Lemos *et al. BMC Microbiol* (2021) 21:294
https://doi.org/10.1186/s12866-021-02354-4

BMC Microbiology

RESEARCH

Open Access

Genome-centric metagenomics reveals insights into the evolution and metabolism of a new free-living group in Rhizobiales

Leandro Nascimento Lemos[1], Fabíola Marques de Carvalho[1], Alexandra Gerber[1], Ana Paula C. Guimarães[1], Celio Roberto Jonck[2], Luciane Prioli Ciapina[1] and Ana Tereza Ribeiro de Vasconcelos[1*]

Abstract

Background: The Rhizobiales (Proteobacteria) order is an abundant and diverse group of microorganisms, being extensively studied for its lifestyle based on the association with plants, animals, and humans. New studies have demonstrated that the last common ancestor (LCA) of Rhizobiales had a free-living lifestyle, but the phylogenetic and metabolism characterization of basal lineages remains unclear. Here, we used a high-resolution phylogenomic approach to test the monophyly of the Aestuariivirgaceae family, a new taxonomic group of Rhizobiales. Furthermore, a deep metabolic investigation provided an overview of the main functional traits that can be associated with its lifestyle. We hypothesized that the presence of pathways (e.g., Glycolysis/Gluconeogenesis) and the absence of pathogenic genes would be associated with a free-living lifestyle in Aestuariivirgaceae.

Results: Using high-resolution phylogenomics approaches, our results revealed a clear separation of Aestuariivirgaceae into a distinct clade of other Rhizobiales family, suggesting a basal split early group and corroborate the monophyly of this group. A deep functional annotation indicated a metabolic versatility, which includes putative genes related to sugar degradation and aerobic respiration. Furthermore, many of these traits could reflect a basal metabolism and adaptations of Rhizobiales, as such the presence of Glycolysis/Gluconeogenesis pathway and the absence of pathogenicity genes, suggesting a free-living lifestyle in the Aestuariivirgaceae members.

Conclusions: Aestuariivirgaceae (Rhizobiales) family is a monophyletic taxon of the Rhizobiales with a free-living lifestyle and a versatile metabolism that allows these microorganisms to survive in the most diverse microbiomes, demonstrating their adaptability to living in systems with different conditions, such as extremely cold environments to tropical rivers.

Keywords: Rhizobiales, Integration of genomic public data, Aestuariivirgaceae, Evolution, Uncultivated lineages

Background

The Rhizobiales (Proteobacteria) order is abundant, diverse and widespread in several environments [1]. Due to their association with plant, animal, and human diseases, and their economic impact, many microorganisms of this group have been widely studied applying molecular biology technologies (metagenomics, ARISA/T-RFLP, geochips, 16S rRNA sequencing). In plants, Rhizobiales order includes symbionts that establish mutualistic and pathogenic relationships. *Rhizobium, Bradyrhizobium, Azorhizobium* and others genera form a symbiotic association with legumes and are responsible for the nitrogen fixation process (for a review see [2]) while *Agrobacterium* belongs to the pathogenic

*Correspondence: atrv@lncc.br
[1] Bioinformatics Laboratory, National Laboratory of Scientific Computing (LNCC), Av. Getúlio Vargas, 333 - Quitandinha, Petrópolis, RJ 25651-076, Brazil
Full list of author information is available at the end of the article

© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

group [3]. Members of the Rhizobiales order have been found in association with lichens [4], as a third member of this ecological relationship. The genera *Brucella* and *Bartonella* are associated with animal and human diseases [5]. In marine environments, Rhizobiales have been associated with diseases in corals [6], however, it has not been proven as the causative agent and could be only an opportunistic bacteria identified in diseased tissues**.** In water column microbiomes surrounding the giant kelp *Macrocystis pyrifera*, Rhizobiales abundance was associated with an increased carbon dioxide (pCO2) [7]. Ng and Chiu [8] observed that an increase in Rhizobiales may be associated with the increase of nutrients that lead to hypoxia and acidification of the oceans.

To date (August 2021), 6983 Rhizobiales genomes are available in the Genome Taxonomy Database (GTDB – [9]), which include nitrogen-fixing plant symbionts (*Rhizobium* and *Bradyrhizobium*), plant and human pathogens (Candidatus *Liberibacter* and *Brucella*) or free-living in soil (*Methylobacterium*). However, some of these genomes deposited in public repositories represent new taxonomic groups and have not been individually explored in the evolutionary and metabolic context. To complement microbiological studies and highlight new discoveries of evolution and metabolism of new taxonomic groups, the reconstruction of genomes from metagenomes samples has been applied in several microbiome datasets [10–13]. Briefly, metagenomic reads were assembled into contigs and then contigs were clustered into individual populations, where each population represents a potential microbial genome [14]. The main advantage of this approach is to access taxonomic and metabolic information of microorganism groups that lack cultivated reference genomes. This includes the description of new archaeal and bacterial lineages [12] and their roles in several microbiomes. Recent advances in assembly and binning algorithms have provided accurate and biological validations predicted in silico results of taxonomic groups discovered by reconstruction of genomes from metagenomes, which were later cultivated and validated by the use of cultivation methods [15].

New taxa have been affiliated to the order Rhizobiales, which include the Aestuariivirgaceae (Rhizobiales) family proposed by [16] during the description and whole-genome-sequence of the *Aestuariivirga litoralis* species. This group was first described as part of an investigation to understand estuarine sediments' microbiome, highlighting significant phenotypic and genomic characterization findings. Furthermore, initial phylogeny analysis based on 16S rRNA and protein marker genes showed that his group should represent a new family [16]. However, an investigation using additional genomes is necessary to corroborate the monophyly of this group, once its phylogenetic position remains unclear. Besides, a deep metabolic investigation can provide new insights into the functional traits and lifestyle of Aestuariivirgaceae in terrestrial and water environments.

In this study we used Metagenome-Assembled Genomes (MAGs) and whole-genome-sequenced bacterial isolates to test the monophyly and to describe metabolic profile of the Aestuariivirgaceae family that can be associated with its lifestyle. We hypothesized that the presence of pathways (e.g., Glycolysis/Gluconeogenesis) and the absence of pathogenic genes would be associated with a free-living lifestyle in Aestuariivirgaceae.

Results and discussion

To test the monophyly and to predict the putative central metabolism of the Aestuariivirgaceae (Rhizobiales) family, we used a dataset with 19 whole-genome sequenced bacterial isolates and Metagenome-Assembled Genomes (MAGs) (Table 1). Firstly, we reconstructed a new metagenome-assembled genome (MAG - named METAPETRO_BR_BIN_54) using marine sediment metagenomes (Supplementary Table 1). Specifically, METAPETRO_BR_BIN_54 has 93.7% of completeness and 2.17% of contamination (Table 1). According to Minimum information about a metagenome-assembled genome of bacteria and archaea (MIMAG) standards [14] and CheckM classification [17], MAGs with more than 90% of completeness and less than 5% of contamination are considered high-quality and near-complete genomes. We reinforce that 2.17 represents genomes with lower percentages of contamination. To complete these analyses, we also add 18 genomes [11, 12, 16, 18–23] deposited in public sequence repositories (Table 1), which were not explored deeply in the context of this investigation. Also according to MIMAG standards [14], these genomes were assigned with high-quality or medium-quality drafts (Table 1). We found Aestuariivirgaceae members in a broad of several environments (Table 1), such as terrestrial (soil, permeable sediments, and phosphatic stromatolites formations) and aquatic (marine sediments, artificial well, wastewater treatment plant, High Arctic freshwater, and Amazon Basin River), demonstrating their adaptability to living in systems with different conditions, such as extremely cold environments to tropical rivers.

From 19 genomes, a total of 13 unique species were identified, which includes *Aestuariivirga litoralis* described by Li and collaborators [16]. High-resolution taxonomy prediction based on the rank-normalized GTDB taxonomy with the criteria of relative evolutionary divergence (RED) and ANI indicated the presence of 8 unique species of the genera *Aestuariivirga* (*Aestuariivirga litoralis*, *Aestuariivirga sp902826365*,

Table 1 Genomic features of Aestuariivirgaceae (Rhizibiales; Proteobacteria) genomes isolated or reconstructed using metagenomes

Genome	Number of contigs	Size (Mbp)	Environment	Taxonomy (GTDB)	Respiration	Completeness/ Contamination	Genome Accession	Reference
Aestuariivirga litoralis	26	4.2	Water	g__Aestuariivirga; s__Aestuariivirga litoralis	Aerobic	98.5/0.4	GCF_003234965.1	[16]
Palsa_927	294	2.59	Palsa	g__Aestuariivirga; s__Aestuariivirga sp003151375	Aerobic	82.2/2.4	GCA_003151375.1	[12]
METAPETRO_BR_BIN_54	756	5.06	Marine sediment	g__JABDJG01;s__	Aerobic	93.7/2.17	This study	This study.
SCPDY	37	7.1	Storage Tank/Water	g__Nordella; s__Nordella sp005502925	Aerobic	98.9/0.6	GCF_005502925.1	[11]
X2C	106	7.1	Artificial well/Hydra/Water	s__Nordella sp005502925	Aerobic	100.0/0.2	GCF_005502975.1	[11].
X1A	48	7.1	Artificial well/Hydra/Water	g__Nordella; s__Nordella sp005502925	Aerobic	98.9/0.6	GCF_005502345.1	[11]
AP_21	745	3.8	Soil	g__Nordella; s__Nordella sp005884715	Aerobic	82.68/0.69	GCA_005884715.1	[12]
Bin_29_15	222	3.2	High Arctic freshwater	g__Aestuariivirga; s__Aestuariivirga sp009885825	Aerobic	95.79/1.64	GCA_009885825.1	[19]
ES-bin-180	684	2.8	Nearby exposed soil of glacier	g__Aestuariivirga; s__Aestuariivirga sp014380505	Aerobic	66.54/1.72	GCA_014380505.1	[20]
RU_4_15	370	3.0	Phosphatic stromatolites formations	g__Aestuariivirga; s__Aestuariivirga sp012032065	Aerobic	72.2/2.48	GCA_012032065.1	[21]
SS_bin_17	495	4.3	Permeable (sandy) sediments	g__JABDJG01; s__JABDJG01 sp013002595	Aerobic	89.03/1.20	GCA_013002595.1	[22]
AM_0226	58	2.9	Amazon Basin river	g__Aestuariivirga; s__Aestuariivirga sp900298995	Aerobic	98.79/0.34	GCA_900298995.1	[18]
Loclat_bin-06399	820	3.9	Water Lake	g__Aestuariivirga; s__Aestuariivirga sp903930095	Aerobic	88.97/4.26	GCA_903930095.1	[47]
Loc080925-5m_bin-0050	321	2.7	Water Lake	g__CABJBCQ01; s__CABJBCQ01 sp903951595	Aerobic	95.0/1.96	GCA_903944365.1	[47]
Loc080925–4m_bin-0358	264	2.7	Water Lake	g__CABJBCQ01; s__CABJBCQ01 sp903951595	Aerobic	99.13/3.26	GCA_903951595.1	[47]
Loc080925–4-5m_bin-0281	264	2.7	Water Lake	g__CABJBCQ01; s__CABJBCQ01 sp903951595	Aerobic	99.13/3.26	GCA_903958745.1	[47]
RBC017	344	2.9	Wastewater treatment plant	g__Aestuariivirga; s__Aestuariivirga sp902826365	Aerobic	88.08/0.6	GCA_902826365.1	[23]
RBC019	527	2.8	Wastewater treatment plant	g__Aestuariivirga; Aestuariivirga sp902826365	Aerobic	82.96/3.06	GCA_902826675.1	[23]
RBC065	393	2.7	Wastewater treatment plant	g__Aestuariivirga; Aestuariivirga sp902826365	Aerobic	79.98/1.3	GCA_902826905.1	[23]

Aestuariivirga sp003151375, *Aestuariivirga sp009885825*, *Aestuariivirga sp012032065*, *Aestuariivirga sp014380505*, *Aestuariivirga sp900298995*, and *Aestuariivirga sp903930095*). *Nordella* genus was represented by two unique species (*Nordella* sp005502925 and *Nordella* sp005884715). This species was identified for the first

time using 16S rRNA gene sequence analysis in an ecological interaction with an amoeba from a water tank [24]. We also identified genomes assigned with the genus *JABDJG01* (*JABDJG01 sp013002595* and *JABDJG01 sp.*) and *CABJBCQ01*. Both genera have not been described in previous studies and the taxonomy name reflects the proposal used by the Genome Taxonomy Database. To clarify the phylogenetic position and to test the monophyly of the Aestuariivirgaceae, we used a high-resolution phylogenomic approach based on the alignment and concatenation of single-copy marker genes (Fig. 1). Our results revealed a clear separation of Aestuariivirgaceae family into a distinct clade of other Rhizobiales families (Bootstrap $\gtrsim$ 95%), indicating that it could seem to be a basal group and may have split early. The formation of this clade validates the monophylic origin of the Aestuariivirgaceae family, which was proposed by Li and collaborators [16]. Our phylogenetic results were the same that predicted by GTDBTk to estimate the taxonomy assignment (Table 1), where *Aestuariivirga sp902826365*, *Nordella sp005502925*, and *CABJBCQ01 sp903951595* were represented by more than one genome.

We found functional traits that may be useful in the ecological niche preferences of Aestuariivirgaceae (Fig. 2). Firstly, the most abundant general functions were associated with Amino Acid Metabolism and Transport, Functions Unknown, Energy Production and Conversion and Carbohydrate metabolism and transport (Fig. 2A). A similar pattern was observed in other Alphaproteobacteria members as described by Pini and collaborators [25]. As expected, many of these functions are also essential for central and accessory metabolism of Aestuariivirgaceae (Fig. 2B). The production of pyruvate from glucose uptake via the Embden-Meyerhof-Parnas (Glycolysis) pathway appeared to be a general trait of the Aestuariivirgaceae members. In addition, we also do not discard a possibility to also use Pentose Phosphate pathway as alternative via to uptake sugars. Yang, Heath & Setubal [26] pointed out that the LCA of all Rhizobiales showed any genes associated with Glycolysis/Gluconeogenesis. In this case, Aestuariivirgaceae metabolism would reflect a basal metabolism of Rhizobiales. The presence of Embden-Meyerhof-Parnas (Glycolysis) pathway also suggest that Aestuariivirgaceae family is well adapted to survive in environments rich in organic matter, as such marine sediments, soils [10, 12], estuarine ecosystems [16] and rivers [18], where the organic matter derived from biological biomass is abundant. Furthermore, *Nordella sp005884715* (AP_21 genome) has potential to perform pyruvate fermentation to lactate generation, which would represent adaptation and alternative metabolism to survive in soils (Fig. 2B). Machine learning predictions revealed with a high-confidence (>0.7) the presence of D-glucose uptake (Fig. 2C) in ten species, corroborating our previous prediction analysing "gene-by-gene" in the metabolic reconstruction. We also infer that *Aestuariivirga litoralis* may living associated with particulate carbon in estuarine ecosystems, where organic matter degradation could continue via Embden-Meyerhof-Parnas (Glycolysis), but we also have not discarded its occurrence in a free-living water column. The same seems to be probably in the other *Aestuariivirga*, *Nordella* and *JABDJG01* and *CABJBCQ01* species described here, and reconstructed from soils, rivers, lakes and sediments, where organic matter is rich.

Still, regarding central metabolism and energy acquisition, member of the Aestuariivirgaceae family showed the main enzymes of the Electron Transport Chain and oxidative phosphorylation (Fig. 2B), including Ubiquinol-cytochrome c reductase cytochrome b/c1 (K00410) and Cytochrome c oxidase cbb3 (K00404), which are key-enzymes in the process to generate ATP using oxygen with final electron acceptor [27]. This result indicates that unlike other non-nitrogen-fixing Rhizobiales, such as Candidatus *Liberibacter asiaticus* and Candidatus *Liberibacter solanacearum* [28], the Aestuariivirgaceae genomes described here have the potential for aerobic respiration. As with glucose uptake metabolism, machine learning predictions also revealed with high-confidence (>0.7) the presence of aerobic metabolism in all Aestuariivirgaceae investigated here (Fig. 2C). Furthermore, Li and collaborators [16] already validated experimentally this metabolic function in *Aestuariivirga litoralis*. Probably, many of the functional predictions described here may reflect the ecological role of these species in their environments, but it also needs experimental validations to better highlight all these predictions. Some new taxonomic groups were firstly described using assembly/binning approaches, and then in additional studies their putative functions were validated. The main recent example is the new archaea super-phylum Asgard archaea discovered in 2015 [29], where evolutionary and functional predictions were done by sequence analyses and 5 years

(See figure on next page.)

Fig. 1 Phylogenomic tree showing the evolutionary position of the Aestuariivirgaceae (Proteobacteria, Rhizobiales) species. The phylogenomic tree was inferred using the alignment and the concatenation of bacterial single-copy core genes (SCGs) (Supplementary Table 3) [39, 40] under the Jones-Taylor-Thorton model and CAT approximation with 20 rate categories. The Aestuariivirgaceae genomes studied here are assigned with a pink color. The nodes that showed a bootstrap support $\geq$70% are assigned with a black point in the tree. Green circle indicates the outgroup used in the phylogenomic analysis

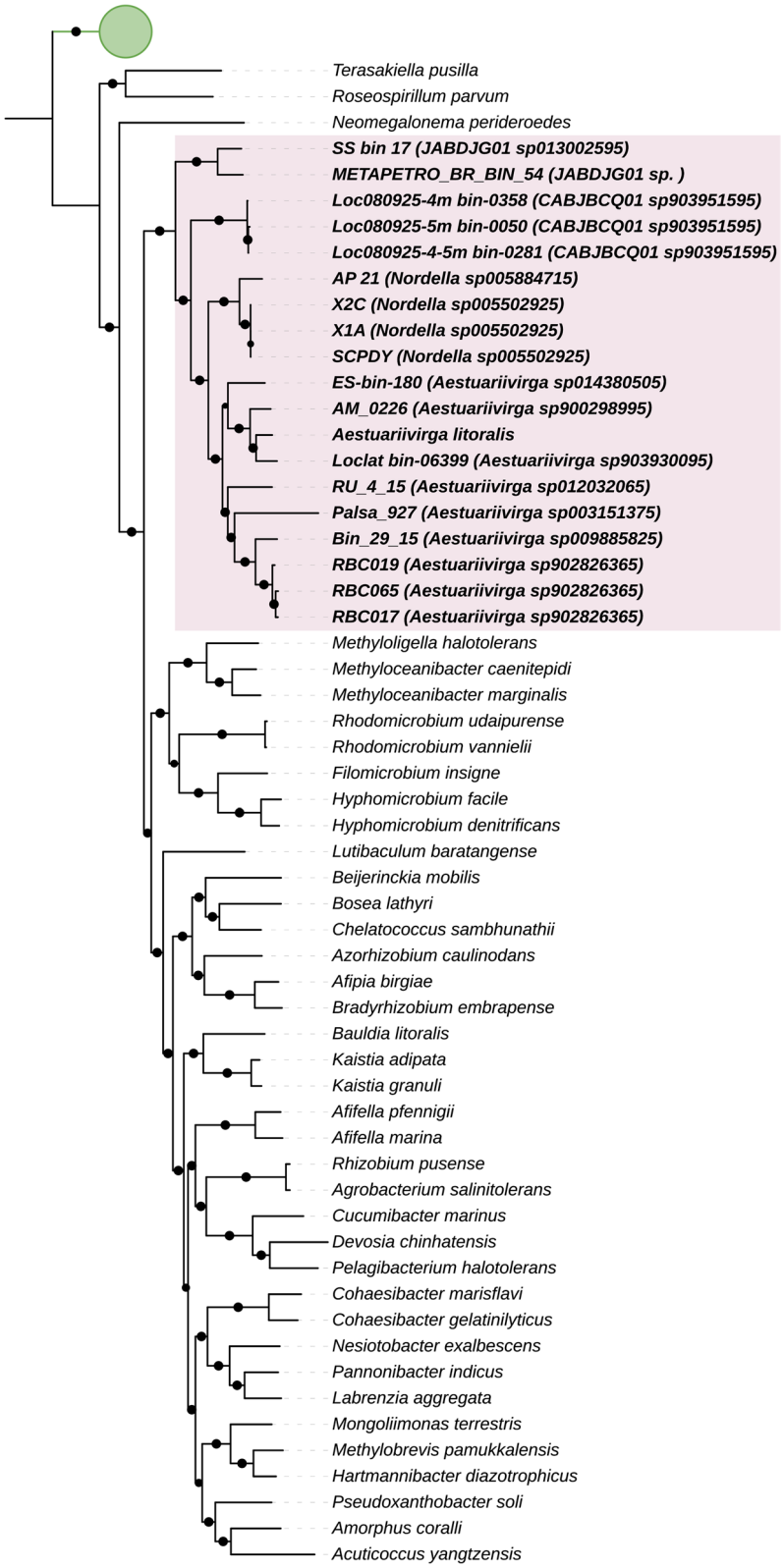

Fig. 1 (See legend on previous page.)

later the first Asgard archaea Candidatus *Prometheoarchaeum syntrophicum* was cultivated [15].

Alternative metabolism to obtain energy could be present in Aestuariivirgaceae (Fig. 2B), but their presence is limited by homology unclear (I) or fragmented metabolic pathway predictions (II). The first case (I), which was related with homology unclear, was the presence of Alkane 1-monooxygenase (alkB - K00496) in *JABDJG01 spp.*(METAPETRO-BIN-54) and *Nordella sp005502925* (X2C, X1A and SCPDY) species. Both sequences showed a sequence identity of 45 and 40% respectively, and the presence of Alkane 1-monooxygenase (alkB - K00496) in both genomes could indicate a potential to use alkanes as growth substrates [30]. The presence of alkanes was not quantified in our sediment samples (METAPETRO-BIN-54) and also was not reported in the previous studies where the *Nordella sp005502925* (X2C, X1A and SCPDY) species genome were reconstructed [11]. Alternatively, regarding fragmented metabolic pathways (II), we also speculate that some Aestuariivirgaceae species could use a final electron acceptor derived from the nitrogen and sulfur cycles. We found an incomplete set of nitrogen cycle genes (por example, nitrite reductase/K00368/Denitrification and nitrate reductase/K00371/Nitrification), suggesting its potential to use nitrogen in respiration. In both cases shown here, we stressed that further studies are needed to investigate whether these functions are really active or only represent distant homologous genes or fragmented metabolic pathways.

Members of Aestuariivirgaceae showed an abundance of two-component proteins of OmpR family and response regulators of nitrogen (NtrC family) and cell cycle, contributing to the signal transduction process (Supplementary Table 2). Sec preprotein translocases seem to be a also useful mechanism for intracellular trafficking of majority bacterial Aestuariivirgaceae, with apparent general export pathway composed of a complex of SecD, SecE, SecF, SecG and SecY in the cytoplasmic membrane [31]. Furthermore, we also found genes of secretion and vesicular transport of effector molecules. As for the transference of genetic material between cell-to-cell interactions and T4SS enzymes, only *Aestuariivirga sp003151375* (Palsa_927) and *Nordella sp005502925* (SCPDY, X1A and X2X) showed potential to use bacterial conjugation (Supplementary Table 2). As for motility, *JABDJG01 sp.* (METAPETRO_BR_BIN_54), *Aestuariivirga sp903930095* (Loclat_bin-06399), *Aestuariivirga sp902826365* (RBC017, RBC019 and RBC065), *Aestuariivirga sp012032065* (RU_4_17) and *JABDJG01 sp013002595* (SS_bin_17) showed a functional flagella (Fig. 2B e 2C). Although the flagella absence has been reported for some Rhizobiales, we can infer that the Che and DviK proteins in the Aestuariivirgaceae family species can help circumvent a lack of motility [32, 33].

Finally, the absence of general phenotype traits associated with nitrogen fixation (e.g., nitrogeneses - *nif*) and pathogenicity (*virB/D*) (Supplementary Table 2), which is present in many Rhizobiales, would suggest a free-living lifestyle in the Aestuariivirgaceae members. This hypothesis agrees with previous results described by Wang and collaborators [34], which showed Rhizobiales has an ancient origin (~ 1500 Mya), and the last common ancestor of this order indicates that the free-living lifestyle was the base of their evolutionary trajectory. The phylogenetic relationship of Aestuariivirgaceae with free-living bacteria (such as *Hyphomicrobium*) observed in this study, leads us to hypothesize that the family members described here are probable free-living bacteria.

Conclusion

In this study, we validate the monophyly of the Aestuariivirgaceae (Rhizobiales) family using phylogenomic methods, suggesting a basal split early taxonomic group. Together with functional annotation, we hypothesized that the presence of specific pathways (e.g., Glycolysis/Gluconeogenesis) and the absence of pathogenic genes in Aestuariivirgaceae could indicate a free-living lifestyle, similar to the Last Common Ancestor (LCA) of all Rhizobiales. These findings also reveal the presence of a versatile metabolism, from sugar degradation to hydrocarbon bioremediation, that allows these microorganisms to survive in the most diverse microbiomes, including soil and groundwater systems. Lastly, additional studies based on metatranscriptomics in environmental samples and culturomics of new Aestuariivirgaceae members will be necessary to identify and quantify gene functions predicted here.

(See figure on next page.)

Fig. 2 Functional profile of the Aestuariivirgaceae (Rhizobiales; Proteobacteria) family. **A** Abundance of general functions in each individual Aestuariivirgaceae genome. **B** Central metabolism of the Aestuariivirgaceae members. The model indicates the major putative functional predictions of the key pathways of Aestuariivirgaceae genomes. The pathways are highlighted by a pink colour and the question mark (?) symbol indicates incomplete pathways. A complete list of genes encoded by Aestuariivirgaceae genomes can be found in the Supplementary Table 2. Abbreviations: TCA, tricarboxylic acid cycle; ATP, Adenosine triphosphate. **C** Phenotype traits predicted by machine learning inferences

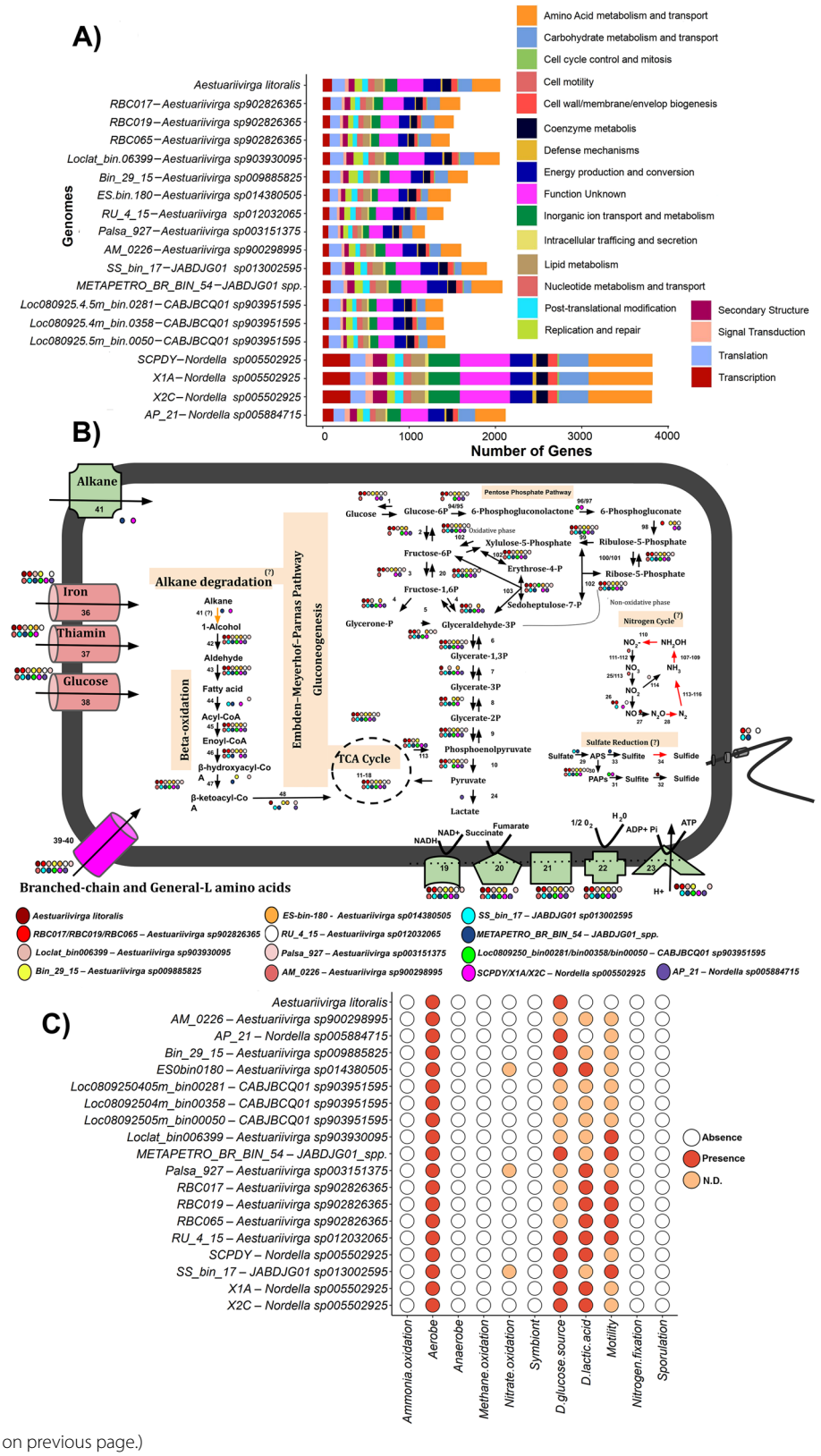

Fig. 2 (See legend on previous page.)

Methods

Sequencing and assembly of marine sediment metagenomes

The total DNA from 28 marine sediment samples (0–2 cm (2,5 a 37 m) depth) collected across Brazilian southwest islands was extracted using the Quick-DNA Miniprep Kit (Zimo Research). Metagenomics libraries were constructed using the Nextera DNA Flex Library Prep Kit (Illumina) according to the manufacturer's protocol. Sequencing was performed on an Illumina NextSeq 500 platform (2 X 150 bp) (Illumina, San Diego, CA) at Computational Genomics Unity Darcy Fontoura de Almeida (UGCDFA) of the National Laboratory of Scientific Computation (LNCC) (Petrópolis, RJ, Brazil). The marine sediment metagenomes were used to assemble genomes from metagenomes (MAGs) following these steps: Firstly, the Trimmomatic [35] was used to remove sequencing adapters and low-quality reads. Then, reads were assembled using Megahit [36]. Only contigs greater than 2500 bp were used in the binning step using Metabat2 [37]. To check the quality control of each individual potential genome (MAGs), we used the CheckM software [17] to estimate the completeness and contamination metrics. To estimate the taxonomy identification, we used the GTDB-tk software [38]. We used only MAGs with medium-quality draft (Completeness $\geq$50.0 and Contamination $\leq$5.0%) [14] in the taxonomic assignment.

Aestuariivirgaceae (Rhizobiales; Proteobacteria) genomes available in the public database

All microbial genomes assigned as Aestuariivirgaceae family were retrieved from the Genome Taxonomy Database (GTDB) (July 2021) [9]. To selected and build an representative dataset with good quality genomes, we follow these criteria: firstly, we selected all genomes presenting a medium-quality draft (Completeness $\geq$50.0 and Contamination $\leq$5.0%) based on the Minimum information about a single amplified genome (MISAG) standards [14].

Phylogenomic analysis

To estimate the phylogenetic position of the Aestuariivirgaceae family into the Rhizobiales order, we used a phylogenomic approach based on the alignment concatenation of 139 bacterial single-copy core genes (SCGs) (Supplementary Table 3) [39, 40]. Nineteen Aestuariivirgaceae genomes were used (Table 1) plus 39 Rhizobiales genomes and three other Bacteria (*Coraliomargarita akajimensis*, *Acidobacterium capsulatum* and *Escherichia coli*, which were used as outgroup). Each single-copy gene marker was identified using the HMM database from Campbell and collaborators [39] in Anvi'o software [40]. Each protein dataset was aligned using Muscle [41]. We excluded ambiguously aligned regions (−gt = 0.50) using trimAl v1.2 [42]. The alignments were concatenated to estimate the phylogeny using the JTT + CAT model in FastTree 2.0 software [43].

Functional genome annotation

Each genome was annotated using an automated annotation workflow (SABIA) [44] to identify the open reading frame (ORF) and assign all functions based on the fast orthology assignment and precomputed eggNOG v5.0 clusters implemented in the eggNOG-mapper [45]. COG Functional Categories were used to summarize general functions and KEGG KO was used to investigate the main metabolic pathways. Machine learning inferences were used to predict the phenotype traits of each individual genomes using PhenDB [46].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12866-021-02354-4.

Additional file 1 : **Supplementary Table 1**. General informations about the metagenomic dataset used in this study to reconstruct METAPETRO_BR_BIN_54 genome. **Supplementary Table 2.** General features of metabolic pathways. **Supplementary Table 3**. Bacterial single-copy core genes (SCGs) from Campbell et al. [39] used in the phylogenomic analysis. (From Campbell et al. [39]).

Acknowledgements

Not applicable.

Authors' contributions

L.N.L., C.R.J., L.P.C., and A.T.R.V. contributed to the study conception and design. Data collection and analysis were performed by L.N.L., F.M.C., A.G., A.P.C.G., C.R.J., L.P.C., and A.T.R.V. The first draft of the manuscript was written by L.N.L., F.M.C., L.P.C., and A.T.R.V. The final draft of manuscript was reviewed by L.N.L., F.M.C., C.R.J., L.P.C., and A.T.R.V. All authors read and approved the final version of the manuscript.

Funding

A.T.R.V. is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (303170/2017-4) and FAPERJ (E-26/202.903/20). This study was supported by Petrobras (Process number: 2018/00190-8). Termo de cooperação n° 5900.0109896.18.9/SAP: 4600579545 - Sequenciamento de DNA e análises bioinformáticas para metagenômica - Petrobras. The authors acknowledge the National Laboratory for Scientific Computing (LNCC/MCTI, Brazil) for providing HPC resources of the SDumont supercomputer, which have contributed to the research results reported within this paper. URL: http://sdumont.lncc.br.

Availability of data and materials

Metagenome-assembled genome (MAG) METAPETRO-BIN-54 was deposited in DDBJ/ENA/GenBank under the accession JAEKFU000000000. The version described in this paper is version JAEKFU010000000. Additional Aestuariivirgaceae genomes were deposited in public sequence repositories (GCF_003234965.1, GCA_003151375.1, GCF_005502925.1, GCF_005502975.1, GCF_005502345.1, GCA_005884715.1, GCA_009885825.1, GCA_014380505.1, GCA_012032065.1, GCA_013002595.1, GCA_900298995.1, GCA_903930095.1, GCA_903944365.1, GCA_903951595.1, GCA_903958745.1, GCA_902826365.1, GCA_902826675.1 and GCA_902826905.1).

Declarations

Ethics approval and consent to participate
This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no conflicts of interest.

Author details
[1]Bioinformatics Laboratory, National Laboratory of Scientific Computing (LNCC), Av. Getúlio Vargas, 333 - Quitandinha, Petrópolis, RJ 25651-076, Brazil. [2]Leopoldo Americo Miguez de Melo Research Center (CENPES-Petrobras), Rio de Janeiro, RJ, Brazil.

Received: 27 August 2021 Accepted: 6 October 2021
Published online: 28 October 2021

References
1. Carvalho FM, Souza RC, Barcellos FG, Hungria M, Vasconcelos ATR. Genomic and evolutionary comparisons of diazotrophic and pathogenic bacteria of the order Rhizobiales. BMC Microbiol. 2010;10:37. https://doi.org/10.1186/1471-2180-10-37.
2. Clúa J, Roda C, Zanetti ME, Blanco FA. Compatibility between legumes and rhizobia for the establishment of a successful nitrogen-fixing symbiosis. Gene. 2018;9(3):125. https://doi.org/10.3390/genes9030125.
3. Barton I, Fuqua C, Platt T. Ecological and evolutionary dynamics of a model facultative pathogen: Agrobacterium and crown gall disease of plants. Environ Microbiol. 2018;20(1). https://doi.org/10.1111/1462-2920.13976.
4. Bates ST, Cropsey GWG, Caporaso JG, Knight R, Fierer N. Bacterial communities associated with the lichen symbiosis. Appl Environ Microbiol. 2011;77(4):1309–14. https://doi.org/10.1128/AEM.02257-10.
5. Kosoy M, Goodrich I. Comparative ecology of Bartonella and Brucella infections in wild carnivores. Front Vet Sci. 2019;5:322. https://doi.org/10.3389/fvets.2018.00322.
6. Rosales S, Clark AS, Huebner LK, Ruzicka RR, Muller EM. Rhodobacterales and Rhizobiales are associated with stony coral tissue loss disease and its suspected sources of transmission. Front Microbiol. 2020;11:681. https://doi.org/10.3389/fmicb.2020.00681.
7. Minich JJ, Morris MM, Brown M, Doane M, Edwards MS, Michael TP, et al. Elevated temperature drives kelp microbiome dysbiosis, while elevated carbon dioxide induces water microbiome disruption. PLoS One. 2018. https://doi.org/10.1371/journal.pone.0192772.
8. Ng JCY, Chiu MY. Changes in biofilm bacterial communities in response to combined effects of hypoxia, ocean acidification and nutrients from aquaculture activity in three fathoms cove. Mar Pollut Bull. 2020;156:1–12. https://doi.org/10.1016/j.marpolbul.2020.111256.
9. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy blleased on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36:996–1004. https://doi.org/10.1038/nbt.4229.
10. Diamond S, Andeer PF, Li Z, Crits-Christoph A, Burstein D, Anantharaman K, et al. Mediterranean grassland soil C–N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms. Nat Microbiol. 2019;4:1356–67. https://doi.org/10.1038/s41564-019-0449-y.
11. Pedron R, Esposito A, Bianconi I, Pasolli E, Tett A, Asnicar F, et al. Genomic and metagenomic insights into the microbial community of a thermal spring. Microbiome. 2019;7:8. https://doi.org/10.1186/s40168-019-0625-6.
12. Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB, Zayed AAF, et al. Genome-centric view of carbon processing in thawing permafrost. Nature. 2018;560:49–54. https://doi.org/10.1038/s41586-018.
13. Levy-Booth DJ, Hashimi A, Roccor R, Liu L-Y, Renneckar S, Eltis LD, et al. Genomics and metatranscriptomics of biogeochemical cycling and degradation of lignin-derived aromatic compounds in thermal swamp sediment. ISME J. 2020:1–15. https://doi.org/10.1038/s41396-020-00820.
14. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35:725–31. https://doi.org/10.1038/nbt.3893.
15. Imachi H, Nobu MK, Nakahara N, Morono Y, Ogawara M, Takaki Y, et al. Isolation of an archaeon at the prokaryote–eukaryote interface. Nature. 2020;577:519–25. https://doi.org/10.1038/s41586-019-1916-6.
16. Li X, Salam N, Li J, Chen Y-M, Yang Z, Han M, et al. Aestuariivirga litoralis gen. nov., sp. nov., a proteobacterium isolated from a water sample, and proposal of Aestuariivirgaceae fam. nov. Int J Syst Evol Microbiol. 2019;69:299–306. https://doi.org/10.1099/ijsem.0.003087.
17. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55. https://doi.org/10.1101/gr.186072.114.
18. Santos C Jr, Logares R, Henrique-Silva F. Degradation of terrestrial organic matter by aquatic microbial genomes in the Amazon River. Re Square. 2020. https://doi.org/10.21203/rs.3.rs-32535/v2.
19. Ruuskanen M, Colby G, Pierre K, Louis V, Aris-Brosou S, Poulain A. Microbial genomes retrieved from high Arctic lake sediments encodefor adaptation to cold and oligotrophic environments. Limnol Oceanogr. 2019;S1:S233–47.
20. Zeng Y, Chen X, Madsen A, Zervas A, Nielsen T, Andrei A, et al. Potential rhodopsin- and bacteriochlorophyll-based dual phototrophy in a high Arctic glacier. mBio. 2020;11(6):e02641–20.
21. Waterworth S, Isemonger E, Rees E, Dorrington R, Kwan C. Conserved bacterial genomes from two geographically isolated peritidal stromatolite formations shed light on potential functional guilds. Environ Microbiol. 2021;13(2):126–37.
22. Chen YJ, Leung PM, Wood JL, et al. Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. ISME J. 2021. https://doi.org/10.1038/s41396-021-00988-w.
23. Spasov E, Tsuji JM, Hug LA, et al. High functional diversity among Nitrospira populations that dominate rotating biological contactor microbial communities in a municipal wastewater treatment plant. ISME J. 2020;14:1857–72. https://doi.org/10.1038/s41396-020-0650-2.
24. La Scola B, Barrassi L, Raoult D. A novel alpha-Proteobacterium, Nordella oligomobilis gen. nov., sp. nov., isolated by using amoebal co-cultures. Res Microbiol. 2004;155(1):47–51. https://doi.org/10.1016/j.resmic.2003.09.012.
25. Pini F, Galardini M, Bazzicalup M, Mengoni A. Plant-bacteria association and symbiosis: are there common genomic traits in alphaproteobacteria? Genes. 2011;2(4):1017–32. https://doi.org/10.3390/genes2041017.
26. Yang K, Heath LS, Setubal JC. REGEN: ancestral genome reconstruction for bacteria. Genes (Basel). 2012;3:423–43. https://doi.org/10.3390/genes3030423.
27. Marreiros BC, Calisto F, Castro PJ, Duarte AM, Sena FV, Silva AF, et al. Exploring membrane respiratory chains. Biochim Biophys Acta (BBA) - Bioenerg. 2016;1857:1039–67. https://doi.org/10.1016/j.bbabio.2016.03.028.
28. Lin H, Lou B, Glynn JM, Doddapaneni H, Civerolo EL, Chen C, et al. The complete genome sequence of 'Candidatus Liberibacter solanacearum', the bacterium associated with potato Zebra Chip disease. PLoS One. 2011;6:e19135. https://doi.org/10.1371/journal.pone.0019135.
29. Spang A, Saw J, Jørgensen S, Zaremba-Niedzwiedzka K, Martijn J, Lind A, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. 2015;521:173–9.
30. van Beilen JB, Wubbolts MG. Genetics of alkane oxidation by pseudomonas oleovorans. Biodegradation. 1994;5(3–4):161–74. https://doi.org/10.1007/BF00696457.
31. Wickner W, Driessen AJ, Hartl FU. The enzymology of protein translocation across the *Escherichia coli* plasma membrane. Annu Rev Biochem. 1991;60:101–24. https://doi.org/10.1146/annurev.bi.60.070191.000533.
32. Stephens BB, Loar SN, Alexandre G. Role of CheB and CheR in the complex chemotactic and aerotactic pathway of Azospirillum brasilense. J Bacteriol. 2006 Jul;188(13):4759–68. https://doi.org/10.1128/JB.00267-06.
33. Heindl JE, Crosby D, Brar S, Pinto JF, Singletary T, Merenich D, et al. Reciprocal control of motility and biofilm formation by the PdhS2

two-component sensor kinase of Agrobacterium tumefaciens. Microbiol (Reading). 2019 Feb;165(2):146–62. https://doi.org/10.1099/mic.0.000758.
34. Wang S, Meade A, Lam H-M, Luo H. Evolutionary timeline and genomic plasticity underlying the lifestyle diversity in Rhizobiales. MSystems. 2020;5. https://doi.org/10.1128/mSystems.00438-20.
35. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. https://doi.org/10.1093/bioinformatics/btu170.
36. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674–6. https://doi.org/10.1093/bioinformatics/btv033.
37. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7. https://doi.org/10.7717/peerj.735.
38. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics. 2020;36:1925–7. https://doi.org/10.1093/bioinformatics/btz848.
39. Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, Woyke T, et al. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. PNAS. 2013;110:5540–5. https://doi.org/10.1073/pnas.1303090110.
40. Eren M, Esen O, Quince C, Vineis J, Morrison H, Sogin M, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3:e1319. https://doi.org/10.7717/peerj.1319.
41. Edgar R. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinforma. 2004;5(113):1–19. https://doi.org/10.1186/1471-2105-5-113.
42. Capella-Gutiérrez S, Silla-Martínez J, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.
43. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490. https://doi.org/10.1371/journal.pone.0009490.
44. Almeida LGP, Paixão R, Souza RC, da Costa GC, Barrientos FJA, dos Santos MT, et al. A system for automated bacterial (genome) integrated annotation—SABIA. Bioinformatics. 2004;20:2832–3. https://doi.org/10.1093/bioinformatics/bth273.
45. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol. 2017;34:2115–22. https://doi.org/10.1093/molbev/msx148.
46. Feldbauer R, Schulz F, Horn M, Rattei T. Prediction of microbial phenotypes based on comparative genomics. BMC Bioinformatics. 2015;16(Suppl 14):S1.
47. Buck M, Garcia SL, Fernandez L, et al. Comprehensive dataset of shotgun metagenomes from oxygen stratified freshwater lakes and ponds. Sci Data. 2021;8:131. https://doi.org/10.1038/s41597-021-00910-1.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC