AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

RESEARCH ARTICLE

Comparative Genomics of the Genus *Pseudomonas* Reveals Host- and Environment-Specific Evolution

Zaki Saati-Santamaría,[a,b,c] Riccardo Baroncelli,[d] Raúl Rivas,[a,b,e] Paula García-Fraile[a,b,e]

[a]Departamento de Microbiología y Genética, Universidad de Salamanca, Salamanca, Spain
[b]Institute for Agribiotechnology Research (CIALE), Villamayor, Salamanca, Spain
[c]Institute of Microbiology of the Czech Academy of Sciences, Vídeňská, Prague, Czech Republic
[d]Department of Agricultural and Food Sciences (DISTAL), University of Bologna, Bologna, Italy
[e]Associated Research Unit of Plant-Microorganism Interaction, USAL-CSIC (IRNASA), Salamanca, Spain

ABSTRACT Each Earth ecosystem has unique microbial communities. *Pseudomonas* bacteria have evolved to occupy a plethora of different ecological niches, including living hosts, such as animals and plants. Many genes necessary for the *Pseudomonas*-niche interaction and their encoded functions remain unknown. Here, we describe a comparative genomic study of 3,274 genomes with 19,056,667 protein-coding sequences from *Pseudomonas* strains isolated from diverse environments. We detected functional divergence of *Pseudomonas* that depends on the niche. Each group of strains from a certain environment harbored a distinctive set of metabolic pathways or functions. The horizontal transfer of genes, which mainly proceeded between closely related taxa, was dependent on the isolation source. Finally, we detected thousands of undescribed proteins and functions associated with each *Pseudomonas* lifestyle. This research represents an effort to reveal the mechanisms underlying the ecology, pathogenicity, and evolution of *Pseudomonas*, and it will enable clinical, ecological, and biotechnological advances.

IMPORTANCE Microbes play important roles in the health of living beings and in the environment. The knowledge of these functions may be useful for the development of new clinical and biotechnological applications and the restoration and preservation of natural ecosystems. However, most mechanisms implicated in the interaction of microbes with the environment remain poorly understood; thus, this field of research is very important. Here, we try to understand the mechanisms that facilitate the differential adaptation of *Pseudomonas*—a large and ubiquitous bacterial genus—to the environment. We analyzed more than 3,000 *Pseudomonas* genomes and searched for genetic patterns that can be related with their coevolution with different hosts (animals, plants, or fungi) and environments. Our results revealed that thousands of genes and genetic features are associated with each niche. Our data may be useful to develop new technical and theoretical advances in the fields of ecology, health, and industry.

KEYWORDS *Pseudomonas*, environmental microbiology, genomics, host-cell interactions, microbial ecology

Pseudomonas is a large bacterial genus whose members are adapted to live in many diverse biological niches, such as plants (1–3), mammals (4), reptiles (5, 6), insects (7–10), nematodes (11), humans (12), rivers (13, 14), soils (13), and anthropogenic environments (15), among others (16). Due to the ecological, clinical, and biotechnological importance of *Pseudomonas* bacteria, many research efforts target their functions, such as those involved in the modulation of nutrient cycles (17–19) or the production of secondary metabolites (20, 21) and those responsible for their behavior as beneficial or pathogenic commensals of higher hosts (1, 22–25). Despite this information, many genes and metabolic pathways of *Pseudomonas*

Editor Olaya Rendueles Garcia, Institut Pasteur

Address correspondence to Zaki Saati-Santamaría, zakisaati@usal.es.
The authors declare no conflict of interest.

[This article was published on 10 November 2022 with errors in the ordering of the supplemental material files. The supplemental material was updated in the current version, posted on 1 December 2022.]

Received 24 June 2022
Accepted 24 October 2022
Published 10 November 2022

remain undescribed, as does much of the genetic basis of its adaptation and specialization to different lifestyles.

Discovering novel microbial functions is a complex task, for which laborious wet lab experiments are usually required (i.e., *in vivo* transcriptome sequencing [RNA-seq], transposon mutagenesis and phenotype evaluation, and targeted gene editing) (22, 26–29). Even so, these techniques may cause large studies to yield undesired results when the targeted gene/pathway is not properly selected or the first hypothesis is not adequate. Conversely, comparative genomics is rising as a powerful methodology that helps to unveil genes associated with phenotypes or ecological features, reducing research risks and aiding in relevant breakthroughs in understanding microbe mechanisms (30–33). The reduced sequencing costs and the development and easy use of DNA databases have allowed the scientific community to sequence and share thousands of microbial genomes worldwide. Thus, the study of publicly available genomes helps reveal microbial evolution and adaptation in a low-cost and profitable way.

We aimed to explore the potential ecological functions of *Pseudomonas* and its genomic adaptability to diverse lifestyles and to discover novel genes and functions that participate in the interaction of these bacteria with the environment. We created a database of high-quality publicly available genomes of 3,274 *Pseudomonas* strains with known isolation sources. Then, we used pangenomic and comparative genomic strategies to find differential features among genomes grouped by habitat or host. This work is reinforced by the large genome data set of closely related strains used, providing powerful findings that advance the knowledge of *Pseudomonas* ecology and evolution.

RESULTS

Obtaining a curated pangenome from high-quality *Pseudomonas* genomes. With the aim of studying genes or functions related to the adaptation of *Pseudomonas* to different hosts or niches, we retrieved 11,167 *Pseudomonas* genome sequences from public databases. These genome sequences were filtered to retain only high-quality genomes from *Pseudomonas* strains for which the isolation origin is available publicly. Genomes too phylogenetically distant from the remaining *Pseudomonas* genomes (there were some clades that were located in an extremely far branch, see the GitHub repository for more details online at https://github.com/zakisaati/Pseudomonas_pangenome) were also eliminated due to their possible misidentification. This filtering led to 3,274 high-quality genomes with trustworthy metadata (Materials and Methods; Fig. 1; see Table S1 in the supplemental material). These genomes have a mean coding region density of 71%, 129 contigs, an N_{50} value of 1,042,151 and L_{50} value of 15 (Table S1) and were classified in 393 different species according to the Genome Taxonomy Database (GTDB). *Pseudomonas aeruginosa* is the most represented species (n = 1,661), followed by *Pseudomonas avellanae* (n = 177), *Pseudomonas amygdali* (n = 78), and *Pseudomonas syringae* (n = 50). A total of 26 species encompass $\leq$49 and $\geq$10 genomes of the data set. A total of 170 species have $\leq$9 and $\geq$2 representatives. The remaining 192 genomes belong to unique species. Notably, this analysis is based on the GTDB nomenclature, which splits some valid species into several groups, considering that some strains should represent different taxa. In case those groups were categorized into the current valid taxonomic names, a few species would be more represented. For instance, *Pseudomonas fluorescens* groups would sum 103 genomes; *P. syringae*, 93 genomes; *Pseudomonas chlororaphis*, 73 genomes; *Pseudomonas putida*, 48 genomes; and *Pseudomonas stutzeri*, 39 genomes; among others. Of those strains belonging to the same species, only 71 share average nucleotide identity (ANI) values of $>$99,999% (ANI matrixes are available at Zenodo, https://doi.org/10.5281/zenodo.7105218).

We built a *Pseudomonas* pangenome, obtaining 326,707 protein clusters (fasta file available at Zenodo, https://doi.org/10.5281/zenodo.7105218) from a total of 19,056,667 coding sequences (CDSs) (70% similarity, 80% coverage). We found a very narrow core genome comprising 65 genes, while most of the genes shaped the accessory genome (see Table S2 in the supplemental material). Each genome has 1,085 to 1,645 soft-core genes (genes present in the 95% of the genomes), which represents 18 to 43% of the total number of genes.

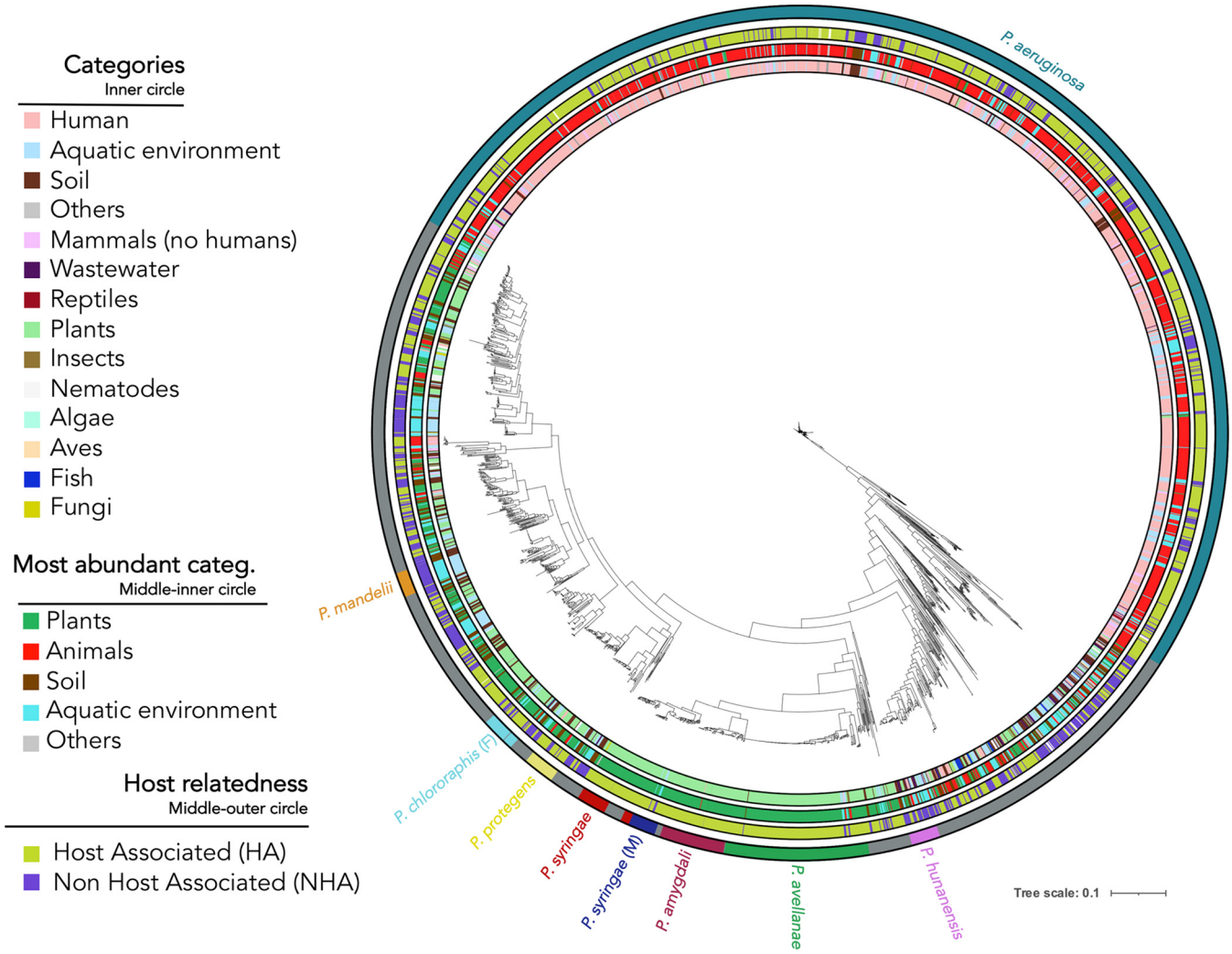


FIG 1 Phylogeny and niche distribution of 3,274 *Pseudomonas* genomes. The phylogenetic tree was built with the 3,274 *Pseudomonas* genomes depicting the isolation source for each strain. Each isolation source is labeled with a different color. There are 3 differently colored circles that represent the isolation sources at different levels. The inner circle includes all the different categories. In the following circle, these categories are merged into 5 categories. Then, the third circle comprises the host relatedness (yes versus no). Finally, the outer circle shows the most represented species (>30 genomes) in this genome collection according to the GTDB.

The pangenome curve did not reach a plateau, suggesting that a fraction of the diversity of *Pseudomonas* genes remains cryptic (curve available at Zenodo, https://doi.org/10.5281/zenodo.7105218).

Divergent niche specialization in *Pseudomonas*. We manually classified the genome collection into several categories based on the isolation source (Fig. 1; see Table S3 in the supplemental material). This classification also included broad categories comprising different isolation sources, such as "animals," "host associated," (HA), and "non-host associated" (NHA).

Next, we constructed a phylogenomic tree to determine whether the different genomic categories showed evolutionary divergence related to the isolation origin. As shown in Fig. 1, some phylogenetic clades, mainly those including human-related and some plant-related *Pseudomonas* genomes, were associated mostly with specific environmental niches. This finding suggests that species from these clades experienced specialization events related to their association with hosts, while some other categories may be represented by less niche-specific species. Indeed, the number of genes per genome differed significantly among the isolation sources (see Fig. S1 in the supplemental material).

This evolutionary tendency was also preserved in regard to the functional (Clusters of Orthologous Genes [COG] content) profiles of *Pseudomonas* (Fig. 2; see Fig. S2 in the

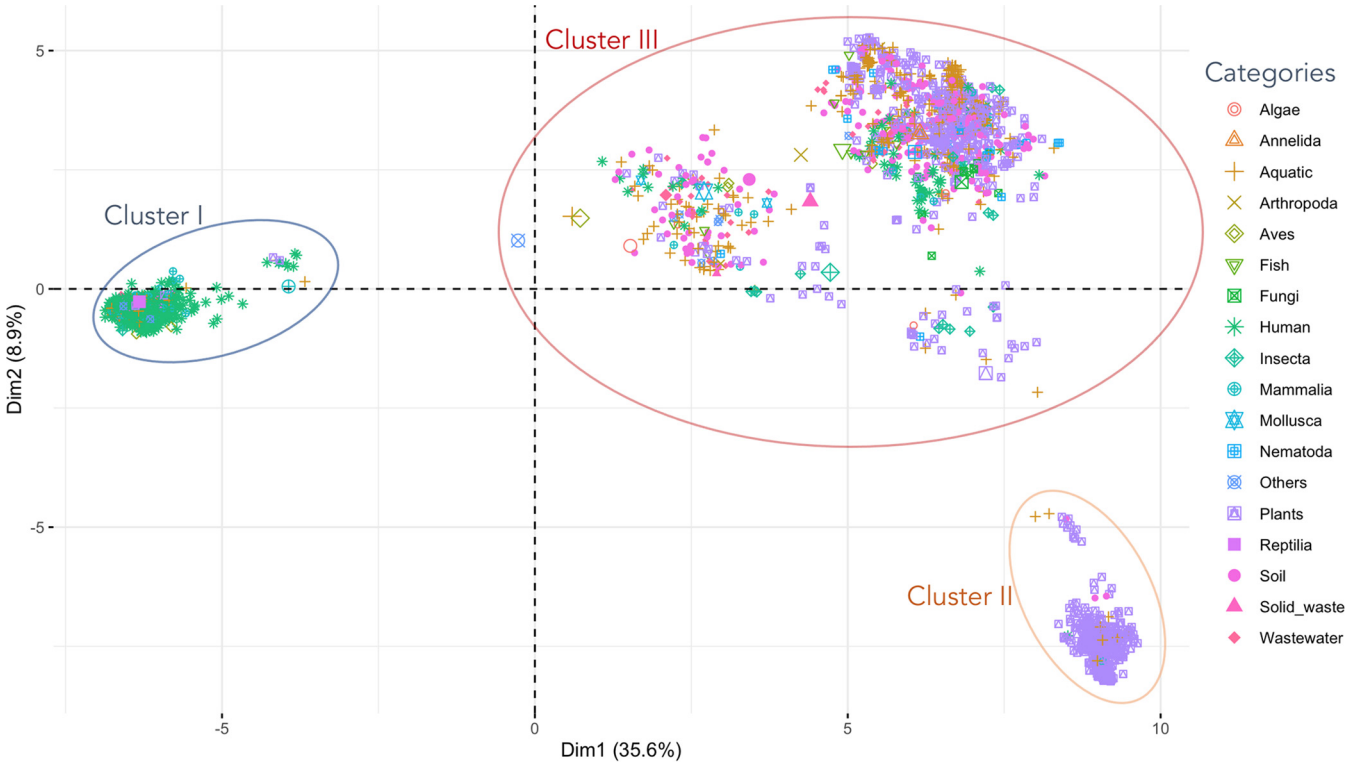


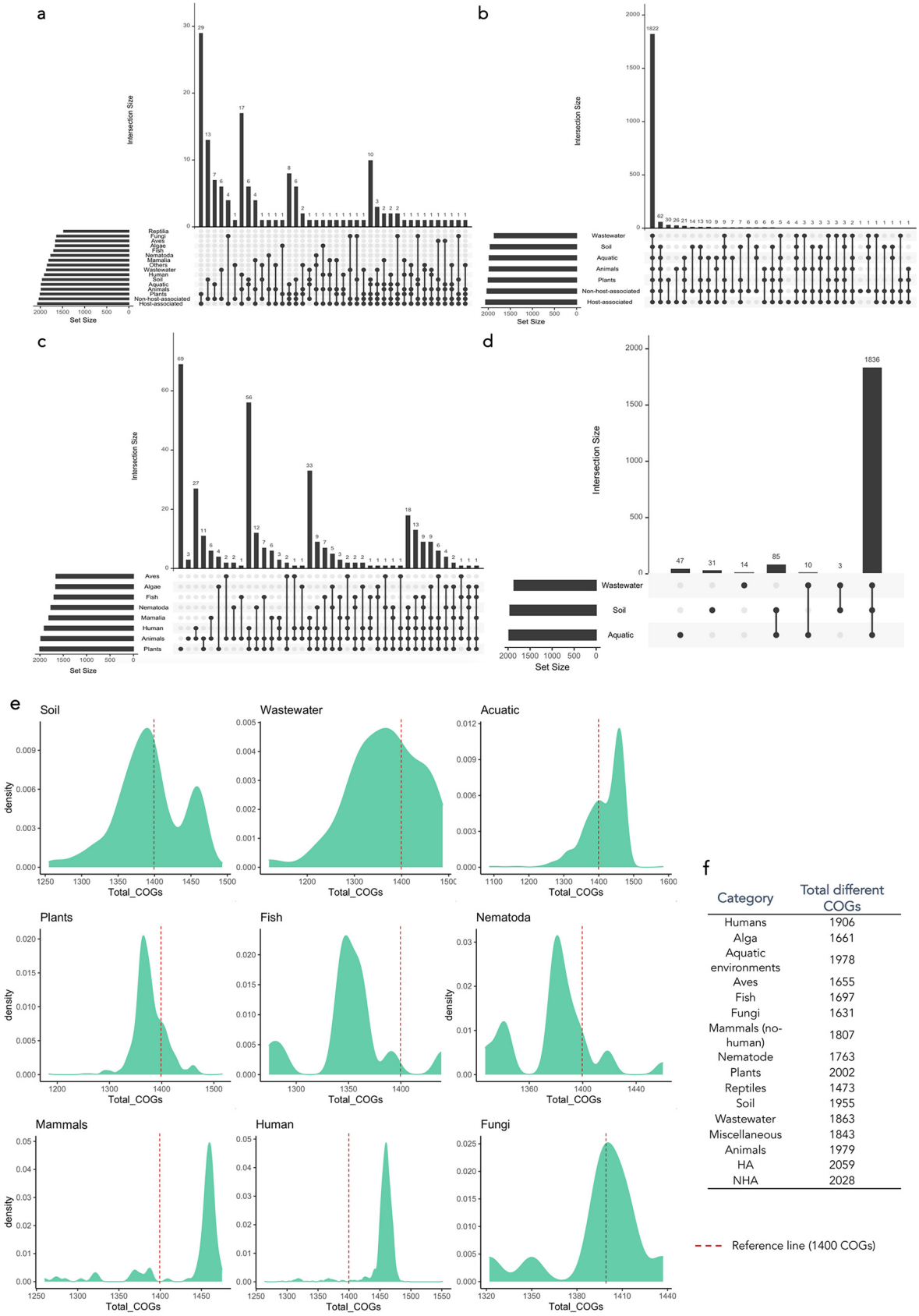
FIG 2 *Pseudomonas* genomes sharing similar habitats have functional resemblance. The principal-component analysis (PCA) is based on the nonredundant presence of Clusters of Orthologous Groups of proteins (COGs) of each of the 3,274 *Pseudomonas* genomes of this study. Each symbol represents an isolation category. We depict 3 main functional clusters of genomes that comprise (i) human-related *Pseudomonas*, (ii) some plant-related *Pseudomonas* (mainly epiphytic and phytopathogenic strains), and (iii) the remaining *Pseudomonas* genomes.

supplemental material). We found 3 main functional clusters of genomes that comprised (i) human-related *Pseudomonas*, (ii) some plant-related *Pseudomonas*, and (iii) the remaining *Pseudomonas* genomes. Interestingly, the second genome cluster included epiphytic and phytopathogenic *Pseudomonas*, while rhizospheric and plant-beneficial *Pseudomonas* members were included in the third and most diverse cluster (Fig. 2).

Diversity of *Pseudomonas* functions. We examined the diversity of nonredundant COG functions within each of the genome categories. HA genomes are slightly more functionally diverse (higher number of nonredundant COGs) than NHA genomes (Tukey honestly significant difference [HSD], adjusted *P* value [p-adj] = 2.22e-16) (Fig. 3; see Table S4 in the supplemental material). The genomes with the highest COG diversity are those within the categories "human" and "mammals." Plant-associated *Pseudomonas* strains are less functionally diverse than those associated with mammals, humans, fungi, and soil (Tukey HSD, p-adj $<$ 0.05); however, this category encompasses the highest diversity of unique functions (2,002 different COGs) (Fig. 3).

Despite the considerable amount of genetic diversity found in the pangenome analyses of *Pseudomonas*, where we found a small core pangenome (65 genes present in all the genomes), 1,822 COGs were shared among the broadest categories (animals, plants, HA, NHA, aquatic environments, soil, and wastewater) (Fig. 3), which implies a large core functional pangenome (COG functions present in all the strains) in *Pseudomonas*. This result means that there is a large genetic variability in those *Pseudomonas* genes that are related with core metabolic functions.

Insights into the metabolism of carbohydrates and proteins. Carbohydrates are the main carbon source for most living organisms. The ability to use carbohydrates present in a certain niche is crucial for the successful adaptation of most bacteria to a given ecosystem. Each host or environment harbors different amounts and diversities of carbohydrates. We aimed to study the potential of the different *Pseudomonas* strains to metabolize carbohydrates. To do so, we annotated the genomes with the Carbohydrate Active EnZymes



Category	Total different COGs
Humans	1906
Alga	1661
Aquatic environments	1978
Aves	1655
Fish	1697
Fungi	1631
Mammals (no-human)	1807
Nematode	1763
Plants	2002
Reptiles	1473
Soil	1955
Wastewater	1863
Miscellaneous	1843
Animals	1979
HA	2059
NHA	2028

FIG 3 The lifestyle of *Pseudomonas* strains is a key factor that determines their metabolic range. (a, b, c, d) Intersection plots showing the number of nonredundant Clusters of Orthologous Groups of proteins (COGs) in each category and the unique COGs shared by two or more categories. (e) Density plots relating the number of genomes with the COG content. The red dashed line represents 1,400 COGs as a reference to compare different plots. (f) Table of total unique COGs within each isolation source category.

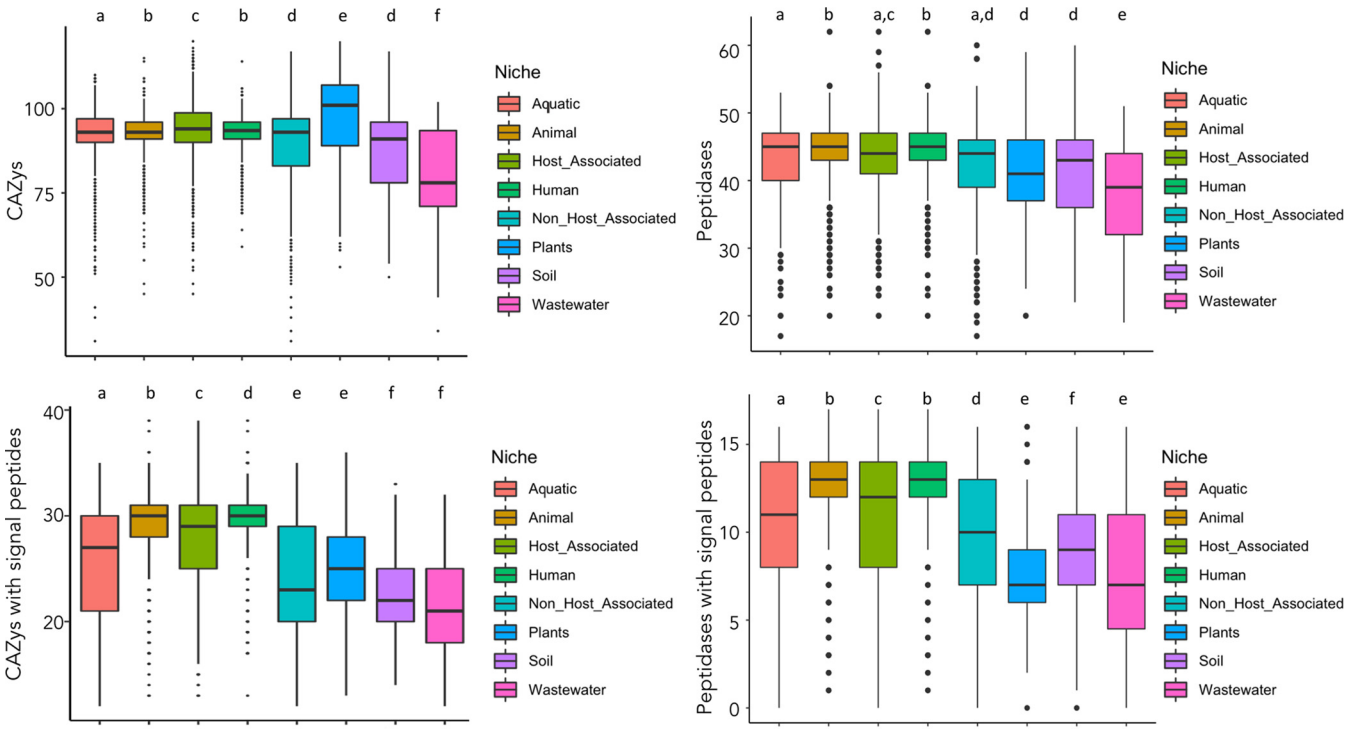


FIG 4 The potential to metabolize carbohydrates and proteins is shaped by the isolation source. This figure includes boxplots that represent the content of either CAZy and peptidase per each genome and isolation source, as well as the enzymes with signal peptides. Different letters represent groups with significant differences (p adj < 0.05).

(CAZy) database, which returned a total of 305,915 proteins involved in carbohydrate metabolism (see Table S5 in the supplemental material). Our results suggest a niche-dependent potential to metabolize these compounds (Fig. 4). The *Pseudomonas* strains isolated from plants, which are hosts with large amounts of complex carbohydrates, showed a higher CAZy content ($>$100 CAZys/genome), while wastewater-associated *Pseudomonas* represented the category with the lowest number of CAZys ($<$80 CAZys/genome) (Fig. 4).

Peptidases catalyze the cleavage of a vast variety of peptides and are associated with many diverse biological functions. Thus, the adaptability of bacteria can be influenced by this proteolytic ability. Here, we show that the lifestyle of the *Pseudomonas* strains significantly influences the peptidase content (Fig. 4). This finding denotes that the mean content of peptidases on the studied strains is significantly related to the isolation source (Tukey HSD; p-adj < 0.05). For example, the strains living in wastewater environments carry a lower number of peptidases than those isolated from other niches (Fig. 4).

Both CAZys and peptidases can be secreted or placed within the outer membrane, enabling a broader set of interactions with the environment and surrounding microbes. Signal peptides are markers that allow proteins to enter secretion pathways. Consequently, the presence of signal peptides on these enzymes can be crucial for the adaptability, fitness, and interaction of *Pseudomonas* within its ecosystem. Interestingly, despite having the highest CAZy content, plant-associated *Pseudomonas* showed far fewer secreted CAZys than the strains belonging to the animal, human, and aquatic categories (Fig. 4). Similarly, the presence of peptidases with signal peptides per genome was the lowest in the group "plants" (Fig. 4). This result suggests that plant-associated *Pseudomonas* strains use most of their degradative arsenal to metabolize carbohydrates and peptides in the cytoplasm in order to obtain energy, carbon, and nitrogen, while a larger proportion of enzymes with signal peptides encoded by the other strains may be dedicated to interactions with the environment.

Stress resistance. In certain environments, bacterial survival requires resistance to environmental threats (heat, presence of antibiotics, metals, and biocides; see Materials and Methods). Thus, we looked for mechanisms enabling *Pseudomonas* to resist these stress conditions, and we analyzed their distribution (see Table S6 in the supplemental material).

TABLE 1 Number of proteins or functions associated with the isolation origin of *Pseudomonas* strains

	No. of:			
Category	Protein clusters[a] (HP[b])	COGs[a]	Resistance-related proteins[c]	CAZys[c]
Aquatic environment	4.841 (3.327)	93	16	145
Fish	663 (472)	3	1	34
Fungi	263 (115)	7	4	53
Humans	7.060 (4.308)	465	58	108
Mammals (nonhuman)	2.563 (1.337)	81	6	64
Nematodes	677 (409)	24	2	54
Plants	10.916 (7.214)	211	12	344
Soil	3.476 (1.767)	142	13	254
Insects	77 (19)	-	2	96
Wastewater	524 (283)	17	7	70
Animals	6.837 (4.202)	431	55	107
Host associated	6.975 (4.173)	315	34	162
Not host associated	9.099 (5.382)	225	20	309

[a]$P < 10^{-6}$ (Benjamini-Hochberg correction).
[b]HP, hypothetical protein.
[c]$P < 10^{-2}$ (Benjamini-Hochberg correction).

Among the *Pseudomonas* strains studied in this work, those isolated from humans possess the most genetic machinery related to the mentioned threats. In contrast, plant-associated strains possessed the smallest number of stress resistance mechanisms, even smaller than the number observed for strains isolated from bulk soil (p adj < 0.05) (see Fig. S3 in the supplemental material), suggesting that the plant environment protects bacteria from environmental stresses or microbiological competition. Overall, HA *Pseudomonas* have a broader resistance potential (mainly antimicrobial resistance) than NHA strains (p adj < 0.05).

Thousands of proteins and functions show specific associations with the environment in *Pseudomonas*. Each microenvironment where *Pseudomonas* bacteria live has unique physicochemical characteristics, available molecules for bacterial nutrition, or, in the case of host-related niches, defense mechanisms. To adapt to the different natural conditions of the niches they inhabit, *Pseudomonas* bacteria should have undergone adaptation events driven by the evolution of their accessory genome, offering some survival advantage. We used a comparative genomics approach to study protein and functional (COGs, resistance-related proteins, and CAZys) enrichment in each niche/isolation category. We used Scoary to detect specific enrichments, yielding thousands of proteins and functions associated with different niches or hosts (Table 1; see Files S3 to S6 in the supplemental material).

We found that a large proportion of the proteins detected with Scoary that seem to be related to the isolation source do not have any functional annotation (hypothetical proteins). Additionally, considering host relatedness, we found that more proteins were associated with the NHA category, while the number/diversity of COGs and resistance-related proteins associated with hosts (HA) was higher. Focusing on clear isolation sources, the genomes isolated from plants were associated with the most proteins. Similarly, human-related strains showed more unique associations of COGs and stress resistance-related proteins. Additionally, *Pseudomonas* strains isolated from plants, soils, and natural aquatic environments were associated with the highest number of CAZys (Table 1; Table S5). Many protein clusters are significantly associated with more than one niche (see Fig. S4 in the supplemental material). For instance, 4,052 proteins are associated with "animals," "humans," "mammalia," and "hosts," probably due to a bias toward the higher number of human isolates present in our animal/mammalia/host data sets which implies an overlap between these niches. Similarly, 165 proteins are significantly associated with both "soil" and "plants" categories, which could be due to isolation or metadata bias, or even due to some role in the rhizosphere environment. We also compared the proteins associated with HA and NHA categories and found substantial similarity (see Fig. S5 in the supplemental material), suggesting that these associations do not include large groups of extremely evolutionarily distant proteins.

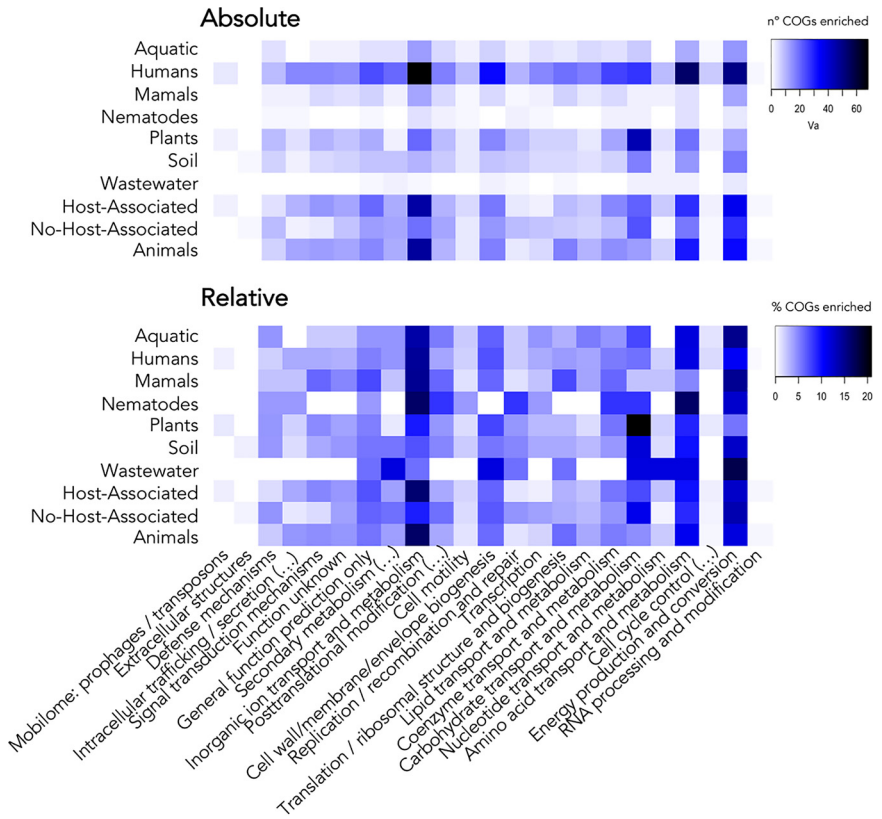


FIG 5 The metabolic profile based on the niche-associated COGs in *Pseudomonas* reveals metabolic categories enriched in each group of genomes. The heatmaps represent the different categories of Clusters of Orthologous Groups of proteins (COGs) enriched in *Pseudomonas* genomes. At the top is shown the sum of significantly enriched COGs falling within a certain category. At the bottom, each count of COGs has been relativized to the total number of COGs associated with each host or niche.

Focusing on the metabolic pathways or metabolism categories in which niche-associated COGs are classified (Fig. 5), we show that carbohydrate metabolism is specifically enriched in plant-associated strains. Additionally, human- and aquatic environment-related *Pseudomonas* strains have a large proportion of enriched COGs associated with the transport and metabolism of inorganic ions. The largest proportion of HA COGs was observed for energy metabolism or the metabolism of inorganic ions and amino acids.

Additionally, among the enriched proteins mentioned above, some have already been suggested to function within *Pseudomonas* habitats. For example, the HCN-ABC operon appears to be associated with humans (34), the metapyrocatechase enzyme (also named catechol 2,3-dioxygenase, responsible for toluene degradation) (35) was found to be enriched in wastewater environments, and 1-aminocyclopropane-1-carboxylic acid (ACC) deaminase is enriched in *Pseudomonas* strains isolated from plant environments (36) (File S3). Additionally, we found proteins involved in vitamin B12 synthesis associated with insect niches; B12 supply has been suggested to play an endosymbiont role in insects (37, 38), but this role has never been proven in *Pseudomonas*.

In contrast, many of the genes found to be enriched in some groups of *Pseudomonas* have never been described as relevant for interactions with the environment or the host. For instance, we found that a protein similar to the biofilm dispersion protein (BdlA) is strongly enriched in those isolates associated with nematodes (Benjamini-Hochberg's [B-H] p-adj = 4.39×10^{-10}), which suggests its potential importance for the removal of biofilm in these hosts. Similarly, the YdeP protein, which is suggested to be related with acid resistance, is enriched in the isolates associated with insects (B-H p-adj = 7.86×10^{-14}). The protein with the best significant association to plant-niche encodes an L-glyceraldehyde 3-phosphate reductase (B-H p-adj = 2.11×10^{-305}). Similarly, the enrichment of a blue-light photoreceptor

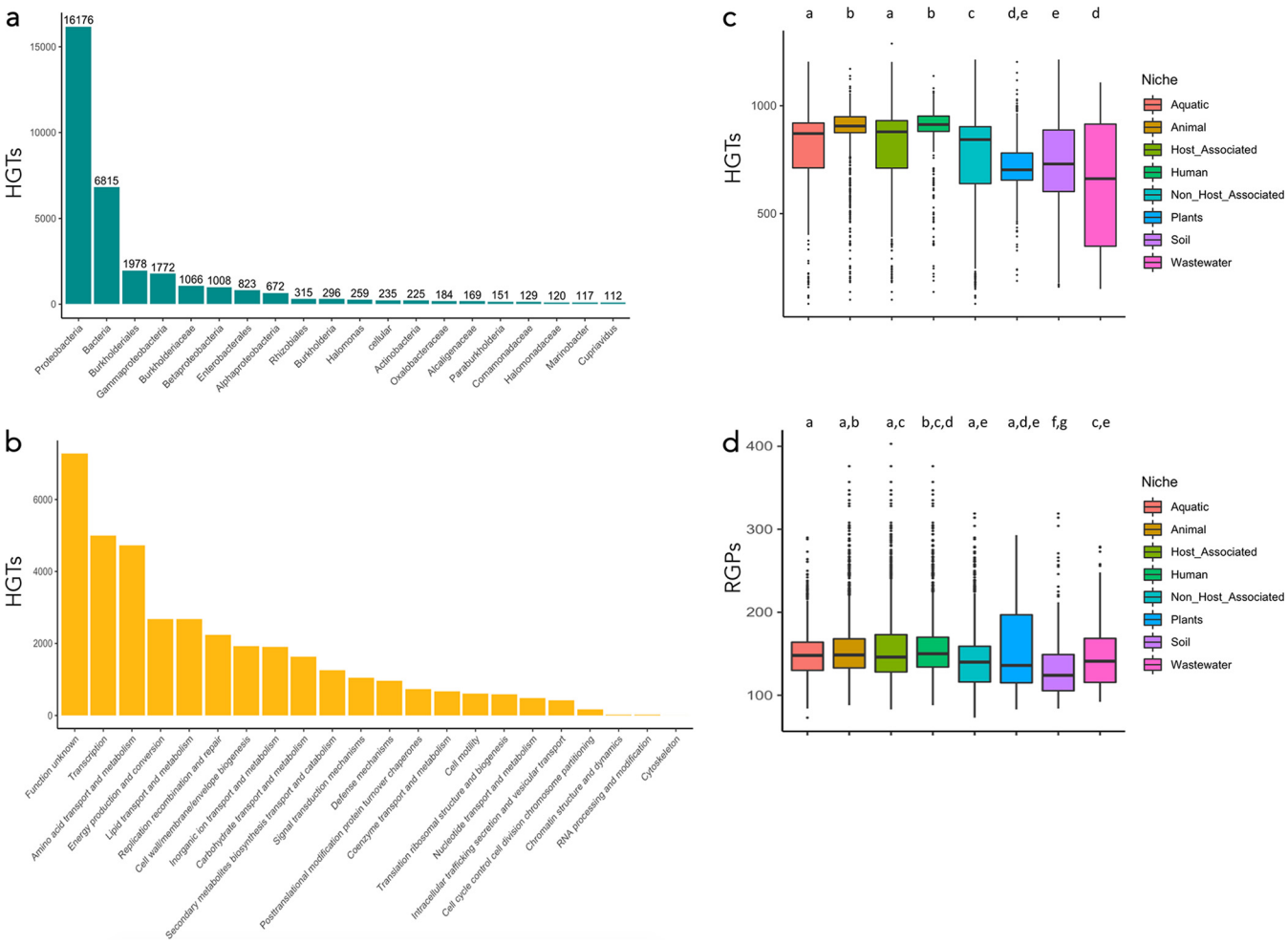


FIG 6 Gene dynamics in the *Pseudomonas* genomes. (a) This figure represents the number of detected HGT genes classified based on the taxa that is supposed to be the source of the horizontal transfer. Only the 20 most abundant taxa are displayed. (b) Bars represent the number of genes derived from HGT events according to the categories in which the encoded protein is classified based on the COG annotation. (c, d) Boxplots showing the number of HGT and RGP events per genome and classified into the different isolation sources. Different letters represent groups with significant differences (p adj < 0.05).

in soil associated pseudomonads (B-H p-adj = 8.57×10^{-51}) might be linked with the dynamics or the response of bacterial cells to the soil depth. More examples are the YjcH protein and a xanthine permease in human-associated *Pseudomonas*, the transcriptional regulator *XynR* in plant-associated strains, and an operon homologous to the Yop virulon in fish-related strains, or in a broader sense, any of the thousands of hypothetical proteins associated with each niche.

Genetic dynamics. Since horizontal gene transfer (HGT) influences bacterial adaptability to novel environmental conditions, we searched for *Pseudomonas* genes that may have been acquired from other microbial taxa via this process. A total of 11.5% of the representative sequences of the *Pseudomonas* pangenome were detected as potential horizontally transferred genes. Of these genes, the majority could have a bacterial origin, primarily from *Proteobacteria*, followed by *Burkholderiales* (Fig. 6). Additionally, there are also proteins for which the origin is suggested to be domain *Archaea*, superphylum *Terrabacteria*, or phylum *Firmicutes*, among other groups, even though these proteins are present in a smaller proportion within the pangenome (Fig. 6).

The composition of the different microbiomes or the features of a certain niche may imply differential dynamics in the transfer of genes. In this regard, HA strains of *Pseudomonas* showed slightly higher gene transfer rates than NHA strains (Fig. 6), which may be influenced by the antibiotic resistance spread of *P. aeruginosa* (HA) strains. The most notable difference is that between the animal (mean, 885 HGTs/genome) and plant (mean, 718 HGTs/genome)

categories (Tukey HSD p-adj = 1.62×10^{-11}). The *Pseudomonas* strains isolated from soils and wastewater showed high variability in their HGT content (Fig. 6).

With regard to the functional annotation of the representative genes/proteins of the pangenome likely acquired by HGT events, 134 were annotated as resistance-related proteins (55.4% of the total number of resistance-related proteins detected). Moreover, several invertases (proteins that switch antibiotic-resistance regulatory genes, among other functions [39]) appear to be acquired from *Klebsiella* species. Similarly, 850 of 3,245 CAZys would have been horizontally acquired. Additionally, the classification of these proteins within COG categories reveals that most of them do not have an assigned functional category, whereas the assigned proteins belong mainly to the "transcription," "amino acid and transport metabolism," "energy production and conversion," and "lipid transport and metabolism" categories (Fig. 6).

Commonly, clusters of genes acquired by HGT are located in regions of genome plasticity (RGPs), such as genomic islands or plasmids. We looked for RGPs through the graphical and partition-driven methods implemented in PPanGGOLiN (Materials and Methods). We detected 499,509 RPGs in total (152,6 RGPs/genome). We found that lifestyle impacts the RGP content per genome (Fig. 6). The *Pseudomonas* strains isolated from soils showed the fewest RGPs, those isolated from humans had intermediate numbers of RGPs/genome, and those isolated from plants had the most RGPs.

DISCUSSION

Earth ecosystems and living organisms (hosts such as animals or plants) are strongly influenced by their microbiomes (40–43). Thus, understanding the functions of bacteria in their biological niches is of utmost importance and can be profitable for the study of animal or plant illnesses and for the development of biofertilizers and bioremediation agents, among others. (3, 9, 27, 44, 45). A common integrant of the microbiomes inhabiting many diverse habitats and hosts is the genus *Pseudomonas* (16). The variability of niches where *Pseudomonas* can survive makes the study of both the ecological functions and the metabolic potential of these bacteria very appealing. To perform deep analyses of the ecological functions of bacteria belonging to this genus, we implemented a comparative genomics study of more than 3,000 *Pseudomonas* strains, revealing their probable roles in their isolation niches and providing an understanding of their likely evolution and adaptation mechanisms.

We found that plant-associated *Pseudomonas* strains have smaller genomes than other strains of this genus. This finding may indicate a more intimate cross-kingdom interaction and specialization within these hosts (46). Nonetheless, this tendency does not stand out when HA and NHA genomes are compared. We also found that HGT events affect 10 to 20% of the genes in each *Pseudomonas* genome. Most of these events involved bacteria from the same phylum (*Proteobacteria*), which usually represents a relevant proportion of the bacterial communities of the common habitats of *Pseudomonas* (47–52). More than one-half of the detected clusters of resistance genes could have been horizontally acquired. This finding supports those of Freschi et al. (53), who showed that HGT events drive the gain of antibiotic resistance and virulence in *P. aeruginosa*.

Our findings indicate that strains associated with humans have a larger number of resistance-related genes than other strains. Nevertheless, this difference could be the result of biased annotation; while many *P. aeruginosa* resistance genes are included in the available databases, the resistance genes belonging to less-studied environmental *Pseudomonas* strains may have never been elucidated.

Here, we show that plant-related *Pseudomonas* strains have enriched genetic machinery for carbohydrate metabolism, probably to adapt to the complex carbohydrate content in the plant environment (54, 55). Levy et al. (56) reached similar conclusions when comparing genomes belonging to different phyla associated or not associated with plants. The enrichment of signal peptides among CAZys encoded by animal-related *Pseudomonas* strains may be due to the involvement of these enzymes in biofilm formation and/or the degradation of host tissues. In contrast, the CAZys of strains isolated from plant hosts seem to be located mainly in the cytoplasm, suggesting an enhanced ability to metabolize

carbohydrates for use as a C or energy source. Similarly, we found variation in the content of secreted peptidases in the categories of genomes examined in this work, suggesting a role of these proteins in the lifestyle adaptations of *Pseudomonas*. These findings agree with those of Nguyen et al. (57), who found segregation of extracellular peptidases of bacteria and archaea according to their habitats.

We found an enrichment of inorganic ion metabolism in human-associated *Pseudomonas*, which was due mainly to the enrichment of proteins related to H^+ transport and iron metabolism. The first process may aid in the adaptability of *P. aeruginosa* to the variability of H^+ levels in lung tissues (58, 59). Additionally, the enrichment of iron-related functions is related to siderophore production by this pathogenic species, which also serves as a mechanism of niche competition (60).

Adding to previous comparative genomic works aiming to find undescribed microbial genes with ecological functions (31–33, 61, 62), we discovered thousands of such genes that may have key functions in the environmental interaction or the adaptation of *Pseudomonas* (File S3 to S6). Interestingly, the annotation processes identified many of them as hypothetical proteins. Here, we provide a catalogue of genes with likely relevant roles in the lifestyle of *Pseudomonas* bacteria, allowing researchers to direct efforts toward deeply investigating their functions and to discover new bacterial functionalities that currently remain hidden (27, 63–65). Those proteins with the highest odds ratio and lowest Benjamini-Hochberg's (B-H) adjusted *P* value found on each of the tables presented in File S3 should be considered strongly related with the *Pseudomonas* lifestyle in each niche. Hence, to further use our data, the main findings might be prospected carefully based on the interest of the research. For instance, we suggest using the representative sequences for the *Pseudomonas*-pangenome proteins (https://doi.org/10.5281/zenodo.7105218) to compare (i.e., blastp searches) our data with particular proteins of interest and thereby translate the ecological importance found for the query protein (if any).

Despite the large genome data set, which provides us high statistical confidences, there are some issues that may obscure the results, which are as follows: (i) it is possible that the genome metadata obtained from databases were not sufficiently detailed or even incorrect; (ii) the isolation of bacterial strains in a niche does not always imply a strict adaptation of that strain into the niche, since it may be a transient cell in the environment or even a contaminant; and (iii) biases inherent to *in silico* methods (i.e., difficulty to cluster proteins/genes with similar functions properly) may draw some wrong conclusions. Also, some of the results may be just the consequence of phylogenetic signals that may bias the presence/absence of genes/proteins in some groups of genomes (i.e., the category of human isolates is comprised mainly by *P. aeruginosa*). Thus, for better confidence to use our results, they should be validated experimentally (e.g., genetic modifications and transcriptional information).

In summary, we show the genomic adaptability patterns of *Pseudomonas* strains to different lifestyles. For example, plant-associated *Pseudomonas* strains dedicate the largest number of genes to the metabolism of carbohydrates, but the involved proteins are likely located in the cytosol, in contrast to other strains that present a higher proportion of CAZys in the outer envelope or excrete them. Additionally, contrary to plants, the human/animal environment seems to add pressure to resist stresses, although this issue may be biased toward the better understanding of clinically relevant genes. Furthermore, the association of *Pseudomonas* with higher hosts increases the probability of gene exchange through horizontal transfer. Overall, our results will facilitate studies focusing on the evolutionary dynamics, ecology, biotechnology, and clinical relevance of bacteria. New insights into genes or functions associated with isolation niches can inspire scientific applications in infectious disease diagnosis and treatment or even the development of engineered strains with biotechnological uses.

MATERIALS AND METHODS

Creation of a collection of genomes with a known ecological niche of isolation. We downloaded a total of 11,167 genomes of the bacterial genus *Pseudomonas* with their metadata from the JGI-IMG database (https://img.jgi.doe.gov/), *Pseudomonas* Genome DB (https://www.pseudomonas.com), EzBioCloud Genome Database (https://help.ezbiocloud.net/ezbiocloud-genome-database/), and NCBI genome database (https://www.ncbi.nlm.nih.gov/genome/). We manually inspected the list of genomes to remove all those that did not have concise information about the isolation source or those that, according to the name of the strain,

were redundant between genomes from different databases. The quality of the genomes was evaluated with QUAST (v5.2.0) (66) and BUSCO (v5.4.3) (67). Genomes with less than 95% completeness and/or that were highly fragmented were removed for this study. We built a phylogenetic tree with all remaining genomes with the UBCG program (68), which extracts, concatenates, and aligns 92 housekeeping genes from each genome and then builds the tree. UBCG was set to use codon-based alignment with FastTree (maximum likelihood; GTR + CAT model) and a gene support index threshold of 95%. The cutoff for gap-containing positions was set at 50%. The phylogeny was visualized with the iTOL program (69), where genomes that were placed out of the phylogenetic tree (see picture in extended methods described online at https://github.com/zakisaati/Pseudomonas_pangenome) were identified and removed from the analysis. Finally, we created a new phylogenetic tree with the remaining 3,274 genomes, and we depicted the isolation metadata in iTOL.

Pangenome analysis. We annotated the 3,274 genomes with Prokka (v1.14.6) (70). Once the genomes were annotated, the files in gene feature format (GFF) were used to perform the pangenome calculations and comparative genomics analyses. To do these analyses, we ran PPanGGOLiN (71) (v1.1.96) following the instructions of the developers and using the default values at each step of the process except for the MMseqs2 (72) clustering of proteins, for which an identity percentage of 70% was chosen. PPanGGOLiN uses a sophisticated method that defines partition nodes to build pangenome graphs used to classify gene families into persistent (conserved in the majority of genomes), shell (present at intermediate frequencies in the genome collection), and cloud (present at low frequency) (see the methods described by Gautreau et al. [71]). Scripts from this program were used to generate a matrix of the presence/absence of protein families. We obtained pangenome statistics from this program. Then, we used the roary_plots.py script (https://sanger-pathogens.github.io/Roary/) to generate graphs of the pangenomes, using the phylogenetic tree built with the UBCG program as the basis.

Search for protein functions. We annotated the proteomes with the dbCAN2 program (73) (standalone v2.0.11; https://github.com/linnabrown/run_dbcan) to search for enzymes related to carbohydrate metabolism (CAZys). We retained only the CAZys detected by at least 2 of the 3 algorithms (HMMER, DIAMOND, and HotPep) used by the program. Peptidases were detected through a DIAMOND (74) search (E value threshold, 10^{-3}) against the MEROPS database (75). The search for proteins related to resistance to antimicrobials, biocides, and other abiotic stresses was carried out by annotating the representative sequences of the groups of orthologous proteins (obtained with PPanGGOLiN) with the AMRFinderPlus (76) program (v3.9.8). The signal peptide search was performed via SignalP (v5.0b) (77). COG terms were retrieved from Prokka annotations.

Gene dynamics. Possible HGT events were estimated using the HGTector program (v2.9b3) (78), with the following flags: method = diamond, E value = 1e-10, and tax-unirank = species. The protein database was compiled from all protein sequences of NCBI RefSeq genomes of bacteria, archaea, viruses, fungi, and protozoa (1 genome per species) plus all NCBI-defined reference, representative, and type material genomes.

We looked for RGPs with the panRGP (79) algorithm implemented within the PPanGGOLiN program (71). These regions correspond to genomic islands, plasmids, and regions that are missing in multiple strains.

Statistical analyses and measurement of protein enrichment among different isolation categories. We used the Scoary program (v1.6.16) (80) to study the association of genes, proteins, or functions with the isolation source. To do so, a presence/absence matrix of genes/proteins/functions in each genome and a comma-delimited table (.csv) encoded in binary code (0 and 1) were used as input so that each genome was assigned with corresponding metadata. Statistical calculations were performed using a *P* value adjusted with Benjamini-Hochberg's (B-H) method for the correction of multiple comparisons (81). The results tables were investigated to select only those data with *P* values (BH correction) lower than the chosen threshold, which was set to 10^{-2} for comparisons of CAZYs and resistance-related genes and 10^{-6} for the protein and COG function analyses. Genes with functions with odds ratios higher than 1 and *P* values less than the selected threshold were considered to be associated with the isolation source.

We searched for significant variation in the number of different functions of the genomes of *Pseudomonas* associated with distinct niches by using Tukey's test for multiple comparisons (Tukey HSD) in the analysis of variance (ANOVA) framework with the "*stats*" R package (v4.0.2). The differences were visualized with the ggplot2 (v3.3.2) (82) package for R and the UpSetR package (v1.4.0) (83).

We built a sequence similarity network (SSN) comprising proteins significantly associated with HA or NHA categories with the Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST) (84). Then, we visualized this SSN in Cytoscape (v3.7.2) (85).

Taxonomic analysis of the genome collection. We used the GTDB-Tk program (v2.1.1) (86) to classify each genome into a *Pseudomonas* species though the "classify_wf" command. This workflow uses the closest ANI value to locate the user strain into the closest species in the GTDB.

We also compared the pairwise similarity of the genomes (all versus all) by measuring the ANI distances with FastANI (v1.33) (87) and adding the "–matrix" flag.

Data availability. There are files hosted at Zenodo (https://doi.org/10.5281/zenodo.7105218). This repository includes the following: (i) the pangenome rarefaction curve in html format, (ii) the representative sequences of the protein clusters of the *Pseudomonas* pangenome, (iii) a folder with the 3,274 genomes of our study, and (iv) two large files which are the output from executing FastANI on the genome collection.

We also included bioinformatic codes and source data in the GitHub repository created for this article (https://github.com/zakisaati/Pseudomonas_pangenome).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, XLSX file, 1.2 MB.

SUPPLEMENTAL FILE 2, PDF file, 2.5 MB.

SUPPLEMENTAL FILE 3, XLSX file, 12.1 MB.
SUPPLEMENTAL FILE 4, XLSX file, 0.5 MB.
SUPPLEMENTAL FILE 5, XLSX file, 0.4 MB.
SUPPLEMENTAL FILE 6, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

We thank Iñaki Odriozola for help in understanding statistical issues and José Francisco Cobo Díaz for allowing us to use his computing resources.

This work was funded by the Regional Government of Castilla y León, Escalera de Excelencia CLU-2018-04, and cofunded by the P.O. FEDER of Castilla y León 2014 to 2020. Z.S.S. received a grant from the Regional Government of Castilla y León and a grant cofinanced by the European NextGenerationEU, Spanish "Plan de Recuperación, Transformación y Resiliencia," Spanish Ministry of Universities, and the University of Salamanca ("Ayudas para la recualificación del sistema universitario español 2021-2022").

We declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

REFERENCES

1. Shalev O, Ashkenazy H, Neumann M, Weigel D. 2022. Commensal *Pseudomonas* protect *Arabidopsis* thaliana from a coexisting pathogen via multiple lineage-dependent mechanisms. ISME J 16:1235–1244. https://doi.org/10.1038/s41396-021-01168-6.
2. Rieusset L, Rey M, Muller D, Vacheron J, Gerin F, Dubost A, Comte G, Prigent-Combaret C. 2020. Secondary metabolites from plant-associated *Pseudomonas* are overproduced in biofilm. Microb Biotechnol 13:1562–1580. https://doi.org/10.1111/1751-7915.13598.
3. Jiménez-Gómez A, Saati-Santamaría Z, Kostovcik M, Rivas R, Velázquez E, Mateos PF, Menéndez E, García-Fraile P. 2020. Selection of the root endophyte *Pseudomonas brassicacearum* CDVBN10 as plant growth promoter for Brassica napus L. crops. Agronomy 10:1788. https://doi.org/10.3390/agronomy10111788.
4. Zhou W, Qi D, Swaisgood RR, Wang L, Jin Y, Wu Q, Wei F, Nie Y. 2021. Symbiotic bacteria mediate volatile chemical signal synthesis in a large solitary mammal species. ISME J 15:2070–2080. https://doi.org/10.1038/s41396-021-00905-1.
5. Lauer A, Simon MA, Banning JL, Lam BA, Harris RN. 2008. Diversity of cutaneous bacteria with antifungal activity isolated from female four-toed salamanders. ISME J 2:145–157. https://doi.org/10.1038/ismej.2007.110.
6. Brockmann M, Aupperle-Lellbach H, Gentil M, Heusinger A, Müller E, Marschang RE, Pees M. 2020. Challenges in microbiological identification of aerobic bacteria isolated from the skin of reptiles. PLoS One 15:e0240085. https://doi.org/10.1371/journal.pone.0240085.
7. Vodovar N, Vallenet D, Cruveiller S, Rouy Z, Barbe V, Acosta C, Cattolico L, Jubin C, Lajus A, Segurens B, Vacherie B, Wincker P, Weissenbach J, Lemaitre B, Médigue C, Boccard F. 2006. Complete genome sequence of the entomopathogenic and metabolically versatile soil bacterium *Pseudomonas entomophila*. Nat Biotechnol 24:673–679. https://doi.org/10.1038/nbt1212.
8. Ceja-Navarro JA, Vega FE, Karaoz U, Hao Z, Jenkins S, Lim HC, Kosina P, Infante F, Northen TR, Brodie EL. 2015. Gut microbiota mediate caffeine detoxification in the primary insect pest of coffee. Nat Commun 6:7618. https://doi.org/10.1038/ncomms8618.
9. Saati-Santamaría Z, Rivas R, Kolařik M, García-Fraile P. 2021. A new perspective of *Pseudomonas*—host interactions: distribution and potential ecological functions of the genus *Pseudomonas* within the bark beetle holobiont. Biology 10:164. https://doi.org/10.3390/biology10020164.
10. Saati-Santamaría Z, López-Mondéjar R, Jiménez-Gómez A, Díez-Méndez A, Větrovský T, Igual JM, Velázquez E, Kolarik M, Rivas R, García-Fraile P. 2018. Discovery of phloeophagus beetles as a source of *Pseudomonas* strains that produce potentially new bioactive substances and description of *Pseudomonas bohemica* sp. nov. Front Microbiol 9:913. https://doi.org/10.3389/fmicb.2018.00913.
11. Zimmermann J, Obeng N, Yang W, Pees B, Petersen C, Waschina S, Kissoyan KA, Aidley J, Hoeppner MP, Bunk B, Spröer C, Leippe M, Dierking K, Kaleta C, Schulenburg H. 2020. The functional repertoire contained within the native microbiota of the model nematode *Caenorhabditis elegans*. ISME J 14:26–38. https://doi.org/10.1038/s41396-019-0504-y.
12. Boxberger M, Cenizo V, Cassir N, La Scola B. 2021. Challenges in exploring and manipulating the human skin microbiome. Microbiome 9:125. https://doi.org/10.1186/s40168-021-01062-5.
13. Butaitė E, Baumgartner M, Wyder S, Kümmerli R. 2017. Siderophore cheating and cheating resistance shape competition for iron in soil and freshwater *Pseudomonas* communities. Nat Commun 8:414. https://doi.org/10.1038/s41467-017-00509-4.
14. Mulet M, Montaner M, Román D, Gomila M, Kittinger C, Zarfel G, Lalucat J, García-Valdés E. 2020. *Pseudomonas* species diversity along the Danube River assessed by rpoD gene sequence and MALDI-TOF MS analyses of cultivated strains. Front Microbiol 11:2114. https://doi.org/10.3389/fmicb.2020.02114.
15. Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, Maritz JM, Reeves D, Gandara J, Chhangawala S, Ahsanuddin S, Simmons A, Nessel T, Sundaresh B, Pereira E, Jorgensen E, Kolokotronis S-O, Kirchberger N, Garcia I, Gandara D, Dhanraj S, Nawrin T, Saletore Y, Alexander N, Vijay P, Hénaff EM, Zumbo P, Walsh M, O'Mullan GD, Tighe S, Dudley JT, Dunaif A, Ennis S, O'Halloran E, Magalhaes TR, Boone B, Jones AL, Muth TR, Paolantonio KS, Alter E, Schadt EE, Garbarino J, Prill RJ, Carlton JM, Levy S, Mason CE. 2015. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. Cell Syst 1:72–87. https://doi.org/10.1016/j.cels.2015.01.001.
16. Peix A, Ramírez-Bahena MH, Velázquez E. 2018. The current status on the taxonomy of *Pseudomonas* revisited: an update. Infect Genet Evol 57:106–116. https://doi.org/10.1016/j.meegid.2017.10.026.
17. Guo H, Chen C, Lee DJ, Wang A, Ren N. 2015. Denitrifying sulfide removal by *Pseudomonas* sp. C27 at excess carbon supply: mechanisms. Bioresour Technol 180:381–385. https://doi.org/10.1016/j.biortech.2015.01.030.
18. Zhan Y, Yan Y, Deng Z, Chen M, Lu W, Lu C, Shang L, Yang Z, Zhang W, Wang W, Li Y, Ke Q, Lu J, Xu Y, Zhang L, Xie Z, Cheng Q, Elmerich C, Lin M. 2016. The novel regulatory ncRNA, *NfiS*, optimizes nitrogen fixation via base pairing with the nitrogenase gene *nifK* mRNA in *Pseudomonas stutzeri* A1501. Proc Natl Acad Sci U S A 113:E4348–E4356. https://doi.org/10.1073/pnas.1604514113.
19. Nogales J, Mueller J, Gudmundsson S, Canalejo FJ, Duque E, Monk J, Feist AM, Ramos JL, Niu W, Palsson BO. 2020. High-quality genome-scale metabolic modelling of *Pseudomonas putida* highlights its broad metabolic capabilities. Environ Microbiol 22:255–269. https://doi.org/10.1111/1462-2920.14843.
20. Nguyen DD, Melnik AV, Koyama N, Lu X, Schorn M, Fang J, Aguinaldo K, Lincecum TL, Ghequire MGK, Carrion VJ, Cheng TL, Duggan BM, Malone JG, Mauchline TH, Sanchez LM, Kilpatrick AM, Raaijmakers JM, Mot RD, Moore BS, Medema MH, Dorrestein PC. 2016. Indexing the *Pseudomonas* specialized metabolome enabled the discovery of poaeamide B and the bananamides. Nat Microbiol 2:16197. https://doi.org/10.1038/nmicrobiol.2016.197.
21. Saati-Santamaría Z, Selem-Mojica N, Peral-Aranega E, Rivas R, García-Fraile P. 2022. Unveiling the genomic potential of Pseudomonas type strains for discovering new natural products. Microb Genom 8:e000758. https://doi.org/10.1099/mgen.0.000758.

22. Vogel CM, Potthoff DB, Schäfer M, Barandun N, Vorhol JA. 2021. Protective role of the *Arabidopsis* leaf microbiota against a bacterial pathogen. Nat Microbiol 6:1537–1548. https://doi.org/10.1038/s41564-021-00997-7.
23. Sands K, Carvalho MJ, Portal E, Thomson K, Dyer C, Akpulu C, Andrews R, Ferreira A, Gillespie D, Hender T, Hood K, Mathias J, Milton R, Nieto M, Taiyari K, Chan GJ, Bekele D, Solomon S, Basu S, Chattopadhyay P, Mukherjee S, Iregbu K, Modibbo F, Uwaezuoke S, Zahra R, Shirazi H, Muhammad A, Mazarati JB, Rucogoza A, Gaju L, Mehtar S, Bulabula ANH, Whitelaw A, Walsh TR, BARNARDS Group. 2021. Characterization of antimicrobial-resistant Gram-negative bacteria that cause neonatal sepsis in seven low-and middle-income countries. Nat Microbiol 6:512–523. https://doi.org/10.1038/s41564-021-00870-7.
24. Peña JM, Prezioso SM, McFarland KA, Kambara TK, Ramsey KM, Deighan P, Dove SL. 2021. Control of a programmed cell death pathway in *Pseudomonas aeruginosa* by an antiterminator. Nat Commun 12:1702. https://doi.org/10.1038/s41467-021-21941-7.
25. Mahrt N, Tietze A, Künzel S, Franzenburg S, Barbosa C, Jansen G, Schulenburg H. 2021. Bottleneck size and selection level reproducibly impact evolution of antibiotic resistance. Nat Ecol Evol 5:1233–1242. https://doi.org/10.1038/s41559-021-01511-2.
26. Westermann AJ, Vogel J. 2021. Cross-species RNA-seq for deciphering host–microbe interactions. Nat Rev Genet 22:361–378. https://doi.org/10.1038/s41576-021-00326-y.
27. Levy A, Conway JM, Dangl JL, Woyke T. 2018. Elucidating bacterial gene functions in the plant microbiome. Cell Host Microbe 24:475–485. https://doi.org/10.1016/j.chom.2018.09.005.
28. Sargison FA, Fitzgerald JR. 2021. Advances in transposon mutagenesis of *Staphylococcus aureus*: insights into pathogenesis and antimicrobial resistance. Trends Microbiol 29:282–285. https://doi.org/10.1016/j.tim.2020.11.003.
29. Tong Y, Weber T, Lee SY. 2019. CRISPR/Cas-based genome engineering in natural product discovery. Nat Prod Rep 36:1262–1280. https://doi.org/10.1039/c8np00089a.
30. Dewar AE, Thomas JL, Scott TW, Wild G, Griffin AS, West SA, Ghoul M. 2021. Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range. Nat Ecol Evol 5:1624–1636. https://doi.org/10.1038/s41559-021-01573-2.
31. Taib N, Megrian D, Witwinowski J, Adam P, Poppleton D, Borrel G, Beloin C, Gribaldo S. 2020. Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. Nat Ecol Evol 4:1661–1672. https://doi.org/10.1038/s41559-020-01299-7.
32. Chen MY, Teng WK, Zhao L, Hu CX, Zhou YK, Han BP, Song LR, Shu WS. 2021. Comparative genomics reveals insights into cyanobacterial evolution and habitat adaptation. ISME J 15:211–227. https://doi.org/10.1038/s41396-020-00775-z.
33. Sheppard SK, Guttman DS, Fitzgerald JR. 2018. Population genomics of bacterial host adaptation. Nat Rev Genet 19:549–565. https://doi.org/10.1038/s41576-018-0032-z.
34. Chen W, Roslund K, Fogarty CL, Pussinen PJ, Halonen L, Groop PH, Metsälä M, Lehto M. 2016. Detection of hydrogen cyanide from oral anaerobes by cavity ring down spectroscopy. Sci Rep 6:22577. https://doi.org/10.1038/srep22577.
35. Hassan HA, Aly AA. 2018. Isolation and characterization of three novel catechol 2, 3-dioxygenase from three novel haloalkaliphilic BTEX-degrading *Pseudomonas* strains. Int J Biol Macromol 106:1107–1114. https://doi.org/10.1016/j.ijbiomac.2017.08.113.
36. Ravanbakhsh M, Kowalchuk GA, Jousset A. 2019. Root-associated microorganisms reprogram plant life history along the growth–stress resistance tradeoff. ISME J 13:3093–3101. https://doi.org/10.1038/s41396-019-0501-1.
37. Jing TZ, Qi FH, Wang ZY. 2020. Most dominant roles of insect gut bacteria: digestion, detoxification, or essential nutrient provision? Microbiome 8:38. https://doi.org/10.1186/s40168-020-00823-y.
38. Mao M, Bennett GM. 2020. Symbiont replacements reset the co-evolutionary relationship between insects and their heritable bacteria. ISME J 14:1384–1395. https://doi.org/10.1038/s41396-020-0616-4.
39. Jiang X, Hall AB, Arthur TD, Plichta DR, Covington CT, Poyet M, Crothers J, Moses PL, Tolonen AC, Vlamakis H, Alm EJ, Xavier RJ. 2019. Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. Science 363:181–187. https://doi.org/10.1126/science.aau5238.
40. Henry LP, Bruijning M, Forsberg SK, Ayroles JF. 2021. The microbiome extends host evolutionary potential. Nat Commun 12:5141. https://doi.org/10.1038/s41467-021-25315-x.
41. Kappler A, Bryce C, Mansor M, Lueder U, Byrne JM, Swanner ED. 2021. An evolving view on biogeochemical cycling of iron. Nat Rev Microbiol 19:360–374. https://doi.org/10.1038/s41579-020-00502-7.
42. Shu WS, Huang LN. 2022. Microbial diversity in extreme environments. Nat Rev Microbiol 20:219–235. https://doi.org/10.1038/s41579-021-00648-y.
43. Wunder LC, Aromokeye DA, Yin X, Richter-Heitmann T, Willis-Poratti G, Schnakenberg A, Otersen C, Dohrmann I, Römer M, Bohrmann G, Kasten S, Friedrich MW. 2021. Iron and sulfate reduction structure microbial communities in (sub-) Antarctic sediments. ISME J 15:3587–3604. https://doi.org/10.1038/s41396-021-01014-9.
44. Kordes A, Preusse M, Willger SD, Braubach P, Jonigk D, Haverich A, Warnecke G, Häussler S. 2019. Genetically diverse *Pseudomonas aeruginosa* populations display similar transcriptomic profiles in a cystic fibrosis explanted lung. Nat Commun 10:3397. https://doi.org/10.1038/s41467-019-11414-3.
45. Tremblay J, Fortin N, Elias M, Wasserscheid J, King TL, Lee K, Greer CW. 2019. Metagenomic and metatranscriptomic responses of natural oil degrading bacteria in the presence of dispersants. Environ Microbiol 21:2307–2319. https://doi.org/10.1111/1462-2920.14609.
46. Fisher RM, Henry LM, Cornwallis CK, Kiers ET, West SA. 2017. The evolution of host-symbiont dependence. Nat Commun 8:15973. https://doi.org/10.1038/ncomms15973.
47. Louca S, Mazel F, Doebeli M, Parfrey LW. 2019. A census-based estimate of Earth's bacterial and archaeal diversity. PLoS Biol 17:e3000106. https://doi.org/10.1371/journal.pbio.3000106.
48. Louca S, Parfrey LW, Doebeli M. 2016. Decoupling function and taxonomy in the global ocean microbiome. Science 353:1272–1277. https://doi.org/10.1126/science.aaf4507.
49. Bradley PH, Pollard KS. 2017. Proteobacteria explain significant functional variability in the human gut microbiome. Microbiome 5:36. https://doi.org/10.1186/s40168-017-0244-z.
50. Degli Esposti M, Martinez Romero E. 2017. The functional microbiome of arthropods. PLoS One 12:e0176573. https://doi.org/10.1371/journal.pone.0176573.
51. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, Bengtsson-Palme J, Anslan S, Coelho LP, Harend H, Huerta-Cepas J, Medema MH, Maltz MR, Mundra S, Olsson PA, Pent M, Põlme S, Sunagawa S, Ryberg M, Tedersoo L, Bork P. 2018. Structure and function of the global topsoil microbiome. Nature 560:233–237. https://doi.org/10.1038/s41586-018-0386-6.
52. Tal S, Tikhonov E, Aroch I, Hefetz L, Turjeman S, Koren O, Kuzi S. 2021. Developmental intestinal microbiome alterations in canine fading puppy syndrome: a prospective observational study. NPJ Biofilms Microbiomes 7:52. https://doi.org/10.1038/s41522-021-00222-7.
53. Freschi L, Vincent AT, Jeukens J, Emond-Rheault JG, Kukavica-Ibrulj I, Dupont MJ, Charette SJ, Boyle B, Levesque RG. 2019. The *Pseudomonas aeruginosa* pan-genome provides new insights on its population structure, horizontal gene transfer, and pathogenicity. Genome Biol Evol 11:109–120. https://doi.org/10.1093/gbe/evy259.
54. Zhalnina K, Louie KB, Hao Z, Mansoori N, da Rocha UN, Shi S, Cho H, Karaoz U, Loqué D, Bowen BP, Firestone MK, Northen TR, Brodie EL. 2018. Dynamic root exudate chemistry and microbial substrate preferences drive patterns in rhizosphere microbial community assembly. Nat Microbiol 3:470–480. https://doi.org/10.1038/s41564-018-0129-3.
55. Fabryová A, Kostovčík M, Díez-Méndez A, Jiménez-Gómez A, Celador-Lera L, Saati-Santamaría Z, Sechovcová H, Menéndez E, Kolařik M, García-Fraile P. 2018. On the bright side of a forest pest-the metabolic potential of bark beetles' bacterial associates. Sci Total Environ 619–620:9–17. https://doi.org/10.1016/j.scitotenv.2017.11.074.
56. Levy A, Salas Gonzalez I, Mittelviefhaus M, Clingenpeel S, Herrera Paredes S, Miao J, Wang K, Devescovi G, Stillman K, Monteiro F, Alvarez BR, Lundberg DS, Lu TY, Lebeis S, Jin Z, McDonald M, Klein AP, Feltcher ME, Rio TF, Grant SR, Doty SL, Ley RE, Zhao B, Venturi V, Pelletier DA, Vorholt JA, Tringe SG, Woyke T, Dang JL. 2017. Genomic features of bacterial adaptation to plants. Nat Genet 50:138–150. https://doi.org/10.1038/s41588-017-0012-9.
57. Nguyen TT, Myrold DD, Mueller RS. 2019. Distributions of extracellular peptidases across prokaryotic genomes reflect phylogeny and habitat. Front Microbiol 10:413. https://doi.org/10.3389/fmicb.2019.00413.
58. Shah VS, Meyerholz DK, Tang XX, Reznikov L, Abou Alaiwa M, Ernst SE, Karp PH, Wohlford-Lenane CL, Heilmann KP, Leidinger MR, Allen PD, Zabner J, McCray PB, Ostedgaard LS, Stoltz DA, Randak CO, Welsh MJ. 2016. Airway acidification initiates host defense abnormalities in cystic fibrosis mice. Science 351:503–507. https://doi.org/10.1126/science.aad5589.
59. Bhagirath AY, Li Y, Somayajula D, Dadashi M, Badr S, Duan K. 2016. Cystic fibrosis lung environment and *Pseudomonas aeruginosa* infection. BMC Pulm Med 16:174. https://doi.org/10.1186/s12890-016-0339-5.
60. Damron FH, Oglesby-Sherrouse AG, Wilks A, Barbier M. 2016. Dual-seq transcriptomics reveals the battle for iron during *Pseudomonas aeruginosa* acute murine pneumonia. Sci Rep 6:39172. https://doi.org/10.1038/srep39172.
61. Labarre A, López-Escardó D, Latorre F, Leonard G, Bucchini F, Obiol A, Cruaud C, Sieracki ME, Jaillon O, Wincker P, Vandepoele K, Logares R, Massana R. 2021.

Comparative genomics reveals new functional insights in uncultured MAST species. ISME J 15:1767–1781. https://doi.org/10.1038/s41396-020-00885-8.
62. Mageiros L, Méric G, Bayliss SC, Pensar J, Pascoe B, Mourkas E, Calland JK, Yahara K, Murray S, Wilkinson TS, Williams LK, Hitchings MD, Porter J, Kemmett K, Feil EJ, Jolley KA, Williams NJ, Corander J, Sheppard SK. 2021. Genome evolution and the emergence of pathogenicity in avian *Escherichia coli*. Nat Commun 12:765. https://doi.org/10.1038/s41467-021-20988-w.
63. Kobras CM, Fenton AK, Sheppard SK. 2021. Next-generation microbiology: from comparative genomics to gene function. Genome Biol 22:123. https://doi.org/10.1186/s13059-021-02344-9.
64. Brunetti AE, Bunk B, Lyra ML, Fuzo CA, Marani MM, Spröer C, Haddad CFB, Lopes NP, Overmann J. 2022. Molecular basis of a bacterial-amphibian symbiosis revealed by comparative genomics, modeling, and functional testing. ISME J 16:788–800. https://doi.org/10.1038/s41396-021-01121-7.
65. Allen JP, Ozer EA, Minasov G, Shuvalova L, Kiryukhina O, Anderson WF, Satchell KJF, Hauser AR. 2020. A comparative genomics approach identifies contact-dependent growth inhibition as a virulence determinant. Proc Natl Acad Sci U S A 117:6811–6821. https://doi.org/10.1073/pnas.1919198117.
66. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.
67. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
68. Na SI, Kim YO, Yoon SH, Ha SM, Baek I, Chun J. 2018. UBCG: up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. J Microbiol 56:280–285. https://doi.org/10.1007/s12275-018-8014-6.
69. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239.
70. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
71. Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, Dubois M, Perrin A, Médigue C, Calteau A, Cruveiller S, Matias C, Ambroise C, Rocha EPC, Vallenet D. 2020. PPanGGOLiN: depicting microbial diversity via a partitioned pangenome graph. PLoS Comput Biol 16:e1007732. https://doi.org/10.1371/journal.pcbi.1007732.
72. Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026–1028. https://doi.org/10.1038/nbt.3988.
73. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. 2018. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 46:W95–W101. https://doi.org/10.1093/nar/gky418.
74. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176.
75. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. 2018. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res 46:D624–D632. https://doi.org/10.1093/nar/gkx1134.
76. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, Hoffmann M, Pettengill JB, Prasad AB, Tillman GE, Tyson GH, Klimke W. 2021. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep 11:12728. https://doi.org/10.1038/s41598-021-91456-0.
77. Armenteros JJA, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 2019 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37:420–423. https://doi.org/10.1038/s41587-019-0036-z.
78. Zhu Q, Kosoy M, Dittmar K. 2014. HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. BMC Genom 15:717. https://doi.org/10.1186/1471-2164-15-717.
79. Bazin A, Gautreau G, Médigue C, Vallenet D, Calteau A. 2020. panRGP: a pangenome-based method to predict genomic islands and explore their diversity. Bioinformatics 36:i651–i658. https://doi.org/10.1093/bioinformatics/btaa792.
80. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. 2016. Rapid scoring of genes in microbial pangenome-wide association studies with Scoary. Genome Biol 17:238. https://doi.org/10.1186/s13059-016-1108-8.
81. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
82. Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York, NY.
83. Conway JR, Lex A, Gehlenborg N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics 33:2938–2940. https://doi.org/10.1093/bioinformatics/btx364.
84. Zallot R, Oberg N, Gerlt JA. 2019. The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. Biochemistry 58:4169–4182. https://doi.org/10.1021/acs.biochem.9b00735.
85. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303.
86. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. 2020. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36:1925–1927. https://doi.org/10.1093/bioinformatics/btz848.
87. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9.