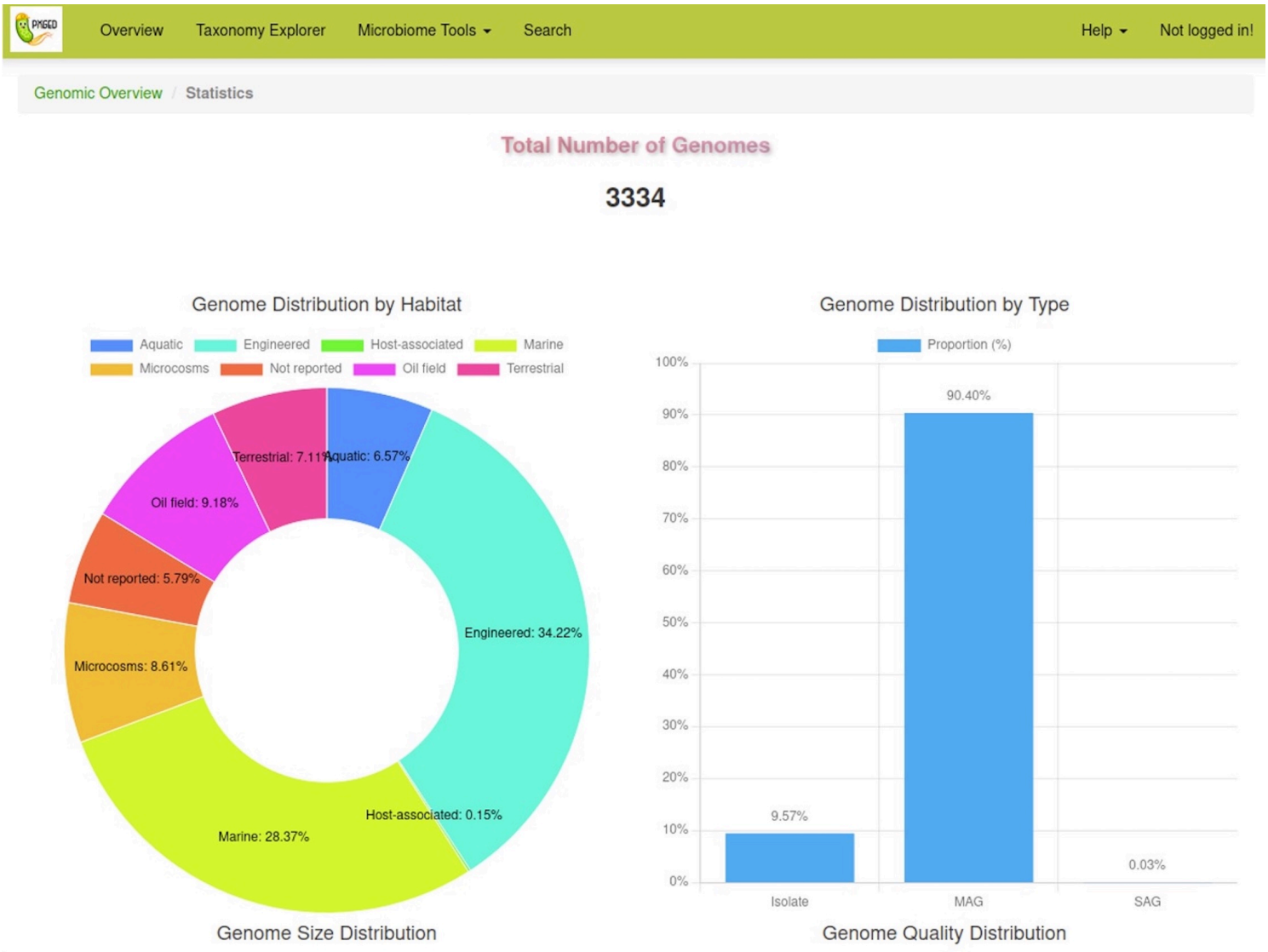


Figure 2

Screenshot of the PaGeD platform's interactive dashboard showing the distribution of genomes across various habitats. The visualization, generated from associated metadata, categorizes genomes by isolation source and habitat type, providing an intuitive overview of microbial diversity in oil-associated environments (URL: https://paged.cpqba.unicamp.br/overview/paged_

statistics/view).

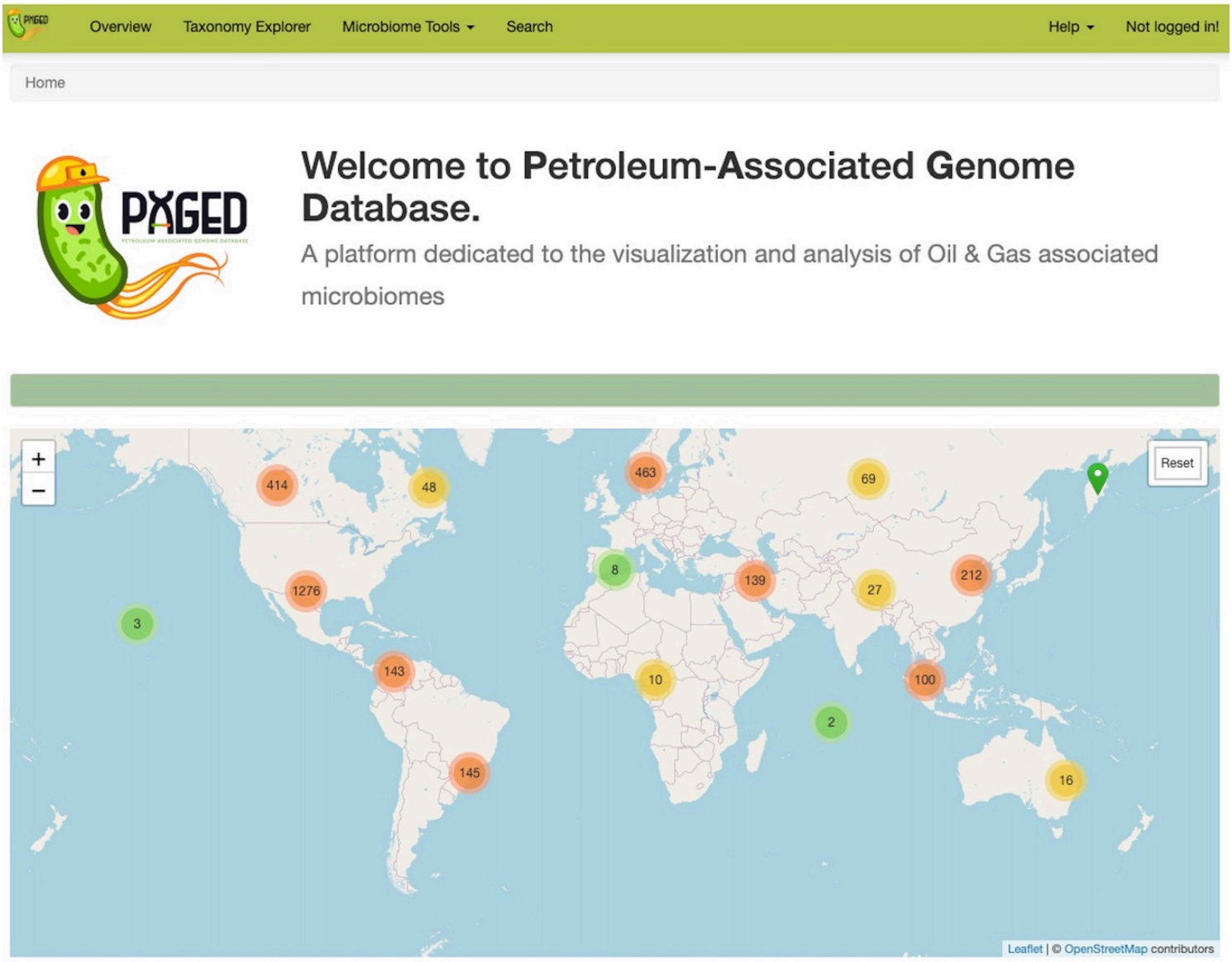


Figure 3

Screenshot of the PaGeD homepage displaying an interactive map of genome geographic origin across 74 countries, six continents, and major oceanic regions (URL: https://paged.cpqba.unicamp.br/).

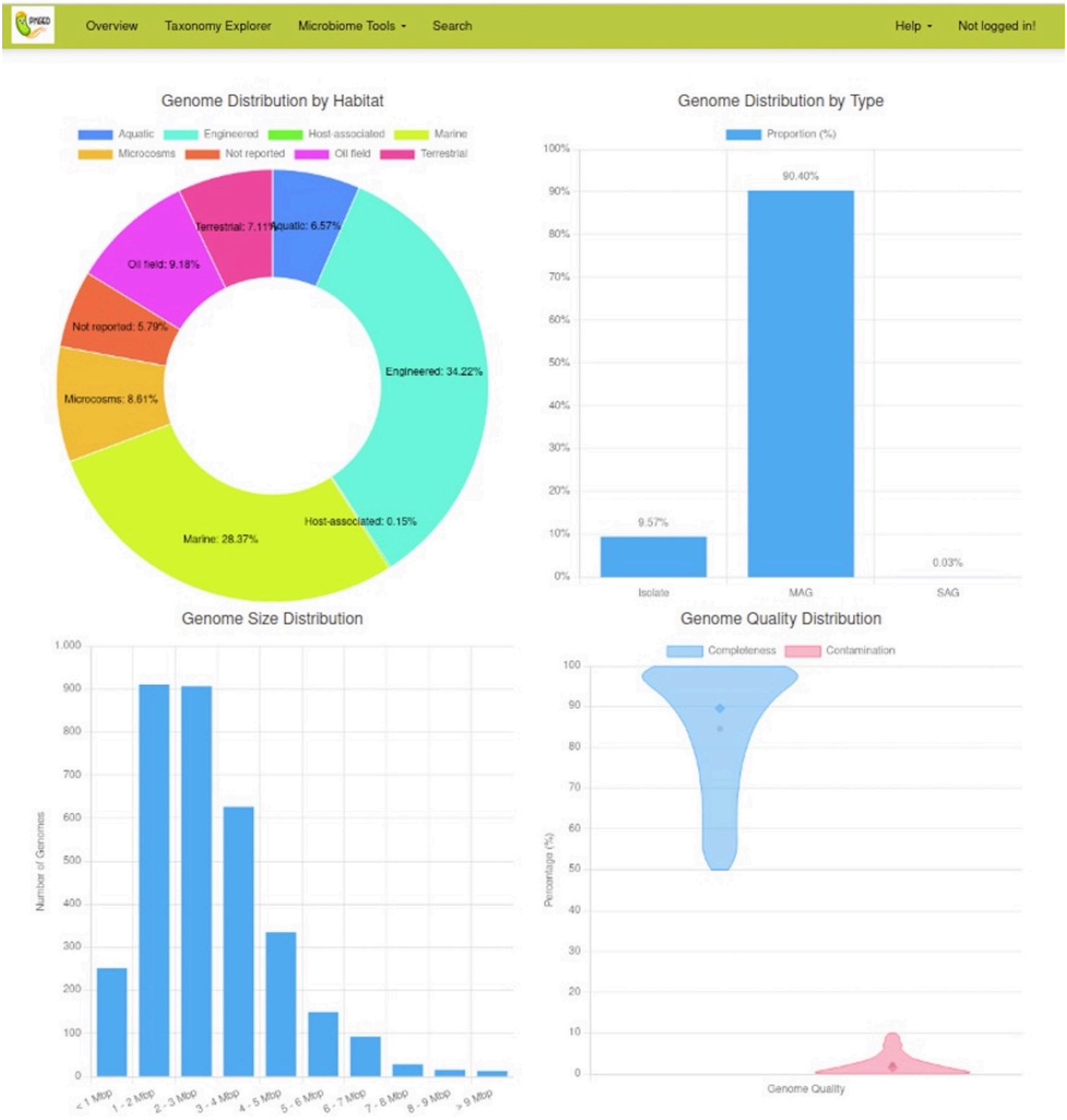


Figure 4

Screenshot of the PaGeD platform displaying quality metrics for the genome dataset. The visualization includes a barplot for genome type distribution, a histogram for genome size distribution, and a violin plot summarizing quality metrics (completeness and contamination. See Table S4 for additional quality statistics (URL: https://paged.cpqba.unicamp.br/overview/paged_statistics/view).

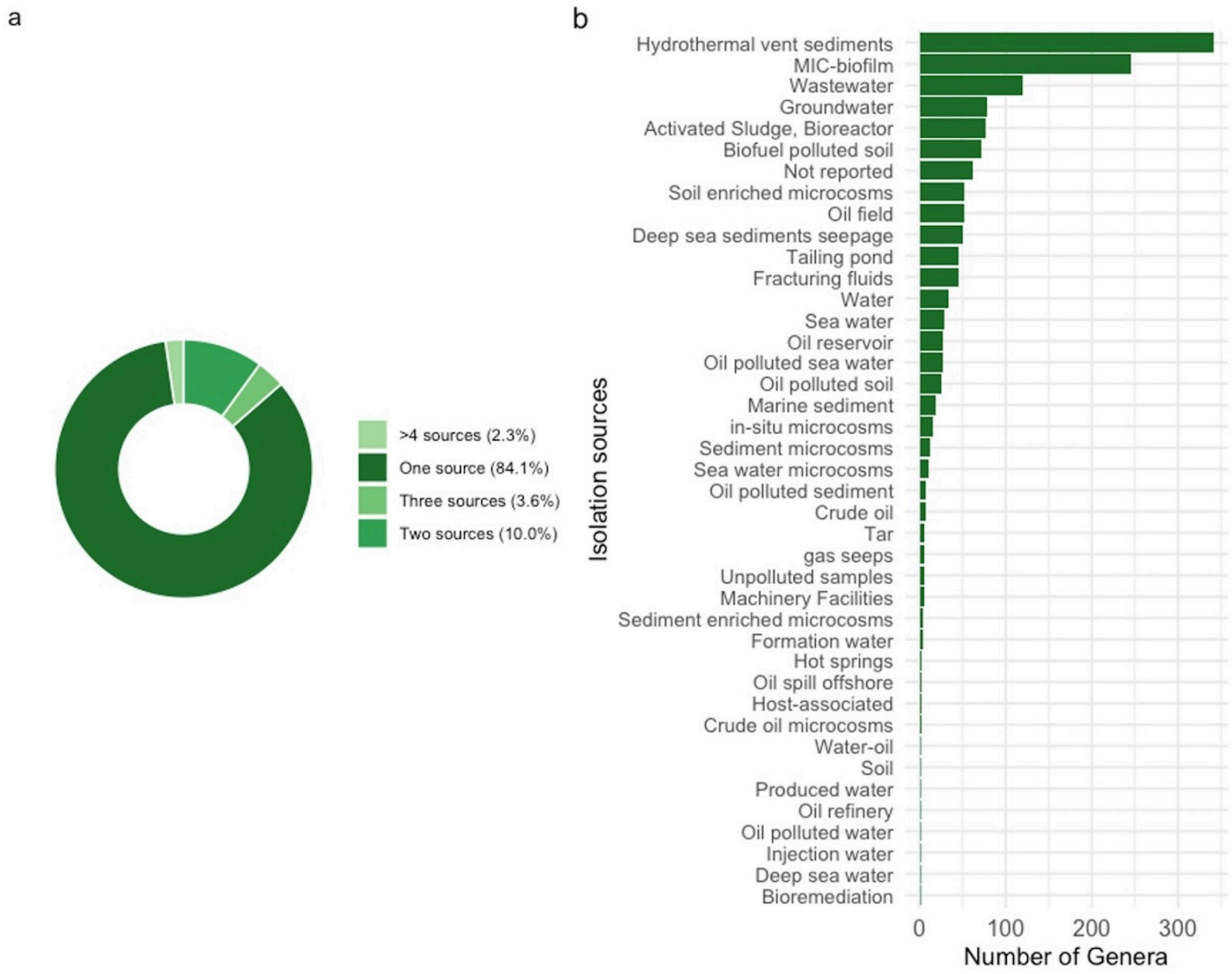


Figure 5

(a) Percentage of genera according to the number of isolation sources. (b) Distribution of genera found in only one isolation source according to the type of environment.



Genome ID	Genome Type	Length	N50	Gc Perc	Num Contigs	Completeness	Contamination	Quality	Representative	Country	Local
PaGeD_00089	Isolate	6349902	6349902	37.30	1	99.33	0.11	High-Quality	False	Colombia	Arauca: Cano Limc
PaGeD_00090	Isolate	6401521	6401521	37.30	1	98.78	0.11	High-Quality	False	Colombia	Arauca: Cano Limc
PaGeD_00095	Isolate	6368297	6368297	37.30	1	99.17	0.11	High-Quality	False	Colombia	Arauca: Cano Limc
PaGeD_01548	Isolate	6776640	6776640	37.00	376	99.29	1.29	High-Quality	False	China	Daqing
PaGeD_01549	Isolate	6050470	6050470	37.30	235	99.66	0.11	High-Quality	False	India	Lakowa Geleky, As
PaGeD_01565	Isolate	6045648	6045648	37.30	187	98.01	0.11	High-Quality	False	India	Saurashtra Region
PaGeD_01594	Isolate	6795250	6795250	37.10	2	99.63	0.28	High-Quality	False	China	Xinjiang
PaGeD_01607	Isolate	6959251	6959251	37.00	2	99.68	1.73	High-Quality	False	China	N/A
PaGeD_01687	Isolate	6315819	6315819	37.30	71	99.68	0.14	High-Quality	False	China	Nankai University
PaGeD_01710	Isolate	7146312	7146312	37.00	81	95.13	6.88	Medium-Quality	False	Mexico	Gulf of Mexico
PaGeD_01713	Isolate	6501414	6501414	37.30	1	99.68	1.08	High-Quality	False	China	Daqing
PaGeD_01714	Isolate	6481697	6481697	37.30	1	99.68	0.11	High-Quality	True	China	Daqing
PaGeD_01715	Isolate	6486336	6486336	37.30	1	99.68	0.67	High-Quality	False	China	Daqing
PaGeD_01805	Isolate	6333060	6333060	37.30	94	99.68	0.11	High-Quality	False	India	Paradip Port Odish
PaGeD_02067	Isolate	6886787	6886787	37.10	60	99.68	0.32	High-Quality	False	Brazil	Rio de Janeiro

Figure 6

Species page for *Pseudomonas aeruginosa*. This page provides comprehensive information of the species, including its associated genomes, geographic distribution (displayed via an interactive map), and genome type (e.g., metagenome-assembled genomes vs. isolate genomes). Additionally, habitat details and direct links to public accession numbers (p.e NCBI, JGI) are available for further validation and exploration of the genomic data (URL: https://paged.cpqba.unicamp.br/taxonomy_

explorer/species_genomes/1621).

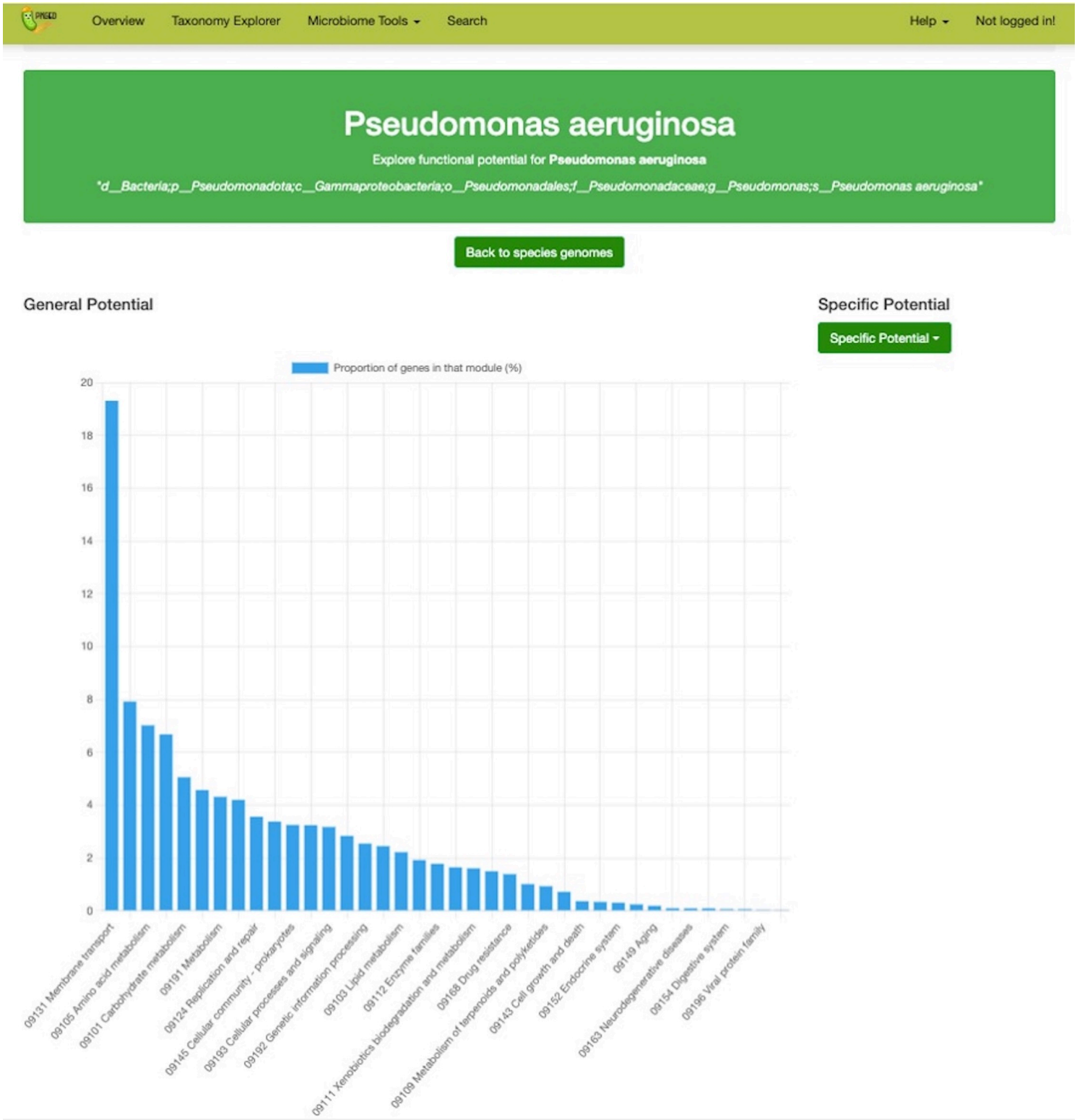


Figure 7

Screenshot of the PaGeD web platform displaying the General Functional Potential interface for *Pseudomonas aeruginosa*. The bar plot shows the relative proportion of genes assigned to KEGG Level 2 functional categories, providing an overview of the metabolic capacities of this species (URL: https://paged.cpqba.unicamp.br/taxonomy_explorer/species_

genomes/1621/potential/general_potential).

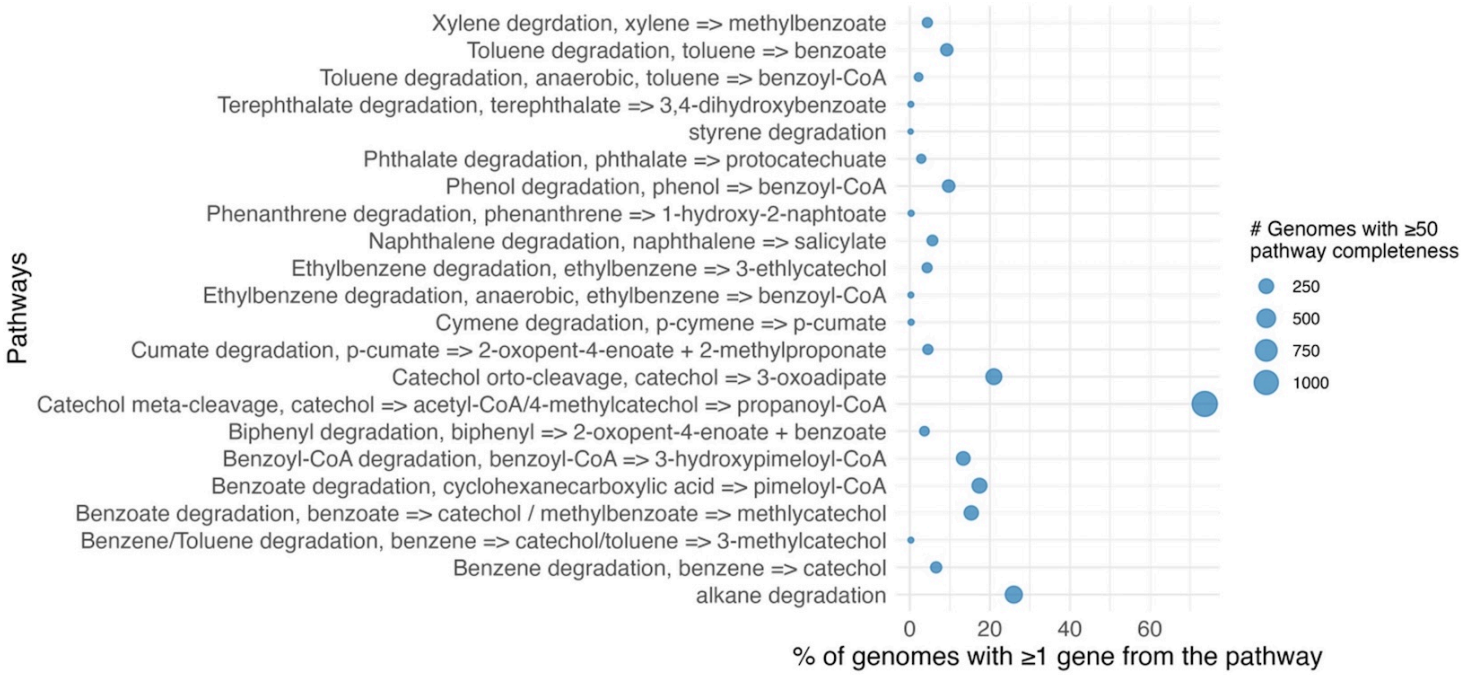


Figure 8

Distribution of hydrocarbon degradation pathways across genomes in PaGeD. The Y-axis lists metabolic pathways involved in hydrocarbon degradation. The X-axis shows the percentage of genomes that encode at least one gene from each pathway. Bubble size indicates the number of genomes with ³50% pathway completeness, providing a combined view of pathway prevalence and functional richness across the dataset.

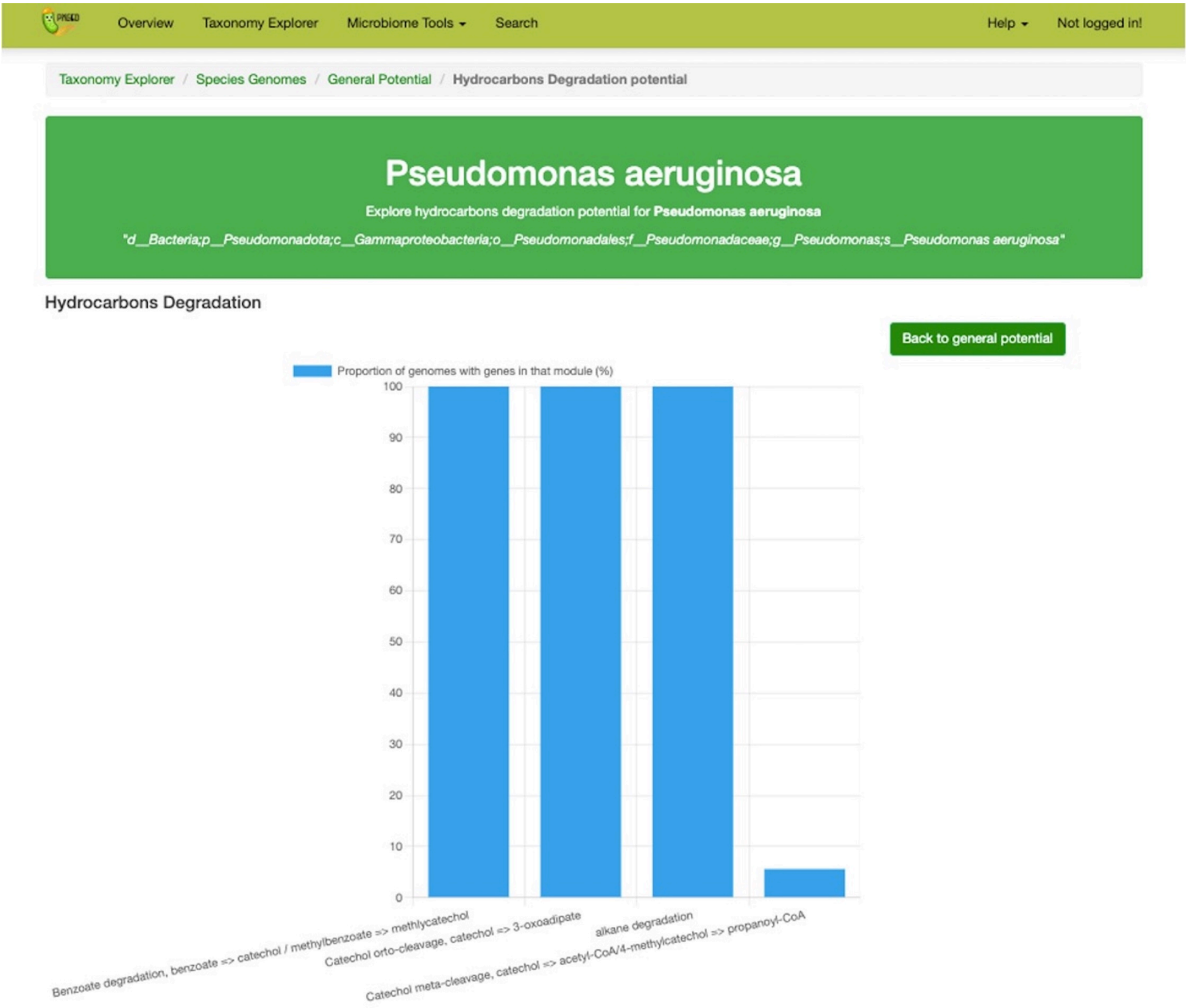


Figure 9

Screenshot of the PaGeD web platform displaying the Hydrocarbon Degradation Potential interface for *Pseudomonas aeruginosa*. The bar plot represents the completeness of curated KEGG modules associated with hydrocarbon degradation pathways, expressed as the percentage of pathway genes encoded by this species (URL: https://paged.cpqba.unicamp.br/taxonomy_

explorer/species_genomes/1621/potential/deghyd).

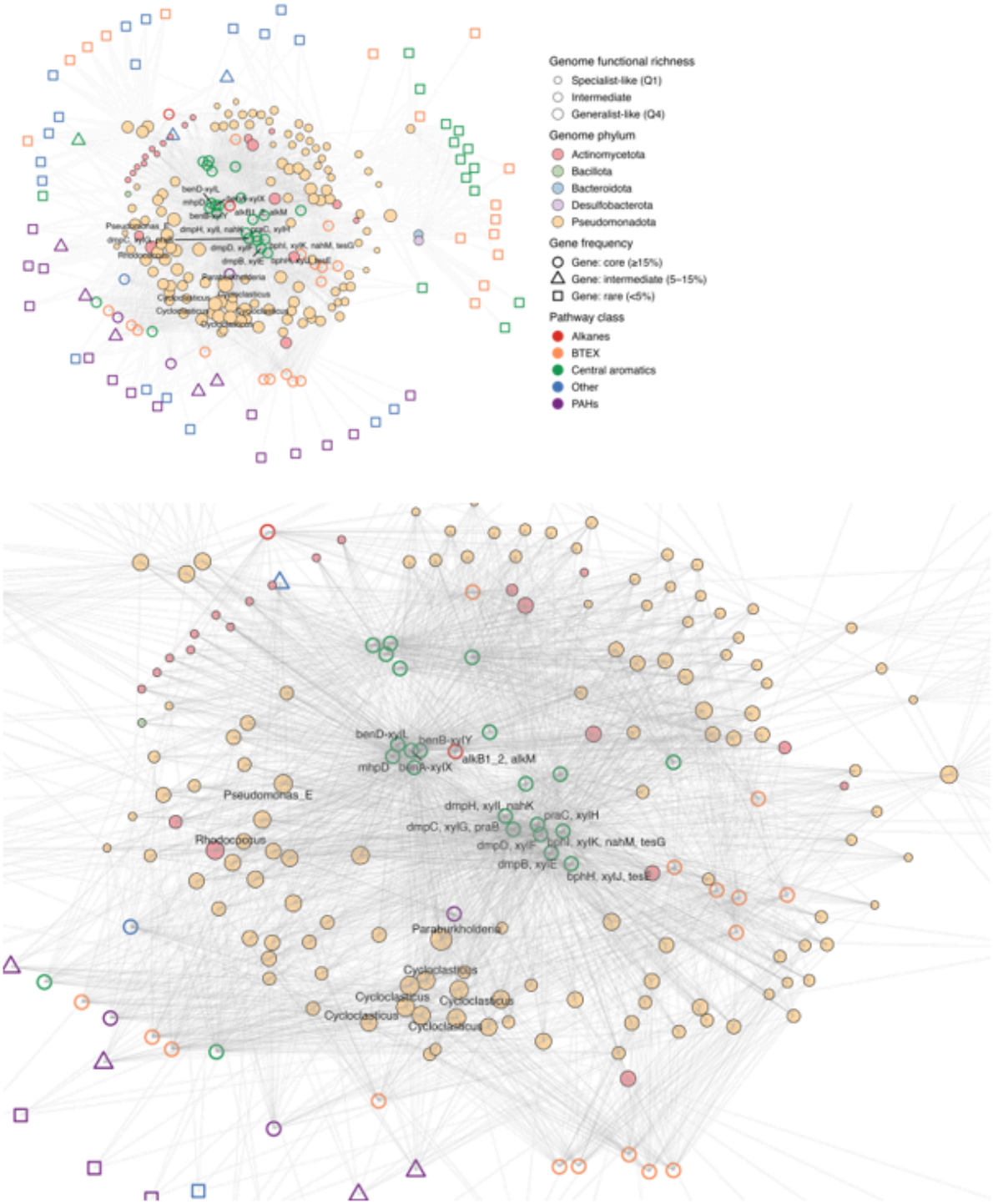


Figure 10

Bipartite network linking hydrocarbon degradation genes and genomes in PaGeD.
(A) Global network showing the organization of hydrocarbon degradation functions across all genomes encoding ≥15 degradation-related KOs. Nodes represent either genomes (filled circles) or genes (hollow symbols). Genome nodes are colored by phylum and scaled by functional richness (number of hydrocarbon-degradation KOs), distinguishing specialist-like (Q1) from generalist-like (Q4) genomes.

Gene nodes are colored by pathway class and shaped according to their frequency across genomes (core ≥15%, intermediate 5–15%, rare <5%). Edges indicate the presence of a gene in a genome. (B) Zoom into the functional core of the network, highlighting core genes and generalist-like genomes. This view reveals a dense hub dominated by central aromatic pathways (e.g., catechol and benzoyl-CoA degradation) and highly connected taxa such as *Pseudomonas*, *Cycloclasticus*, *Paraburkholderia*, and *Rhodococcus*. The contrast between the dense central core and the sparse periphery illustrates the coexistence of broadly conserved central catabolic functions and specialized, substrate-specific pathways. Together, these panels depict a modular architecture in which a small set of highly connected genes and genomes form a functional backbone, while specialist degraders and rare genes occupy peripheral positions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Hidalgoetal2026supplementarymaterialenvmicversion.docx
- Hidalgoetal2026PaGeDsupplementarytables.xlsx