Santos-Júnior et al. Microbiome (2020) 8:151
https://doi.org/10.1186/s40168-020-00930-w

Microbiome

RESEARCH Open Access

Uncovering the genomic potential of the Amazon River microbiome to degrade rainforest organic matter

Célio Dias Santos-Júnior[1,2], Hugo Sarmento[3], Fernando Pellon de Miranda[4], Flávio Henrique-Silva[1*] and Ramiro Logares[5*]

Abstract

Background: The Amazon River is one of the largest in the world and receives huge amounts of terrestrial organic matter (TeOM) from the surrounding rainforest. Despite this TeOM is typically recalcitrant (i.e. resistant to degradation), only a small fraction of it reaches the ocean, pointing to a substantial TeOM degradation by the river microbiome. Yet, microbial genes involved in TeOM degradation in the Amazon River were barely known. Here, we examined the Amazon River microbiome by analysing 106 metagenomes from 30 sampling points distributed along the river.

Results: We constructed the *Amazon River basin Microbial non-redundant Gene Catalogue* (AMnrGC) that includes ~ 3.7 million non-redundant genes, affiliating mostly to bacteria. We found that the Amazon River microbiome contains a substantial gene-novelty compared to other relevant known environments (rivers and rainforest soil). Genes encoding for proteins potentially involved in lignin degradation pathways were correlated to tripartite tricarboxylates transporters and hemicellulose degradation machinery, pointing to a possible *priming effect*. Based on this, we propose a model on how the degradation of recalcitrant TeOM could be modulated by labile compounds in the Amazon River waters. Our results also suggest changes of the microbial community and its genomic potential along the river course.

Conclusions: Our work contributes to expand significantly our comprehension of the world's largest river microbiome and its potential metabolism related to TeOM degradation. Furthermore, the produced gene catalogue (AMnrGC) represents an important resource for future research in tropical rivers.

Keywords: Amazon River, Freshwater bacteria, Biodiversity, Metagenomics, Lignin degradation, Cellulose degradation, Priming effect, Gene catalogue

* Correspondence: dfhs@ufscar.br; ramiro.logares@icm.csic.es
[1]Molecular Biology Laboratory, Department of Genetics and Evolution – DGE, Universidade Federal de São Carlos – UFSCar, Rod. Washington Luis KM 235 - Monjolinho, São Carlos, SP 13565-905, Brazil
[5]Institute of Marine Sciences (ICM), CSIC, Passeig Marítim de la Barceloneta 37-49, ES08003, Barcelona, Catalonia, Spain
Full list of author information is available at the end of the article

Background

Continental waters play a major biogeochemical role by linking terrestrial and marine ecosystems [1]. In particular, rainforest rivers receive large amounts of terrestrial organic matter (TeOM), which may then reach the ocean. TeOM is difficult to degrade (i.e. recalcitrant), being normally processed in rivers by microorganisms, stimulating its conversion to carbon dioxide [2–4]. Therefore, riverine microbiomes should have evolved metabolisms capable of degrading TeOM. Even though the gene repertoire of river microbiomes can provide crucial insights to understand the links between terrestrial and marine ecosystems, as well as the fate of organic matter synthesized on land, very little is known about the genomic machinery of riverine microbes that degrade TeOM.

Microbiome gene catalogues allow the characterization of functional repertoires, linking genes with ecological function and ecosystem services. Recently, large gene catalogues have been produced for the global ocean [5–7], soils [8] and animal guts [9, 10]. In particular, ~ 47 million genes have been reported for the global ocean microbiome [11] and ~ 160 million genes for the global topsoil microbiome [8]. Although functional metagenomics was already performed in the Amazon River [12–18], so far, no comprehensive gene catalogue was generated, which hinders our understanding of the genomic machinery that degrades almost half of the 1.9 Pg C discharged into rivers every year as recalcitrant TeOM [1]. This is particularly relevant in tropical rainforests, like the Amazon forest, which accounts for ~ 10% of the global primary production, fixing 8.5 Pg C per year [19, 20]. The Amazon River basin comprises almost 38% of continental South America [21], and its discharge accounts for 18% of the world's inland-water inputs to the oceans [22]. Despite its relevance for global-scale processes, there is a limited understanding of the Amazon River microbiome.

Large amounts of organic and inorganic particulate material [23] turn the Amazon River into a turbid system. High turbidity reduces light penetration, and consequently, the Amazon River has very low rates of phytoplankton production [24], meaning that TeOM is the major carbon source for microbial growth [25]. High respiration rates in Amazon River waters generate a CO_2 super-saturation that leads to its outgassing to the atmosphere. Overall, Amazon River outgassing accounts for 0.5 Pg C per year to the atmosphere [26], almost equivalent to the amount of carbon sequestered by the forest [19, 20]. Despite the predominantly recalcitrant nature of the TeOM that is discharged into the Amazon River, heterotrophic microbes are able to degrade up to ~ 55% of the lignin produced by the rainforest [27, 28]. The unexpectedly high degradation rates of some TeOM compounds in the river was recently explained by the availability of labile compounds that promote the degradation of recalcitrant counterparts, a mechanism known as *priming effect*, which has been observed in incubation experiments [28].

Determining the repertoire of gene functions in the Amazon River microbiome is one of the key steps to understand the mechanisms involved in the degradation of complex TeOM produced in the rainforest. Given that most TeOM present in the Amazon River is lignin and cellulose [27–31], the functions associated with their degradation were expected to be widespread in the Amazon microbiome. Instead, these functions exhibited very low abundances [16, 17, 32], highlighting our limited understanding of the enzymes involved in the degradation of lignin and cellulose in aquatic systems.

Cellulolytic bacteria use an arsenal of enzymes with synergistic and complementary activities to degrade cellulose. For example, glycosyl hydrolases (GHs) catalyse the hydrolysis of glycoside linkages, while polysaccharide esterases support the action of GHs over hemicelluloses and polysaccharide lyases promote depolymerization [33, 34]. In contrast, lignin is more resistant to degradation [35, 36], since its role is preventing microbial enzymes from degrading labile cell-wall polysaccharides [37]. The microbial production of extracellular hydrogen peroxide, a highly reactive compound, is the first step of lignin oxidation mediated by enzymes, like lignin peroxidase, manganese-dependent peroxidase and copper-dependent laccases [33]. Lignin oxidation also produces a complex mixture of aromatic compounds, which compose the humic fraction of dissolved carbon detected in previous studies in the Amazon River [29, 30]. Our knowledge of bacterial-mediated lignin degradation in the Amazon River is limited; however, it is known that in tropical streams bacterial lignocellulose degradation tends to occur in the entire water column, being slow and also predominantly modulated by bacteria in anoxic regions close to sediments [38–41].

Here, we produced the first gene catalogue of the world's largest rainforest river by analysing 106 metagenomes (~ 500×10^9 base pairs), originating from 30 sampling points covering a total of ~ 2106 km, from the upper Solimões River to the Amazon River plume in the Atlantic Ocean. This gene catalogue was used to examine the genomic machinery of the Amazon River microbiome potentially responsible for metabolizing large amounts of organic carbon originating from the surrounding rainforest. Specifically, we ask: How novel is the gene repertoire of the Amazon River microbiome? Which are the main functions potentially associated with TeOM degradation? Do TeOM degradation-related genes and functions display a spatial distribution pattern? And finally, is there any evidence of *priming effect* in TeOM degradation?

Results

Cataloguing the genes of the Amazon River microbiome

Amazon River genes were predicted after co-assembling 106 metagenomes (Supplementary Tables 1 and 2 in Additional file 1) in groups that shared the same geographic origin (Fig. 1a). We predicted 6,074,767 genes longer than 150 bp, allowing for alternative initiation codons. After redundancy removal by clustering genes with an identity > 95% and an overlap > 90% of the shorter gene, the *Amazon River basin Microbial non-redundant Gene Catalogue* (AMnrGC) included 3,748,772 non-redundant genes, with half of the genes with a length ≥ 867 bp (publicly available in Zenodo, doi: https://doi.org/10.5281/zenodo.1484504). About 52% of the AMnrGC genes were annotated with at least one database, while ~ 86% of the annotated genes were simultaneously annotated using two or more different databases. The recovered gene and functional diversity seemed to be representative of this microbiota as indicated by the leveling off of the rarefaction curves of genes and functions (Fig. 1c).

The Amazon River microbiome differed from other microbiomes

We compared the metagenomic information contained in the Amazon River microbiome with that from the Amazon rainforest soil and other available temperate rivers (Canada watersheds and Mississippi River) using k-mers (Supplementary Table 3 in Additional file 1). The k-mer diversity comparison of these microbiomes indicated that they are different in terms of genomic composition (Fig. 1d), forming groups of heterogeneous constitution (significant β dispersion [that is, average distance of samples to the group centroid]—PERMUTEST, $F = 25.7$, $p < 0.001$). In particular, the k-mer composition of Amazon River samples was markedly different to the other microbiomes (PERMANOVA, $R^2 = 0.10$, $p = 9.99 \times 10^{-5}$; ANOSIM, $R = 0.27$, $p < 0.001$), which suggests that this

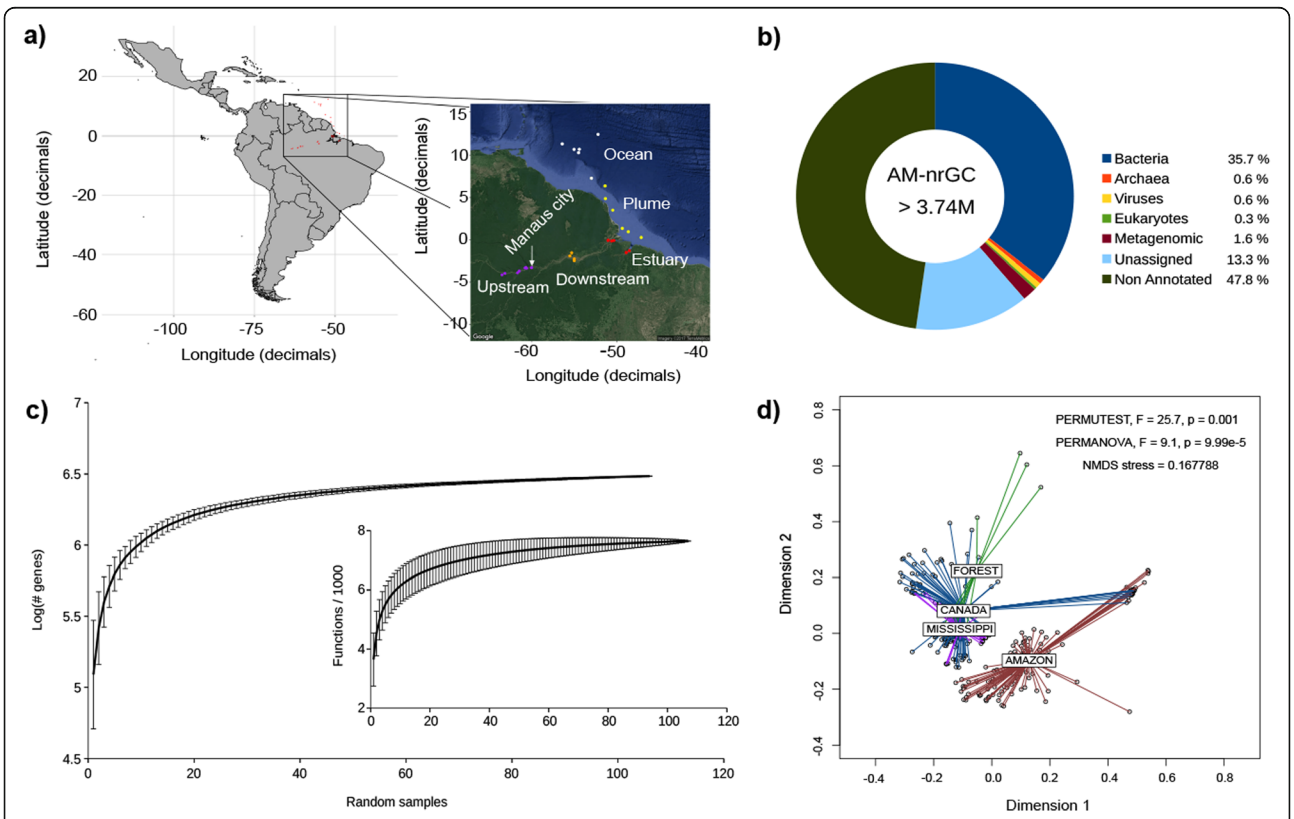

Fig. 1 The Amazon River basin Microbial Non-Redundant Gene Catalogue (AMnrGC). **a** Distribution of the 106 metagenomes used in this work over the five sections of the Amazon River: Upstream (purple dots), Downstream (orange dots), Estuary (red dots), Plume (yellow dots) and coastal Ocean (white dots). **b** Taxonomic classification of the ~ 3.7 million genes in the AMnrGC. "Unassigned" genes were not assigned taxonomy, but they were functionally assigned, differently from "non-annotated" genes, which do not have any ortholog. Those genes displaying orthology to poorly characterized genes found in metagenomes were referred to as "Metagenomic". **c** Rarefaction curves of non-redundant genes and PFAM families (internal plot). Note both point towards saturation. **d** NMDS comparing the Amazon river microbiome with other microbiomes based on information content [k-mer composition]. Amazon River (AMAZON), Amazon forest soil (FOREST), Canada watersheds (CANADA) and Mississippi River (MISSISSIPPI)

basin, or tropical rainforest rivers in general, may contain specific gene repertoires.

The metagenomic composition (k-mer based) of the five sampled sections of the Amazon River (i.e. upstream, downstream, estuary, plume and ocean) displayed significant differences (PERMANOVA test, $F = 2.34$, $p < 9.9e{-}5$; Fig. 2a), indicating that river sections may include different gene assemblages. These groups representing river sections were considered heterogeneous, as there was a significant β dispersion ($F = 7.7$, $p = 1e{-}3$) among metagenomic samples in each group (Fig. 2b). Additionally, the freshwater samples from different river sections (upstream, downstream and estuary) had shorter distances to centroids than those of brackish and marine samples (Fig. 2b). Even though we have used different size fractions to capture free-living or particle-attached microbes, this did not influence the k-mer composition (PERMANOVA test, $F = 3.62$, $p = 0.06$; β dispersion, $F = 3.62$, $p = 0.074$; Fig. 2c).

Gene identification

About 48% of the AMnrGC genes could not be annotated due to lack of orthologs in reference databases. Besides, even though ~ 1.6% of the genes in the AMnrGC were previously found in metagenomic studies, they were poorly characterized, without being assigned to a particular taxon (here referred to as "Metagenomic" genes; Fig. 1b). Genes annotated exclusively through hidden Markov models (HMM) represented 13.3% of the AMnrGC. As the annotation using HMM profiles does not rely on direct orthology to specific sequences, but on orthology to a protein family (which may include mixed taxonomic signal), we could not assign taxonomy to those genes and they are referred to as "Unassigned genes" (Fig. 1b).

Overall, the previous results highlight our limited understanding about the gene composition of the Amazon River microbiome, where most proteins (61.1%) do not have orthologs in main reference databases. Prokaryotic genes (35.7% bacterial and 0.6% archaeal) constituted the majority of the AMnrGC, with only 0.3% and 0.6% of the genes having eukaryotic or viral origin, respectively (Fig. 1b).

Core metabolisms

Functional analysis comprised prokaryotic and eukaryotic genes matching COG, KEGG and/or PFAM databases.

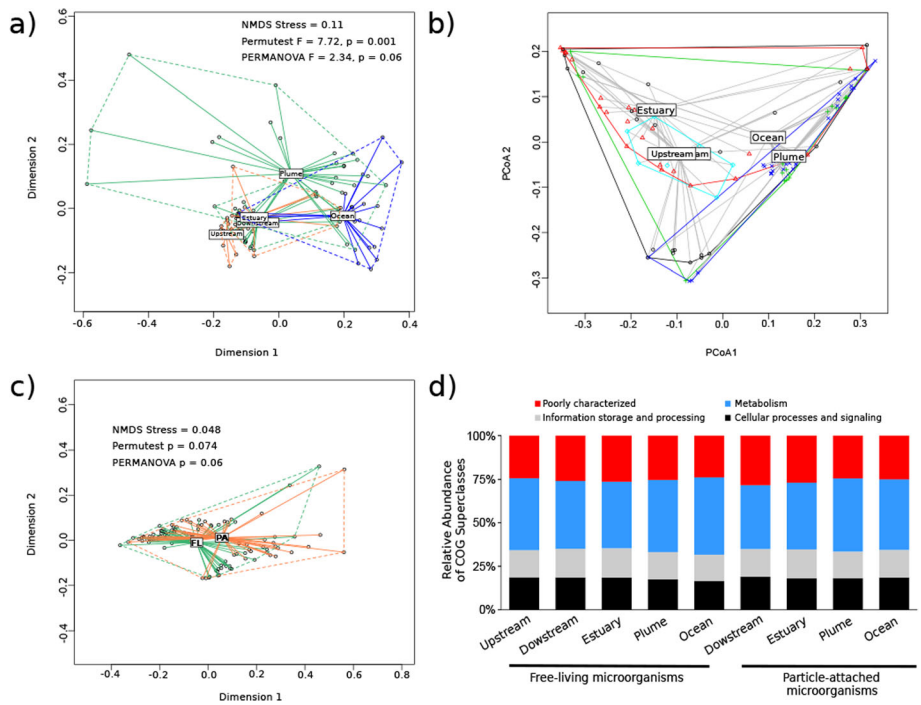

Fig. 2 Metagenomic and COG composition of the studied sections of the Amazon River microbiome. Ordination of metagenomes composing the different river sections based on the Jaccard distances calculated from the presence-absence of k-mers in each sample (**a-c**). NMDS groups were statistically different [PERMANOVA, $F = 2.34$, p value = 9.99e−5] (**a**), displaying intragroup heterogeneity [β dispersion; PERMUTEST, $F = 7.72$, p value = 0.001] (**b**). Metagenomic composition of the Amazon River microbiome according to microbial lifestyle (free-living (FL) vs. particle-attached (PA)) (**c**). NMDS groups were statistically different [PERMANOVA, $F = 3.62$, p value = 0.06], displaying intragroup homogeneity [β dispersion; PERMUTEST, $F = 3.62$, p value = 0.074]. COG composition across size fractions and sections of the Amazon River (**d**). Gene functions grouped into COG superclasses are shown per river section and microbial lifestyle (free-living vs. particle-attached). The Upstream river section is not shown in the particle-attached fraction since it was not sampled

The superclass "Metabolic processes" from the Clusters of Orthologous Genes (COG) database comprises those gene functions belonging either to energy production and conversion, amino acids, nucleotides, carbohydrates, coenzymes, lipids and inorganic ions transport and metabolism, secondary metabolites biosynthesis, transport and catabolism. This superclass was the most abundant in the AMnrGC (35.8% of the genes annotated with COG; Fig. 2d). Genes with unknown function represented 21.4% of the COG annotated proteins. Functionally, microbial lifestyle (i.e. free-living vs. particle-attached) did not influence the COG superclass distribution (Fig. 2d).

Core metabolic functions are those involved in cell or ecosystem homeostasis, normally representing the minimal metabolic machinery needed to survive in a given environment. KEGG and PFAM databases were used to determine the bacterial functional core, allowing also the identification of metabolic pathways. Core functions represented ~ 8% of KEGG and PFAM functions and were mostly related to general carbon metabolism, being predominantly associated with organic matter oxidation to CO_2 and respiration byproducts heading to acetogenic pathways. Apart from core metabolisms, abundant proteins can reveal essential biochemical pathways in microbiomes. The top 100 most abundant functions in the bacterial core were "house-keeping" functions involved in main metabolic pathways (e.g. carbohydrate metabolism, *quorum* sensing, transporters and amino acid metabolism), as well as important protein complexes (e.g. RNA and DNA polymerases and ATP synthase). The non-core metabolism suggests adaptations to a complex environment, including multiple genes related to xenobiotic biodegradation and secondary metabolism (that is, the production and consumption of compounds not directly related to cell survival).

The potential TeOM degradation machinery

A total of 6516 genes from the AMnrGC were identified as taking part in the potential TeOM degradation machinery from the Amazon River microbiome, being divided into cellulose degradation (143 genes), hemicellulose degradation (92 genes), lignin oxidation (73 genes), lignin-derived aromatic compounds transport and metabolism (2324 genes) and tricarboxylate transport (3884 genes) (Figs. 3, 4, and 5). The large number of gene variants associated with the metabolism of lignin-derived compounds and the transport of tricarboxylates (Fig. 4) reflects the variety of molecules generated during the lignin oxidation process in the Amazon River. No significant differences were found in the composition and distribution of genes in samples belonging to different microbial lifestyles (i.e. free-living vs. particle-attached). Eukaryotic contributions to the analysed functions were small (0.5–0.6%); thus, the machinery analysed hereafter is mostly prokaryotic.

Lignin oxidation and deconstruction of cellulose and hemicellulose

TeOM consists of biopolymers, so the first step of its microbial-based degradation consists in converting polymers into monomers. Thus, the identified genes potentially involved in the oxidation of lignin and the degradation of cellulose and hemicellulose were investigated (Fig. 3). We observed a ubiquitous dominance of glycosyl hydrolase GH3, related to cellulose degradation. This function represented 63.2–65.3% of the genes possibly associated with this catabolic step across all river sections (71 ± 8 genes per section) (Fig. 3). In turn, hemicellulose degradation is potentially performed mostly by glycosyl hydrolase GH10 (52-56% of genes) in all river sections (35 ± 6 genes per section). Analysis of gene taxonomy (Fig. 3) indicated that cellulose and hemicellulose hydrolysis could be carried out predominantly by known taxa in fresh (70–81%) and brackish waters (58–79%). Cellulose degradation is likely performed by *Betaproteobacteria* and *Actinobacteria* in freshwaters, while *Bacteroidetes*, *Alphaproteobacteria* and *Gammaproteobacteria* possibly dominate this step in the ocean and plume sections (Fig. 3). A limited fraction of hemicellulose degradation seems to be associated to *Gamma-* and *Deltaproteobacteria*, which display a ubiquitous distribution along the river course. In turn, *Actinobacteria* and *Betaproteobacteria* (in freshwaters), and *Alphaproteobacteria* (in brackish waters) seem to contribute predominantly with functions related to hemicellulose degradation in different river sections (Fig. 3).

We found that lignin oxidation in the Amazon River may be mainly mediated by dye-decolorizing peroxidases (DyPs), as 61.5–71.2% of the genes potentially involved in this step were predominantly associated with freshwater areas. Only laccases (19 ± 2 genes per section) and peroxidases (42 ± 6 genes per section) were found in the Amazon River microbiome, no other families involved in lignin oxidation, like phenolic acid decarboxylase or glyoxal oxidase, were found (Fig. 3). Lignin oxidation appears to be encoded by genes belonging predominantly to taxonomically unassigned taxa (36–42%) in all river sections, as well as *Betaproteobacteria* in freshwaters and *Alpha-* / *Gammaproteobacteria* in brackish waters. Moreover, there is a possible redundancy of functions in *Actinobacteria* and *Alphaproteobacteria*, as they have contrasting abundances in fresh- and brackish waters.

Lignin-derived compound transport

Lignin-derived aromatic compounds need to be transported from the extracellular environment to the cytoplasm prior to their degradation. Transporters that could

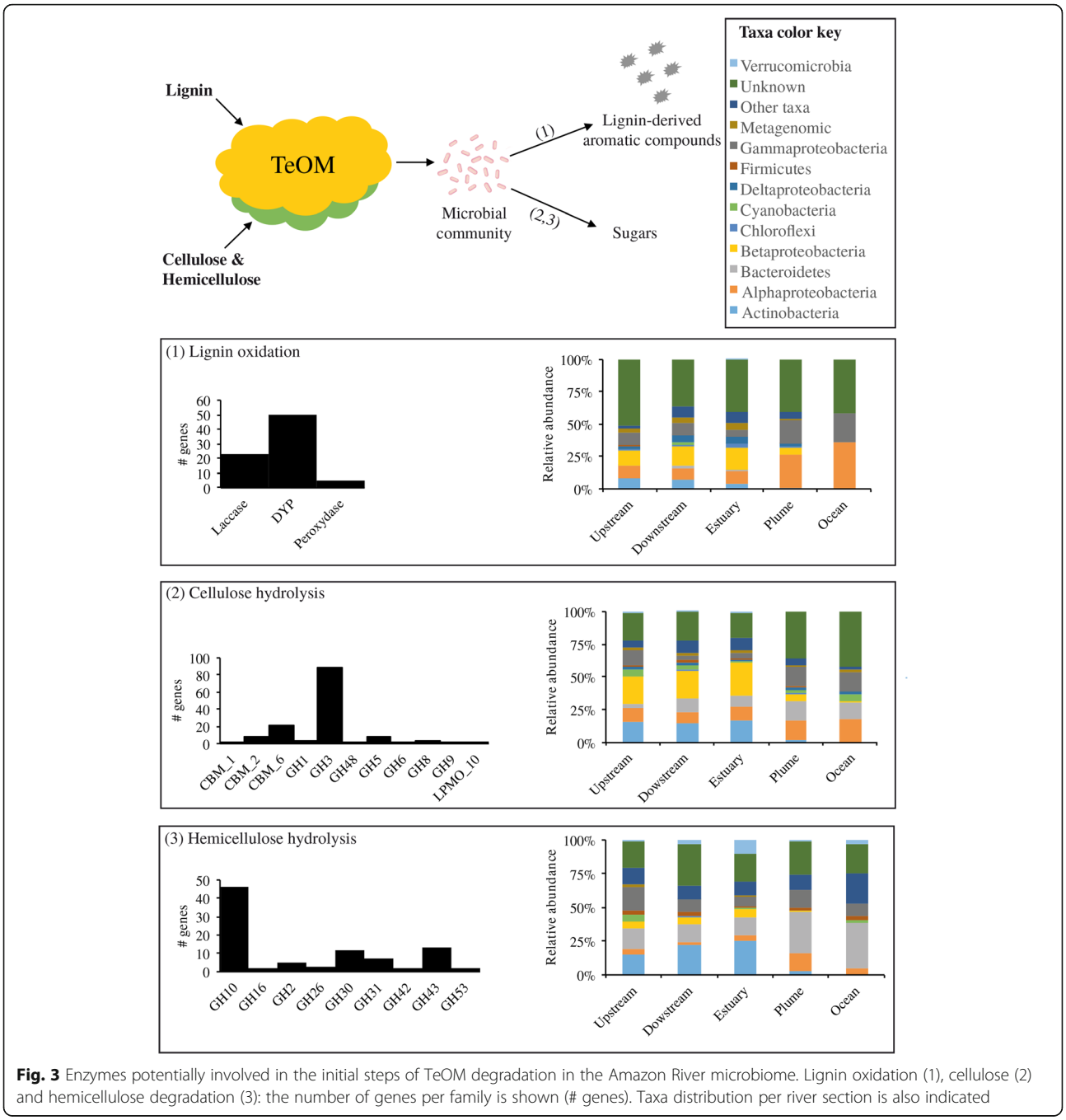

Fig. 3 Enzymes potentially involved in the initial steps of TeOM degradation in the Amazon River microbiome. Lignin oxidation (1), cellulose (2) and hemicellulose degradation (3): the number of genes per family is shown (# genes). Taxa distribution per river section is also indicated

be associated with lignin degradation (AAHS family and ABC transporters) were found in the AMnrGC (Fig. 4). Alphaproteobacterial AAHS and ABC genes had an increased abundance in plume and ocean samples (Fig. 4). Thus, *Alphaproteobacteria* contribute with their functional machinery to the metabolism of lignin-derived compounds in the Amazon River ecosystem. ABC transporters from *Betaproteobacteria* were enriched in freshwater or brackish river sections, while ABC transporters belonging to *Alphaproteobacteria* were enriched in sections with a higher salinity (Fig. 4). Different to the main taxa potentially involved in the degradation of TeOM, *Actinobacteria* had a smaller functional contribution to the transport of lignin-derived compounds, being more uniformly distributed along the river course (Fig. 4).

The tripartite tricarboxylate transporting (TTT) system is composed by three proteins, where *tctC* is responsible for capturing substrates in the extracellular space and bringing them to the transporting channel made by the proteins *tctA* and *tctB*, which recognize the

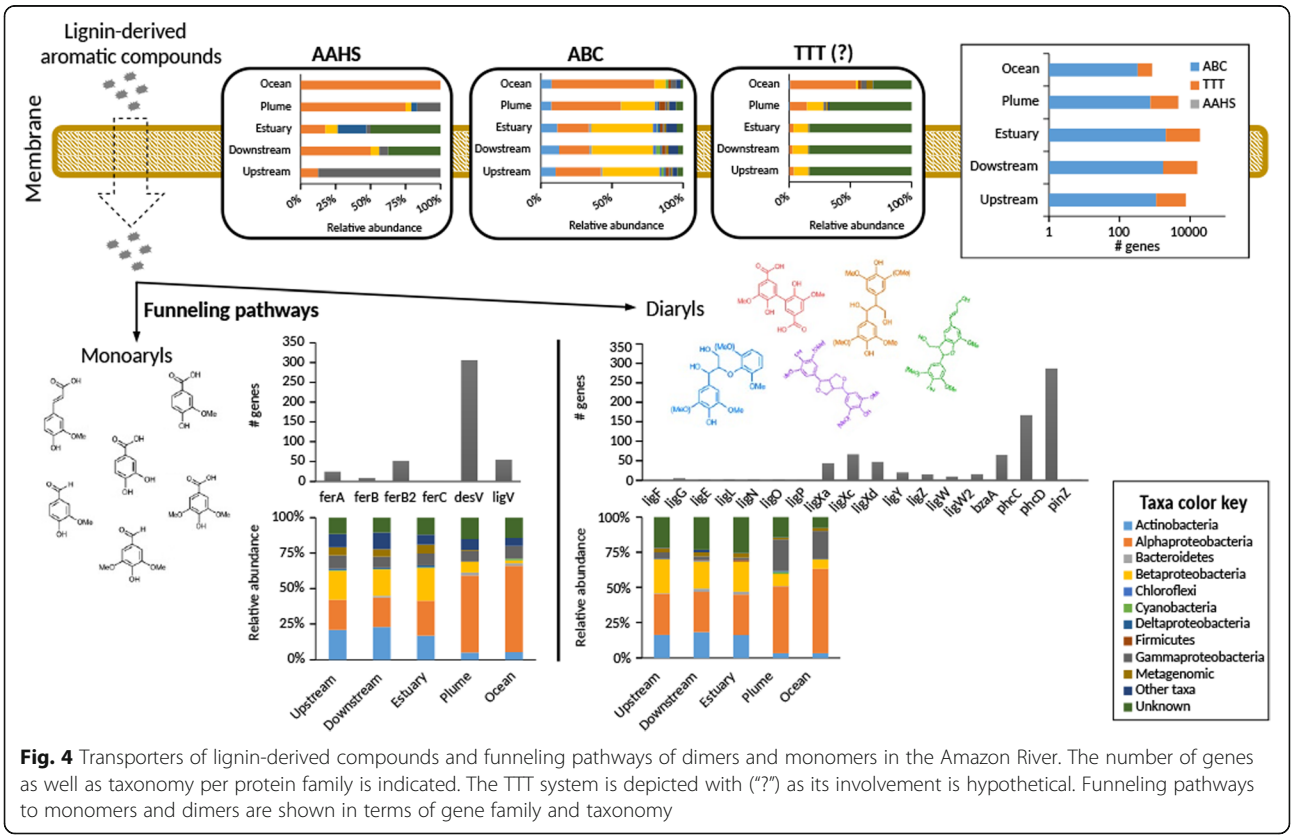

Fig. 4 Transporters of lignin-derived compounds and funneling pathways of dimers and monomers in the Amazon River. The number of genes as well as taxonomy per protein family is indicated. The TTT system is depicted with ("?") as its involvement is hypothetical. Funneling pathways to monomers and dimers are shown in terms of gene family and taxonomy

substrate binding protein and move the substrate across membrane into the cytoplasm. The TTT system is suitable for transporting humic acids, and therefore lignin-derived compounds. There was a large number of gene variants (from 10 to 100) associated with the substrate binding proteins (*tctC*). As each protein is specific to one or a few substrates, possibly there is a huge variety of substrates in the Amazon River ecosystem. The extensive contribution of TTT system genes from unknown taxa reflects our limited understanding about it (Fig. 4). Similar to the other mentioned transporters, TTT transporters from *Alphaproteobacteria* were enriched in plume and ocean samples, while TTT transporters from *Betaproteobacteria* were enriched in freshwater or brackish sections of the river (Fig. 4).

Degradation of lignin-derived aromatic compounds

Following the initial degradation of lignin, diverse aromatic compounds are released. These can be divided into aromatic monomers (monoaryls) or dimers (diaryls), which can be processed through several biochemical steps (also called funneling pathways) until being converted into vanilate or syringate. These compounds can be processed through the ring cleavage pathways to form pyruvate or oxaloacetate, which can be incorporated to the tricarboxylic acid cycle (TCA) of cells, generating energy. All known functions taking place in the metabolism of lignin-derived aromatic compounds were found in the AMnrGC, except the gene *ligD*, a Cα-dehydrogenase for αR-isomers of β-aryl ethers. The possible degradation pathway of lignin-derived compounds in the Amazon River (Figs. 4 and **5**) included 772 and 449 genes potentially belonging to funneling pathways of diaryls and monoaryls, respectively (Fig. 4). Examination of the pathways starting with vanilate and syringate revealed 1059 genes likely to be responsible for the ring-cleavage pathway. Almost 47% of all genes related to the degradation of lignin-derived compounds in the AMnrGC belonged to 4 gene families (*ligH*, *desV*, *phcD* or *phcC*). These genes represent the main steps of intracellular lignin metabolism, which are (1) funneling pathways leading to vanilate/syringate (Fig. 4), (2.1) O-demethylation/C1 metabolism and (2.2) ring cleavage (Fig. 5).

The previously mentioned taxonomic patterns in lignin oxidation as well as in the lignin-derived compound transport were also observed in the genes potentially related to the funneling pathways (Fig. 4) including the posterior steps (Fig. 5). There was an enrichment of *Alphaproteobacteria* functions in the river sections closer to the ocean, while the number of *Actinobacteria* genes decreased in those sections (Figs. 4 and **5**). We also observed a decrease in the relative functional contribution of *Betaproteobacteria*, and an enrichment of functions from *Gammaproteobacteria* with increasing salinity (Figs. 4 and 5).

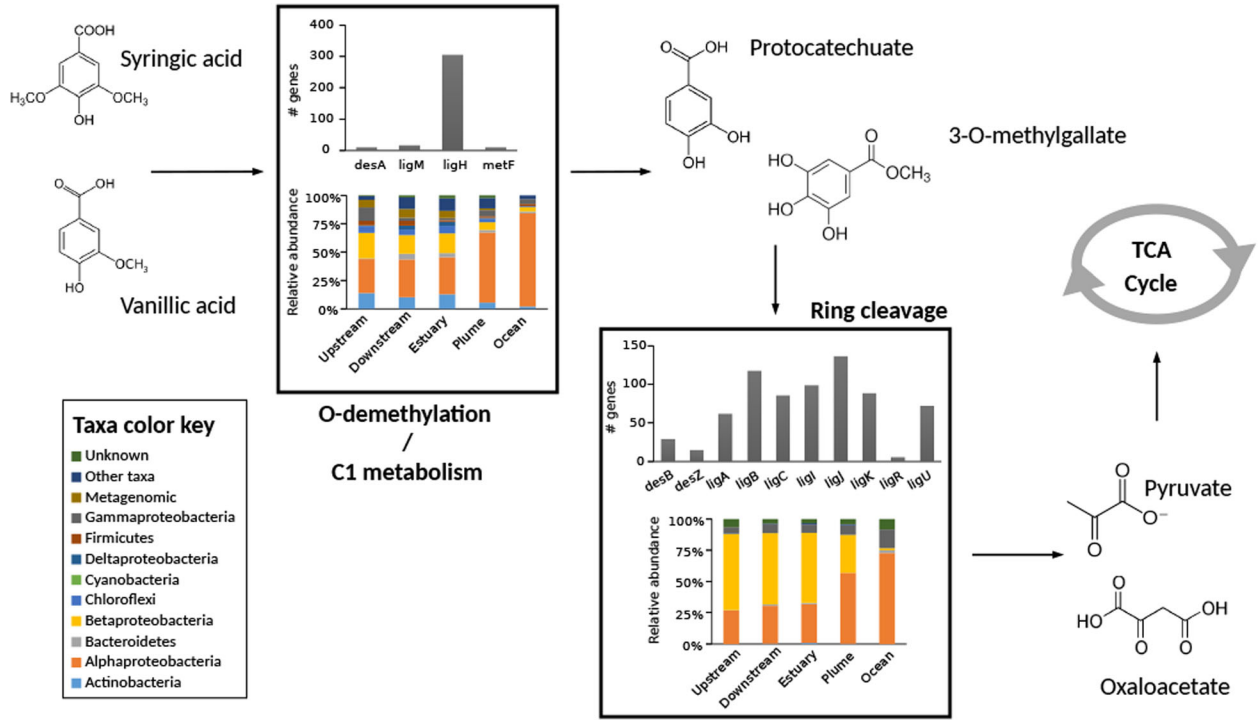

Fig. 5 Main protein families potentially involved in the last steps of the lignin-derived compound metabolism. The O-demethylation and C1 metabolism of compounds is shown in terms of protein families and taxonomy (left), as well as the ring cleavage step that directs the substrates to pyruvate, which enters into the TCA cycle to be converted into ATP and CO_2 (right)

Potential to degrade TeOM among low-rank taxonomic levels

Even though higher taxonomic levels are informative on the main distribution trends of the TeOM degradation potential along the Amazon River, lower taxonomic levels can help linking functions with the actual species or genomes carrying them. Yet, accuracy tends to decrease as genes are taxonomically annotated at lower taxonomic ranks, given that many low-rank taxa are still missing or poorly represented in reference databases. Therefore, we used in most analyses high-rank taxonomic annotations, but we also investigated main trends in the distribution of low-rank taxa that may contribute genes to TeOM degradation in the Amazon River. Specifically, we analysed genes associated to TeOM degradation that could be taxonomically assigned to genera or genomes from unknown genera present in the Genome Taxonomy Database (GTDB) [42]. Only genera or genomes contributing functions in more than half of the samples in each of the river sections as well as in the plume and ocean samples were considered.

Our results point to a limited number of widespread genera or genomes contributing with their functional machineries to the TeOM degradation in the Amazon River system, especially in saline samples from the ocean and plume (Table 1). Mainly two low-rank taxa, HIMB11 (Rhodobacteraceae) and *Pelagibacter*, contributed functions to all TeOM-degradation steps in ocean and plume samples, except for the hydrolysis of cellulose (Table 1). In turn, low-rank taxa contributing genes to TeOM degradation in the Amazon River sections were more diverse than those present in ocean and plume samples, suggesting the existence microbial consortia (Table 1). Genera such as *Ramlibacter*, *Planktophila*, *Methylopumilus*, *Limnohabitans* and *Polynucleobacter* were enriched in TeOM degradation pathways along the Amazon River (Table 1). Overall, there was a clear salinity divide in terms of the main genomes or genera carrying out TeOM degradation in the Amazon River and in the plume and ocean areas.

Spatial distributions

We evaluated whether genes potentially associated with TeOM degradation displayed spatial distribution patterns along the river course (Fig. 6; Supplementary Table 4 in Additional file 1). For this, we used the linear geographic distance of sampling sites to the Amazon River source in Peru. The linear distance to the river source was negatively correlated with the number of genes possibly associated with lignin oxidation ($R_{\text{Pearson's}} = -0.65$, p-val. $= 7.3 \times 10^{-11}$), ring cleavage pathway ($R_{\text{Pearson's}} = -0.63$, p-val. $= 1.2 \times 10^{-11}$), tripartite tricarboxylate transporting

Table 1 Low-rank taxa contributing genes to TeOM degradation in the Amazon River system

Zone	Genera or closest reference genomes from GTDB[a]	TeOM degradation step
Upstream, downstream, and estuary	*Ramlibacter*	Lignin oxidation
	Planktophila	Hemicellulose hydrolysis
	Methylopumilus, *Planktophila*, *Polynucleobacter*	Cellulose hydrolysis
	*Acidovorax*_D, *Cupriavidus*, *Curvibacter*_A, *Fonsibacter*, *Hylemonella*, *Ideonella*_A, *Limnohabitans*, PALSA-911 (Acetobacteraceae), *Polaromonas*, *Polynucleobacter*, *Ramlibacter*, *Reyranella*, SCGC-AAA027-K21 (Burkholderiaceae), UBA3064 (Burkholderiaceae), UBA6679 (Burkholderiaceae), Z2-YC6860 (Xanthobacteraceae)	TTT system
	AAA044-D11 (Nanopelagicaceae), AcAMD-5 (Nanopelagicales), GCA-2737595 (Nanopelagicaceae), *Limnohabitans*, *Nanopelagicus*, *Planktophila*, *Polynucleobacter*, RS62 (Burkholderiaceae), UBA6679 (Burkholderiaceae), UBA7398 (Nanopelagicaceae)	ABC transporters
	Planktophila, *Polynucleobacter*	FP-dimers
	Limnohabitans	FP-monomers
	GCA-2737595 (Nanopelagicaceae), *Methylopumilus*, *Planktophila*, *Fonsibacter*	O-demethylation/C1 metabolism
	*Acidovorax*_D, *Curvibacter*_A, *Limnohabitans*, *Pelomonas*	Ring cleavage
Ocean and plume	HIMB11 (Rhodobacteraceae), *Pelagibacter*	Lignin oxidation
	HIMB11 (Rhodobacteraceae)	Hemicellulose hydrolysis
	D2472 (Gammaproteobacteria), UBA4465 (Cyclobacteriaceae)	Cellulose hydrolysis
	HIMB11 (Rhodobacteraceae), HIMB59 (Alphaproteobacteria), *Pelagibacter*	TTT system
	Pelagibacter, *Pelagibacter*_A, TMED189 (Acidimicrobiia)	ABC transporters
	HIMB11 (Rhodobacteraceae), *Pelagibacter*	FP-dimers
	HIMB11 (Rhodobacteraceae), *Pelagibacter*	FP-monomers
	HIMB11 (Rhodobacteraceae), *Pelagibacter*, SCGC-AAA076-P13 (Gammaproteobacteria)	O-demethylation
	N/A	Ring cleavage

Main prokaryotic genera, or genomes from the Genome Taxonomy Database (GTDB) without assigned genera, contributing genes to TeOM degradation in the Amazon River sections as well as in plume and ocean samples are indicated. Only taxa contributing functions in more than half of the samples of each studied zone are reported
[a]Genera or GTDB reference-genome names are indicated. For reference genome names, the lowest taxonomic level indicated in GTDB is shown in brackets
FP funneling pathways, *TTT* tripartite tricarboxylic transporter, *N/A* not applicable

($R_{\text{Pearson's}} = -0.57$, p-val. = 5.4×10^{-10}) and the AAHS transporters ($R_{\text{Pearson's}} = -0.35$, p-val. = 4.5×10^{-6}) (Fig. 6). This is coherent with a putative trend displayed by gene-function distributions along the river, pointing to lignin oxidation-related functions being replaced by cellulose degradation counterparts in brackish waters. A potential reduction of the microbial gene repertoire related to lignin processing as the river approaches the ocean suggests the ageing of TeOM during its flow through the Amazon River.

AAHS transporters were negatively correlated to the distance to the Amazon River source ($R_{\text{Pearson's}} = -0.35$, p-val. = 4.5×10^{-6}), while ABC transporters were not correlated with this distance ($p > 0.01$) (Fig. 6). Furthermore, AAHS and ABC transporters showed positive correlations to the funneling pathway of diaryls ($R_{\text{Pearson's}} = 0.39$, p-val. = 2.5×10^{-3}) and monoaryls ($R_{\text{Pearson's}} = 0.33$, p-val. = 9.6×10^{-3}), respectively. This suggests specificity in the transport of lignin-derived molecules by those transporter families. Furthermore, AAHS ($R_{\text{Pearson's}} = 0.38$, p-val. = 2.2×10^{-5}) and ABC ($R_{\text{Pearson's}} = 0.54$, p-val. = 6.7×10^{-5}) transporters were positively correlated to the ring cleavage pathway, suggesting that ABC and AAHS transporters are relevant for the metabolism of lignin-derived compounds.

The number of genes in the TTT system displayed a negative correlation with the distance to the Amazon River source ($R_{\text{Pearson's}} = -0.57$, p-val. = 5.4×10^{-10}), suggesting their predominance in freshwater sections of the river (Fig. 6). TTT transporters showed a positive correlation with lignin oxidation genes ($R_{\text{Pearson's}} = 0.67$, p-val. = 6.7×10^{-14}), suggesting they could be transporting lignin-derived products or a TTT coupling with the machinery to oxidize lignin. The TTT system was positively correlated to AAHS ($R_{\text{Pearson's}} = 0.41$, p-val. = 3.5×10^{-5}) and ABC ($R_{\text{Pearson's}} = 0.46$, p-val. = 2×10^{-4}) transporters (Fig. 6) pointing to a possible functional complementarity, as the TTT would transport substrates not transported by the other transporter families.

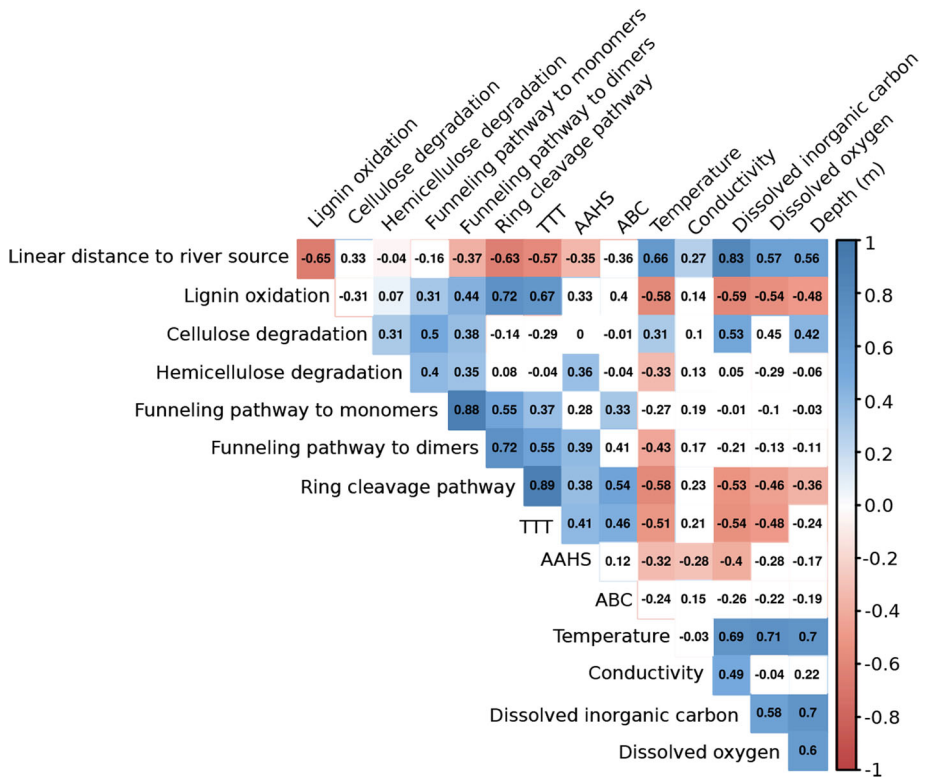

Fig. 6 Correlations among genes associated with the processing of TeOM and their correlation to environmental variables. Correlations between the number of genes associated with lignin oxidation, cellulose and hemicellulose deconstruction, transporting systems (AAHS, ABC and TTT), lignin-derived aromatic compounds processing pathways (Ring cleavage pathways; Funneling pathways of dimers and monomers), and environmental variables (dissolved inorganic carbon—DIC, dissolved oxygen—DO, temperature, conductivity, sample depth—Depth and linear distance from the sampling site to the Amazon River source). Correlation coefficients are shown inside the boxes, and their color indicates the correlation strength. White boxes are non-significant correlations ($p > 0.01$)

The gene machinery associated with the processing of lignin-derived aromatic compounds was positively correlated to the machinery related to lignin oxidation along the river course (Fig. 6), suggesting a co-processing of lignin and its byproducts. In terms of genes, cellulose degradation was not correlated with lignin oxidation ($p > 0.01$), but had a modest positive correlation to hemicellulose degradation ($R_{\text{Pearson's}} = 0.31$, p-val. $= 5.4 \times 10^{-3}$) (Fig. 6), suggesting a coupling between both pathways.

We found correlations between both genes associated with the funneling pathway of dimers ($R_{\text{Pearson's}} = -0.37$, p-val. $= 8 \times 10^{-5}$) and the ring cleavage pathway ($R_{\text{Pearson's}} = -0.63$, p-val. $= 1.2 \times 10^{-11}$), with the distance to the Amazon River source (Fig. 6). This indicates that the degradation of lignin-derived aromatic compounds may follow a similar pattern as the lignin oxidation machinery, being predominantly restricted to upstream sections of the river. Moreover, the number of genes potentially related to hemi-/cellulose degradation was positively correlated to those possibly related to funneling pathways of lignin-derived monomers and dimers. This could reflect a potential co-metabolism of lignin-derived compounds and hemi-/cellulose degradation, instead of lignin oxidation.

The effect of environmental variables in the potential TeOM degradation machinery

The potential processes related to TeOM degradation also seem to be correlated to specific environmental variables (Fig. 6; Supplementary Table 4 in Additional file 1). The machinery related to the oxidation of lignin is potentially more abundant in river sections with lower temperatures ($R_{\text{Pearson's}} = -0.58$, p-val. $= 2 \times 10^{-4}$), lower dissolved inorganic carbon (DIC) ($R_{\text{Pearson's}} = -0.59$, p-val. $= 1 \times 10^{-9}$) and oxygen (DO) ($R_{\text{Pearson's}} = -0.54$, p-val. $= 2 \times 10^{-4}$), and at smaller depths ($R_{\text{Pearson's}} = -0.48$, p-val. $= 1 \times 10^{-3}$). Similarly, the hemicellulose degradation machinery was negatively correlated with temperature ($R_{\text{Pearson's}} = -0.33$, p-val. $= 9 \times 10^{-4}$). In contrast to those mentioned above, the cellulose degradation arsenal was positively correlated to higher temperatures ($R_{\text{Pearson's}} = 0.31$, p-val. $= 8 \times 10^{-4}$), DIC ($R_{\text{Pearson's}} = 0.53$, p-val. $= 1 \times 10^{-7}$) and sampling depth ($R_{\text{Pearson's}} = 0.42$, p-val. $= 5 \times 10^{-5}$).

The transporters of lignin-derived molecules were marginally correlated to the measured environmental variables (Fig. 6). Specifically, ABC transporters did not correlate to any variable; however, AAHS transporters were

negatively correlated to temperature ($R_{Pearson's}$ = − 0.32, p-val. = 6 × 10^{-3}), DIC ($R_{Pearson's}$ = − 0.40, p-val. = 1 × 10^{-4}) and conductivity ($R_{Pearson's}$ = − 0.28, p-val. = 3 × 10^{-3}). TTT transporters seemed to follow a similar trend as lignin oxidation in terms of temperature ($R_{Pearson's}$ = − 0.51, p-val. = 7 × 10^{-7}), DIC ($R_{Pearson's}$ = − 0.53, p-val. = 8 × 10^{-8}) and DO ($R_{Pearson's}$ = − 0.48, p-val. = 2 × 10^{-5}).

The initial steps of the metabolism of lignin-derived molecules did not seem to be correlated to environmental heterogeneity (Fig. 6), except for the functions related to the funneling pathways of dimers that were correlated to temperature ($R_{Pearson's}$ = − 0.43, p-val. = 5 × 10^{-6}). The last steps of the processing of lignin-derived molecules, the ring cleavage, were strongly correlated to environmental heterogeneity. These final steps resembled patterns observed in the lignin oxidation machinery in terms of temperature ($R_{Pearson's}$ = − 0.58, p-val. = 4 × 10^{-8}), DIC ($R_{Pearson's}$ = − 0.53, p-val. = 1 × 10^{-7}), DO ($R_{Pearson's}$ = − 0.46, p-val. = 4 × 10^{-5}) and sampling depth ($R_{Pearson's}$ = − 0.36, p-val. = 4 × 10^{-3}).

Discussion

The AMnrGC significantly expands the comprehension of the metabolic potential of the world's largest river microbiome and is publicly available (https://doi.org/10.5281/zenodo.1484504). The predicted ~ 3.7 M genes are a valuable resource for understanding the functioning of this ecosystem as well as for bioprospecting. Almost half of the genes in the catalogue had no close orthologs, suggesting gene novelty. Yet, this extensive portion of unknown genes (48%) is similar to other environmental microbiomes that featured 40–60% of unknown genes [6–8]. Interestingly, the analysis of k-mers indicated a distinct composition, in terms of genomic information, of the Amazon River microbiome when compared to other rivers and to the Amazon rainforest soil, being coherent with the novel diversity previously found in Brazilian freshwater systems [43]. Altogether, this points to gene novelty and a compositional distinctiveness of the Amazon River microbiome.

Analyses of COG functions pointed to a number of core functions along the Amazon River course, which was supported by the similar distribution of COG superclasses along the different river sections (Fig. 2d). In particular, COG functions within the superclass "Metabolism" were the most abundant in the AMnrGC, as well as in the upper Mississippi River [44]. Core functions included a general carbohydrate metabolism and several transporter systems, mainly ABC transporters. This suggests a sophisticated machinery to process TeOM in the Amazon River, with core metabolisms indicating a general organic matter degradation system, ending in acetogenic pathways.

Lignin-derived aromatic compounds need to be transported from the extracellular milieu to the cytoplasm to be degraded, and different transporting systems can be involved in this process [36, 37, 45, 46]. In particular, previous studies showed that the TTT system was present in high quantities in the Amazon River, and this was attributed to a potential degradation of allochthonous organic matter [14]. Recent findings also suggest a TTT system related to the transport of TeOM degradation byproducts [47, 48]. Little is known about these transporters, but our findings indicate that TTT is an abundant protein family in the Amazon River, suggesting that tricarboxylates are a common carbon source for prokaryotes in these waters. Our results also indicated that the TTT transporters could be linked to the genes potentially related to lignin oxidation, supporting the role of TTT in TeOM degradation.

The taxa found to be potentially involved in TeOM degradation mostly belong to *Proteobacteria* (Table 1), especially *Betaproteobacteria* (such as the genera *Polynucleobacter*, *Methylopumilus* and *Limnohabitans*) and *Alphaproteobacteria* (e.g. HIMB11 and *Candidatus* Pelagibacter). Other important groups include *Actinobacteria* (represented by the genus *Candidatus* Planktophila) and *Bacteroidetes*, all regular freshwater taxa [49]. The participation of *Bacteroidetes* in hemi-/cellulose degradation has been previously reported, being metabolically capable to degrade recalcitrant organic compounds such as humic substances [49, 50]. In the Amazon River, there was an increase of *Gammaproteobacteria* and *Alphaproteobacteria* genes, possibly involved in TeOM degradation, in the river sections closer to the ocean. In turn, there was an increase in the number of TeOM degradation genes from *Actinobacteria* and *Betaproteobacteria* in freshwater sections of the Amazon River. This suggests that salinity shapes the composition of microbial communities in the different sections of the Amazon River and consequently affects their ecology, agreeing with the salinity boundary hypothesis introduced by Logares et al. [51]. The distribution of bacterial taxa along the Amazon River may also reflect taxon-specific preferences for TeOM quality, as previous studies have reported a differential preference of bacteria for fresh vs. aged TeOM [52]. For example, after long incubation experiments, *Actinobacteria*, *Bacteroidetes* and *Betaproteobacteria* showed a preference for fresh TeOM while *Alphaproteobacteria* and *Gammaproteobacteria* displayed a preference for aged organic matter. Furthermore, Sipler et al. [53] found that Arctic coastal *Alphaproteobacteria*, *Bacteroidetes*, *Betaproteobacteria* and *Actinobacteria* were negatively affected by the addition of TeOM in bottle experiments. This suggests that most TeOM degradation occurs in rivers and that coastal microbiomes are less capable to degrade these compounds, being coherent with our results.

Correlations between measured environmental variables and the potential TeOM degradation machinery indicated

that lignin oxidation may happen in regions with low concentration of oxygen and dissolved inorganic carbon, and low temperature. This is coherent with previous reports [38–41]. However, the machinery related to the degradation of lignin-derived compounds seems to be independent of environmental conditions, indicating a potentially ubiquitous degradation of such compounds in the Amazon River. The correlations between environmental variables and the potential degradation of cellulose and hemicellulose still suggest that cellulose and lignin would be degraded in different river sections. Therefore, both processes seem to be decoupled in the Amazon River.

Our results agree with previous experimental evidence of TeOM ageing in the Amazon River [27, 28], which supported a *priming effect* in incubation experiments with recalcitrant and labile organic matter [20]. However, the mechanism behind this *priming effect* remained unexplained. Based on our findings, we hypothesized a model for the potential *priming effect* acting in lignocellulose complexes in the Amazon River (Fig. 7). In this model, there are two different communities co-existing in a consortium: one possibly responsible for hemi-/cellulose degradation and another one likely involved in lignin degradation. The first community would release extracellular enzymes (mainly glycosyl hydrolases from families GH3 and GH10), producing different kinds of carbohydrates. These carbohydrates may provide structural carbon and energy for the entire consortium. The potential lignolytic community would also use the cellulolytic byproducts to grow, which promotes an oxidative metabolism. This oxidative metabolism could trigger the production and secretion of reactive oxygen species (ROS) (Fig. 7). ROS are then used by DyPs and laccases secreted by these putative lignolytic communities to oxidize lignin, exposing more hemi-/cellulose to cellulolytic communities, re-starting the cycle (Fig. 7). Another important role of lignolytic communities could be the degradation of lignin-derived aromatic compounds generated by the lignin oxidation process. Those compounds, if not degraded, can inhibit cellulolytic enzymes and microbial growth [54–57], preventing TeOM degradation. This cycle may be considered as a *priming effect*, where both communities benefit from each other.

Conclusions

The Amazon River is a major carbon link between terrestrial, atmospheric and marine ecosystems. Our work represents a first effort to link the TeOM inputs into the Amazon River with the microbial metabolisms potentially responsible for their degradation. We identified genes and metabolisms that are likely key in TeOM degradation. Furthermore, our results indicate differential distributions of TeOM-related genes that in some cases seem to be driven by environmental heterogeneity. Our work also generated the AMnrGC, an important resource for interrogating the functionality of the Amazon River microbiome as well as for bioprospecting. Given the extent and difficulty of access to many regions of the Amazon River basin, our work is an important first step that could pave the road for future ambitious sampling campaigns that will investigate gene expression, metaproteomics and the capacity of the Amazon River

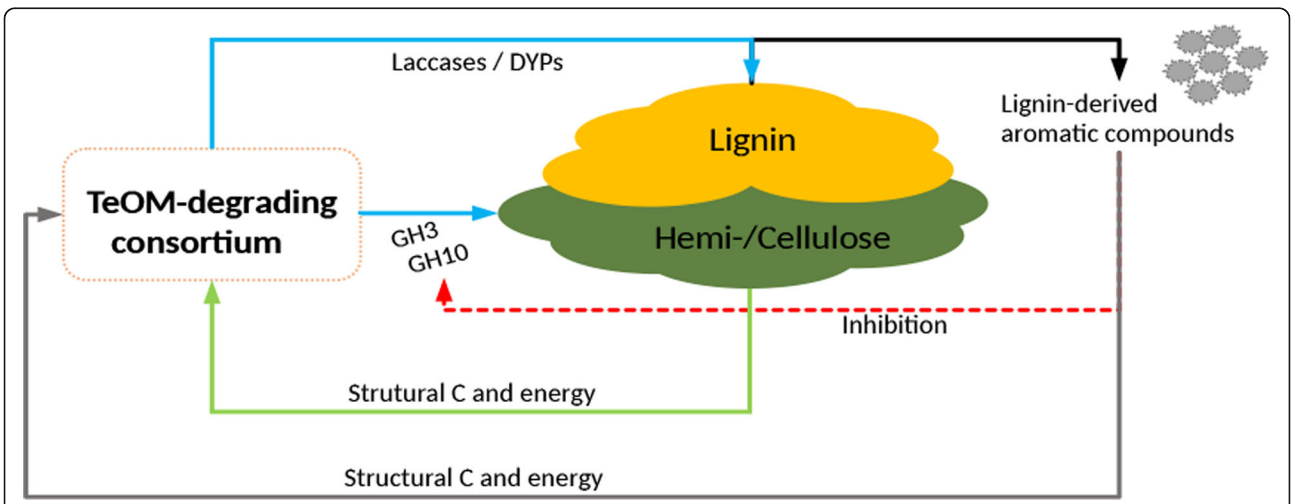

Fig. 7 *Priming effect* model of microbial TeOM degradation in the Amazon River. The cellulolytic communities degrade hemi-/cellulose through secretion of glycosyl hydrolases (mainly GH3/GH10), which release sugars to the environment. These sugars can promote growth of the cellulolytic and lignolytic communities, and during this process, the oxidative metabolism produces reactive oxygen species (ROS). ROS activate the exoenzymes (mainly DYPs and laccases) secreted by the lignolytic community to oxidize lignin. After lignin oxidation, the hemi-/cellulose becomes exposed again, helping the cellulolytic communities to degrade it. During the previous process, several aromatic compounds are formed, which can potentially inhibit cellulolytic enzymes and microbial growth. However, these compounds are consumed by lignolytic microorganisms, reducing their concentration in the environment allowing decomposition to proceed

microbiota to degrade TeOM in field or laboratory experiments.

Materials and methods

We analysed 106 metagenomes [12, 13, 15, 18] from 30 sampling points distributed along the Amazon river basin, with an average coverage of 5.3×10^9 ($\pm 7.4 \times 10^9$) base pairs per metagenome (Supplementary Table 1 in Additional file 1). The sampling points from the Solimões River and lakes in the Amazon River course, located upstream from the city of Manaus, until the Amazon River's plume in the Atlantic Ocean covered ~ 2106 km and were divided into 5 sections (Fig. 1a and Supplementary Table 1 in Additional file 1). These sections were (1) *upstream section* (upstream Manaus city), (2) *downstream section* (placed between Manaus and the start of the Amazon River estuary. It includes the influx of particle-rich white waters from the Solimões River as well as the influx of humic waters from Negro River [58, 59]), (3) *estuary section* (part of the river that meets the Atlantic Ocean), (4) *plume section* (the area where the Ocean is influenced by the Amazon River inputs) and (5) *ocean* (the area with higher salinity surrounding the Plume).

Samples were taken as previously indicated [12, 13, 15, 18]. Depending on the original study, particle-associated microbes were defined as those passing the filter of 300 μm mesh size and being retained in the filter of 2–5 μm mesh size. Free-living microbes were defined as those passing the filter of 2–5 μm mesh size, being retained in the filter of 0.2 μm mesh size. DNA was extracted from the filters as indicated in the original studies [12, 13, 15, 18]. Metagenomes were obtained from libraries prepared with either Nextera or TruSeq kits. Different *Illumina* sequencing platforms were used: Genome Analyzer IIx, HiSeq 2500 or MiSeq. Additional information is provided in Supplementary Table 1 in Additional file 1.

Metagenome analysis

Illumina adapters and poor-quality bases were removed from metagenomes using Cutadapt [60]. Only reads longer than 80 bp, containing bases with Q ≥ 24, were kept. The quality of the reads was checked with FASTQC [61]. Reads from metagenomes belonging to the same sampling points were assembled together using MEGA HIT (v1.0) [62], with the meta-large presets. Only contigs > 1 kbp were considered, as recommended by previous work [63]. Assembly quality was assessed with QUAST [64]. Metagenome assembly yielded 2,747,383 contigs ≥ 1000 base pairs, in a total assembly length of ~ 5.5×10^9 bp with an average N50 of 2064 ± 377 bp (see Supplementary Table 2 in Additional file 1).

Analysis of k-mer diversity over different river zones

A k-mer diversity analysis was used to compare the genetic information of the Amazon River microbiome against that in other microbiomes from Amazon rainforest soil and temperate rivers (Supplementary Table 3 in Additional file 1). Specifically, the Amazon River metagenomes (106) were compared against 37 metagenomes from the Mississippi River [65], 91 metagenomes from three watersheds in Canada [66] and 7 metagenomes from the Amazon forest soil [67]. The rationale to include soil metagenomes was to check whether genomic information in the river could be derived from soil microbiota. K-mer comparisons were run with SIMKA (version 1.4) [68] normalizing by sample size. Low complexity reads and k-mers (Shannon index < 1.5) were discarded before SIMKA analyses. The resulting Jaccard's distance matrix was used to generate a non-metric multidimensional scaling (NMDS) analysis. Permutation tests were used to check the homogeneity of β dispersion in the groups, and permutational multivariate analysis of variance (PERMANOVA/ANOSIM) was used to test the groups' difference. Both analyses were performed using the R package Vegan [69].

Amazon River basin Microbial non-redundant Gene Catalogue (AMnrGC)

Genes were predicted using Prodigal (version 2.6.3) [70]. Only open reading frames (ORFs) predicted as complete, accepting alternative initiation codons, and longer than 150 bp, were considered in downstream analyses. Gene sequences were clustered into a non-redundant gene catalogue using CD-HIT-EST (version 4.6) [71, 72] at 95% of nucleotide identity and 90% of overlap of the shorter gene [5]. Representative gene sequences were used in downstream analyses. GC content per gene was inferred via Infoseq, EMBOSS package (version 6.6.0.0) [73].

Gene abundance estimation

The quality-checked sequencing reads were backmapped against our non-redundant gene catalogue using BWA (version 0.7.12-r1039) [74] and SamTools (version 1.3.1) [75]. Gene abundances were estimated using the software eXpress (version 1.5.1) [76], with no bias correction, as counts per million (CPM). We used a CPM ≥ 1.00 for a gene to be present in a sample, and an average abundance higher than zero ($\mu_{CPM} > 0.0$) for a gene to be present in a river section or water type (i.e. freshwater, brackish water or the mix of them in the plume).

Functional annotation

Representative genes (and their predicted amino acid sequences) were annotated by searching them against KEGG (Release 2015-10-12) [77], COG (Release 2014) [78], CAMERA Prokaryotic Proteins Database (Release

2014) [79] and UniProtKB (Release 2016-08) [80] via the Blastp algorithm implemented in Diamond (v.0.9.22) [81], with a query coverage ≥ 50%, identity ≥ 45%, e-value ≤ $1e^{-5}$ and score ≥ 50. Hits were parsed by score, e-value and identity until the best result was found. KO-pathway mapping was performed using KEGG mapper [82]. HMMSearch (version 3.1b1) [83] was used to search proteins against dbCAN (version 5) [84], PFAM (version 30) [85] and eggNOG (version 4.5) [86] databases, using an e-value ≤ $1e^{-5}$, and posterior probability of aligned residues ≥ 0.9, and no domain overlapping. Accumulation curves were obtained using random progressive nested comparisons with 100 pseudo-replicates for genes and PFAM predictions.

Core metabolisms

We adopted the definition of core metabolic functions as those involved in cell or ecosystem homeostasis, representing the minimal metabolic machinery needed to survive in a given environment. Similar to other works [6, 7], we used the annotations with KEGG and PFAM databases to determine the bacterial functional core. By using gene abundances as CPM as a criterion for counting functions in each sample or river section, we analysed metabolic pathways. Those functions present in at least 80% of the samples were considered as core. KEGG Mapper [87] and MinPath [88] were used to organize the information underlying core functions.

Gene taxonomy

Given that a high number of low-rank taxa are missing or poorly represented in reference databases, taxonomic annotation accuracy tends to decrease as genes are taxonomically annotated at lower taxonomic ranks. For this reason, we used two different approaches to taxonomically annotate genes. *Approach 1* is more conservative, aiming to annotate genes at higher taxonomic ranks (e.g. Class) and therefore being potentially more accurate than the less conservative *Approach 2*, which aims at annotating genes at lower taxonomic ranks (e.g. Genus). The specific methods associated to each approach are indicated below:

- *Approach 1:* High-rank gene taxonomy was assigned considering the best hits (score, e-value and identity; see above) using KEGG (Release 2015-10-12) [77], UniProtKB (Release 2016-08) [80] and CAMERA Prokaryotic Proteins Database (Release 2014) [79]. Taxonomic last common ancestors (LCA) were determined from TaxIDs (NCBI) associated with UniRef100 and KO entries. Information from the CAMERA database was also used to retrieve taxonomy (NCBI TaxID). Taxonomy was assigned using the best hit, of a given protein, obtained across databases. Proteins were annotated as "unassigned" if their taxonomic signatures were mixed, containing representatives from several domains of life, or if they had the function assigned without taxonomic information. Reference sequences with hits to poorly annotated sequences from other metagenomes were referred to as "Metagenomic".
- *Approach 2:* Low-rank taxonomic affiliation was determined using MMseqs2 version 11-e1a1c [89] using default settings, based on the Genome Taxonomy Database (GTDB; publicly available in https://gtdb.ecogenomic.org/) [42].

Potential TeOM degradation machinery

To investigate the potential TeOM degradation, we grouped samples by river section and assessed their gene content. Genes were then searched against reference sequences and protein families known to be involved in TeOM degradation (see Supplementary Table 5 in Additional file 1). In particular, bacterial lignin degradation starts with extracellular polymer oxidation followed by monomers and dimers moving across membranes into the cytoplasm for their ultimate degradation. Protein families related to lignin oxidation (PF05870, PF07250, PF11895, PF04261 and PF02578) were searched among PFAM-annotated genes. The genes related to the metabolism of lignin-derived aromatic compounds were annotated with Diamond (Blastp search mode; v.0.9.22) [81], with query coverage ≥ 50%, protein identity ≥ 40% and e-value ≤ $1e^{-5}$ as recommended by Kamimura et al. [45], using their dataset as reference.

Cellulose and hemicellulose degradation involve glycosyl hydrolases (GH). The most common cellulolytic protein families (GH1, GH3, GH5, GH6, GH8, GH9, GH12, GH45, GH48, GH51 and GH74) [90] and cellulose-binding motifs (CBM1, CBM2, CBM3, CBM6, CBM8, CBM30 and CBM44) [90, 91] were searched in PFAM/dbCAN annotations. In addition, the most common hemicellulolytic families (GH2, GH10, GH11, GH16, GH26, GH30, GH31, GH39, GH42, GH43 and GH53) [91] were searched in the PFAM/dbCAN database. Lytic polysaccharide monooxygenases (LPMO) [91] were also identified using PFAM to investigate the simultaneous deconstruction of cellulose and hemicellulose.

During the degradation of refractory and labile material by exoenzymes, microbes produce a complex mix of particulate and dissolved organic carbon. The use of this mix is mediated by a vast variety of transporter systems [46]. The typical transporters associated with lignin degradation (AAHS family, ABC transporters, MHS family, ITS superfamily and TRAP transporter) were searched with Diamond (v.0.9.22) [81], using query coverage ≥ 50%, protein identity ≥ 40% and e-value ≤ $1e^{-5}$ and a reference dataset previously compiled [45].

Similarly to the fate of hemi-/cellulose degradation byproducts, lignin degradation ends up in the production

of 4-carboxy-4hydroxy-2-oxoadipate, which is converted into pyruvate or oxaloacetate, both substrates of the tricarboxylic acid cycle (TCA) [45]. Recently, several substrate binding proteins (TctC) belonging to the tripartite tricarboxylate transporter (TTT) system were associated with the transport of TeOM degradation byproducts, like adipate [47] and terephthalate [48]. To investigate the metabolism of these compounds, and the possible link between the TTT system and lignin/cellulose degradation, the protein families TctA (PF01970), TctB (PF07331) and TctC (PF03401) were searched in PFAM.

The genes found using the abovementioned strategy were submitted to PSORT v.3.0 [92], to determine the protein subcellular localization (cytoplasm, secreted to the outside, inner membrane, periplasm or outer membrane). We carried out predictions in the three possible taxa (Gram negative, Gram positive and Archaea), and the best score was used to determine the subcellular localization. Genes assigned to an "unknown" location, as well as those with a wrong assignment, were eliminated (for example, genes known to work in extracellular space that were assigned to the cytoplasmic membrane).

The total amount of TeOM degradation genes found per function (lignin oxidation, transport, hemi-/cellulose degradation and lignin-derived aromatic compounds metabolism) in each section of the river were normalized by the maximum gene counts per metagenome. Subsequently, correlograms were produced adding the environmental variables and using Pearson's correlation coefficients calculated with complete pairwise observations using the R packages Corrplot [93] and RColorBrewer [94]. The linear geographic distance of each metagenome to the Amazon River source (i.e. Mantaro River, Peru, 10° 43′ 55″ S / 76° 38′ 52″ W) was also used in this analysis to infer changes in gene counts along the Amazon River course. The sampling site distance to the Amazon River source in Peru was calculated with the R package Fields [95].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40168-020-00930-w.

Additional file 1: Supplementary Tables. (1) Description of the 106 metagenomes used to build the Amazon River basin Microbial non-redundant Gene Catalogue (AMnrGC). The Amazon River basin section shows the group that a sample belongs to according to its geographic location. Other features were obtained from the original publications and SRA codes. "N.A." stands for not available. (2) Co-assembly groups used to build the Amazon River basin Microbial non-redundant Gene Catalogue (AMnrGC). (3) Metagenomes used for K-mer diversity assessment. (4) Correlation and significance between gene content and environmental variables. Pearson's correlation coefficients are shown under the diagonal and correspondent p-values are shown in red above the diagonal. The correlations were calculated using complete pairwise observations. (5) Reference proteins and protein families involved in terrestrial organic matter degradation used to annotate proteins related to lignin oxidation, cellulose and hemicellulose degradation in the AMnrGC.

Abbreviations

AAHS: Aromatic Acid:H^+ Symporter; ABC: ATP-binding cassette transporters; AMnrGC: Amazon River basin Microbial non-redundant Gene Catalogue; C1: One carbon compounds (e.g. methane); COG: Clusters of Orthologous Genes; CPM: Counts per million; DIC: Dissolved inorganic carbon; DO: Dissolved oxygen; DyPs: Dye-decolorizing peroxidases; FP: Funneling pathway; GHn: Glycosyl hydrolase family "n"; KEGG: Kyoto Encyclopedia of Genes and Genomes; LCA: Taxonomic last common ancestors; LPMO: Lytic polysaccharide monooxygenases; MFS: Major facilitator superfamily of transporters; PFAM: Protein family; Pg C: Peta(10^{15}) grams of carbon; p-val.: *p* value; ROS: Reactive oxygen species; TCA: Tricarboxylic acid cycle; Tct: Tripartite transporter component (A, B or C); TeOM: Terrestrial organic matter; TTT: Tripartite tricarboxylate transporter/transporting system

Acknowledgements

Bioinformatics analyses were performed at the MARBITS platform of the Institut de Ciències del Mar (ICM; http://marbits.icm.csic.es) as well as in MareNostrum (Barcelona Supercomputing Center) via grants obtained from the Spanish Network of Supercomputing (RES) to RL. We thank Pablo Sánchez and Lidia Montiel Fontanet for helping with bioinformatics analyses. We also thank the EMM group (https://emm.icm.csic.es) at the ICM-CSIC for all the support and cordiality during the development of part of this work as well as other members of the log-lab (http://www.log-lab.barcelona). We thank the CSIC Open Access Publication Support Initiative through the Unit of Information Resources for Research (URICI) for helping to cover publication fees.

Authors' contributions

CDSJ, FHS and RL designed the study. CDSJ compiled and curated the data and performed bioinformatic analysis. CDSJ, FHS, HS and RL interpreted the results. FHS, RL, FPM and HS supervised and administered the project and provided funding. The original draft was written by CDSJ. All co-authors contributed substantially to manuscript revisions. All author(s) read and approved the final manuscript.

Funding

CDSJ was supported by a PhD scholarship from Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil (CNPq #141112/2016-6). FHS and HS work was supported by Research Productivity grants from CNPq (Process # 311746/2017-9 and #309514/2017-7, respectively). RL was supported by a Ramón y Cajal fellowship (RYC-2013-12554, MINECO, Spain). This work was supported by Petróleo Brasileiro S.A. (Petrobras), as part of a research agreement (#0050.0081178.13.9) with the Federal University of São Carlos, SP, Brazil, within the context of the Geochemistry Thematic Network. Additionally, this work was supported by the projects INTERACTOMICS (CTM2015-69936-P, MINECO, Spain) and MicroEcoSystems (240904, RCN, Norway) to RL and Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP (Process #2014/14139-3) to HS. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 (CAPES #88881.131637/2016-01).

Availability of data and materials

Metagenomes used to construct the Amazon River gene catalogue (AMnrGC) are publicly available (See Supplementary Table 1 in Additional file 1) from the following SRA projects: SRP044326, PRJEB25171 and SRP039390). Publicly available metagenomes used in the k-mer diversity comparison are detailed in Supplementary Table 3 (Additional file 1) and publicly available from the following SRA projects: Amazon forest [PRJNA336764, PRJNA336766, PRJNA337825, PRJNA336700, PRJNA336765], Mississippi River [SRP018728] and Canada watersheds [PRJNA287840]). The AMnrGC and all the associated files are available in a permanent Zenodo repository (https://doi.org/10.5281/zenodo.1484504).

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests
Fernando Pellon de Miranda is employed by Petroleo Brasileiro S.A - Petrobras, Brasil.

Author details
[1]Molecular Biology Laboratory, Department of Genetics and Evolution – DGE, Universidade Federal de São Carlos – UFSCar, Rod. Washington Luis KM 235 - Monjolinho, São Carlos, SP 13565-905, Brazil. [2]Institute of Science and Technology for Brain-Inspired Intelligence – ISTBI, Fudan University, Handan Rd 220, Wu Jiao Chang, Yangpu, Shanghai 200433, China. [3]Laboratory of Microbial Processes & Biodiversity, Department of Hydrobiology – DHB, Universidade Federal de São Carlos – UFSCar, Via Washington Luis KM 235 - Monjolinho, São Carlos, SP 13565-905, Brazil. [4]Centro de Pesquisas e Desenvolvimento Leopoldo Américo Miguez de Mello, Petróleo Brasileiro S.A. (Petrobras), Av. Horácio Macedo 950, Rio de Janeiro, RJ 21941-915, Brazil. [5]Institute of Marine Sciences (ICM), CSIC, Passeig Marítim de la Barceloneta 37-49, ES08003, Barcelona, Catalonia, Spain.

Received: 29 July 2020 Accepted: 6 October 2020
Published online: 30 October 2020

References

1. Cole JJ, Prairie YT, Caraco NF, McDowell WH, Tranvik LJ, Striegl RG, et al. Plumbing the global carbon cycle: integrating inland waters into the terrestrial carbon budget. Ecosystems. 2007;10:172–85..
2. Xenopoulos MA, Downing JA, Kumar MD, Menden-Deuer S, Voss M. Headwaters to oceans: ecological and biogeochemical contrasts across the aquatic continuum: headwaters to oceans. Limnol Oceanogr. 2017; 62:S3–S14.
3. Guenet B, Danger M, Abbadie L, Lacroix G. Primming effect: bridging the gap between terrestrial and aquatic ecology. Ecology. 2010;91:2850–61.
4. Bianchi TS. The role of terrestrially derived organic carbon in the coastal ocean: a changing paradigm and the priming effect. Proc Natl Acad Sci U S A. 2011;108:19473–81.
5. Mende DR, Bryant JA, Aylward FO, Eppley JM, Nielsen T, Karl DM, et al. Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nat Microbiol. 2017;2:1367–73.
6. Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, et al. A global ocean atlas of eukaryotic genes. Nat Commun. 2018;9:373.
7. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015;348: 1261359.
8. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, et al. Structure and function of the global topsoil microbiome. Nature. 2018;560:233–7.
9. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464:59–65.
10. Pan H, Guo R, Zhu J, Wang Q, Ju Y, Xie Y, et al. A gene catalogue of the Sprague-Dawley rat gut metagenome. GigaScience. 2018;7.
11. Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh H-J, Cuenca M, et al. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. Cell. 2019;179:1068-1083.e21.
12. Santos-Júnior CD, Kishi LT, Toyama D, Soares-Costa A, Oliveira TCS, de Miranda FP, et al. Metagenome sequencing of prokaryotic microbiota collected from rivers in the upper Amazon basin. Genome Announc. 2017;5: e01450–16.
13. Toyama D, Kishi LT, Santos-Júnior CD, Soares-Costa A, Souza De Oliveira TC, Pellon De Miranda F, et al. Metagenomics analysis of microorganisms in freshwater lakes of the Amazon basin. Genome Announc. 2016;4:1440–16.
14. Ghai R, Rodriguez-Valera F, McMahon KD, Toyama D, Rinke R, de Oliveira TCS, et al. Metagenomics of the water column in the pristine upper course of the Amazon river. Lopez-Garcia P, editor. PLoS ONE. 2011;6:e23785.
15. Santos-Junior CD, Toyama D, TCS O, Pellon De Miranda F, Henrique-Silva F. Flood season microbiota from the Amazon basin lakes: analysis with metagenome sequencing. Microbiol Resour Announc. 2019;8:e00229–19.
16. Satinsky BM, Smith CB, Sharma S, Ward ND, Krusche AV, Richey JE, et al. Patterns of bacterial and archaeal gene expression through the lower Amazon river. Front Mar Sci. 2017;4:253.
17. Satinsky BM, Smith CB, Sharma S, Landa M, Medeiros PM, Coles VJ, et al. Expression patterns of elemental cycling genes in the Amazon River plume. ISME J. 2017;11:1852–64.
18. Satinsky BM, Zielinski BL, Doherty M, Smith CB, Sharma S, Paul JH, et al. The Amazon continuum dataset: quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010. Microbiome. 2014;2:17.
19. Field B, Randerson F. Primary production of the biosphere: integrating terrestrial and oceanic components. Science. 1998;281:237–40.
20. Malhi Y, Roberts JT, Betts RA, Killeen TJ, Li W, Nobre CA. Climate change, deforestation, and the fate of the Amazon. Science. 2008;319:169–72.
21. Mikhailov VN. Water and sediment runoff at the Amazon River mouth. Water Resour. 2010;37:145–59.
22. Subramaniam A, Yager PL, Carpenter EJ, Mahaffey C, Björkman K, Cooley S, et al. Amazon River enhances diazotrophy and carbon sequestration in the tropical North Atlantic Ocean. Proc Natl Acad Sci U S A. 2008;105:10460–5.
23. Sioli H. The Amazon and its main affluents: hydrography, morphology of the river courses, and river types. Amaz Limnol Landsc Ecol Mighty Trop River Its Basin. Dordrecht: Springer Netherlands: Sioli, H.; 1984. p. 127–165.
24. Wissmar RC, Richey JE, Stallard RF, Edmond JM. Plankton metabolism and carbon processes in the Amazon River, its tributaries, and floodplain waters, Peru-Brazil, May-June 1977. Ecology. 1981;62:1622–33.
25. Mayorga E, Aufdenkampe AK, Masiello CA, Krusche AV, Hedges JI, Quay PD, et al. Young organic matter as a source of carbon dioxide outgassing from Amazonian rivers. Nature. 2005;436:538.
26. Richey JE, Melack JM, Aufdenkampe AK, Ballester VM, Hess LL. Outgassing from Amazonian rivers and wetlands as a large tropical source of atmospheric CO_2. Nature. 2002;416:617–20.
27. Ward ND, Keil RG, Medeiros PM, Brito DC, Cunha AC, Dittmar T, et al. Degradation of terrestrially derived macromolecules in the Amazon River. Nat Geosci. 2013;6:530–3.
28. Ward ND, Bianchi TS, Sawakuchi HO, Gagne-Maynard W, Cunha AC, Brito DC, et al. The reactivity of plant-derived organic matter and the potential importance of priming effects along the lower Amazon River. J Geophys Res Biogeosciences. 2016;121:1522–39.
29. Ertel JR, Hedges JI, Devol AH, Richey JE, Ribeiro M de NG. Dissolved humic substances of the Amazon River system. Limnol Oceanogr. 1986;31:739–54.
30. Seidel M, Dittmar T, Ward ND, Krusche AV, Richey JE, Yager PL, et al. Seasonal and spatial variability of dissolved organic matter composition in the lower Amazon River. Biogeochemistry. 2016;131:281–302.
31. Gagne-Maynard WC, Ward ND, Keil RG, Sawakuchi HO, Da Cunha AC, Neu V, et al. Evaluation of primary production in the lower Amazon River based on a dissolved oxygen stable isotopic mass balance. Front Mar Sci. 2017;4:26.
32. Satinsky BM, Crump BC, Smith CB, Sharma S, Zielinski BL, Doherty M, et al. Microspatial gene expression patterns in the Amazon River plume. Proc Natl Acad Sci U S A. 2014;111:11085–90.
33. Cragg SM, Beckham GT, Bruce NC, Bugg TD, Distel DL, Dupree P, et al. Lignocellulose degradation mechanisms across the Tree of Life. Curr Opin Chem Biol. 2015;29:108–19.
34. Wilhelm RC, Singh R, Eltis LD, Mohn WW. Bacterial contributions to delignification and lignocellulose degradation in forest soils with metagenomic and quantitative stable isotope probing. ISME J. 2019;13:413–29.
35. Kögel-Knabner I. The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter. Soil Biol Biochem. 2002;34:139–62.
36. Bugg TDH, Ahmad M, Hardiman EM, Rahmanpour R. Pathways for degradation of lignin in bacteria and fungi. Nat Prod Rep. 2011;28:1883–96.
37. Janusz G, Pawlik A, Sulej J, Świderska-Burek U, Jarosz-Wilkołazka A, Paszczyński A. Lignin degradation: microorganisms, enzymes involved, genomes analysis and evolution. FEMS Microbiol Rev. 2017;41:941–62.
38. Wantzen KM, Yule CM, Mathooko JM, Pringle CM. 3 - organic matter processing in tropical streams. In: Dudgeon D, editor. Trop Stream Ecol [Internet]. London: Academic Press; 2008 [cited 2020 Feb 26]. p. 43–64. Available from: http://www.sciencedirect.com/science/article/pii/B9780120884490500054.
39. Benner R, Moran MA, Hodson RE. Biogeochemical cycling of lignocellulosic carbon in marine and freshwater ecosystems: Relative contributions of procaryotes and eucaryotes1. Limnol Oceanogr. 1986;31:89–100.
40. Benner R, Opsahl S, Chin-Leo G, Richey JE, Forsberg BR. Bacterial carbon metabolism in the Amazon River system. Limnol Oceanogr. 1995;40:1262–70.
41. Hernes PJ, Benner R. Photochemical and microbial degradation of dissolved lignin phenols: implications for the fate of terrigenous dissolved organic matter in marine environments. J Geophys Res Oceans. 2003;108:3291.

42. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. Nature Publishing Group. 2020:1–8.
43. Tessler M, Brugler MR, DeSalle R, Hersch R, Velho LFM, Segovia BT, et al. A global eDNA comparison of freshwater bacterioplankton assemblages focusing on large-river floodplain lakes of Brazil. Microb Ecol. 2017;73:61–74.
44. Staley C, Gould TJ, Wang P, Phillips J, Cotner JB, Sadowsky MJ. Core functional traits of bacterial communities in the Upper Mississippi River show limited variation in response to land cover. Front Microbiol. 2014;5:414.
45. Kamimura N, Takahashi K, Mori K, Araki T, Fujita M, Higuchi Y, et al. Bacterial catabolism of lignin-derived aromatics: New findings in a recent decade: update on bacterial lignin catabolism. Environ Microbiol Rep. 2017;9:679–705.
46. Poretsky RS, Sun S, Mou X, Moran MA. Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. Environ Microbiol. 2010;12:616–27.
47. Rosa LT, Dix SR, Rafferty JB, Kelly DJ. Structural basis for high-affinity adipate binding to AdpC (RPA4515), an orphan periplasmic-binding protein from the tripartite tricarboxylate transporter (TTT) family in Rhodopseudomonas palustris. FEBS J. 2017;284:4262–77.
48. Hosaka M, Kamimura N, Toribami S, Mori K, Kasai D, Fukuda M, et al. Novel tripartite aromatic acid transporter essential for terephthalate uptake in Comamonas sp. strain E6. Appl Environ Microbiol. 2013;79:6148–55.
49. Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. A guide to the natural history of freshwater lake bacteria. Microbiol Mol Biol Rev MMBR. 2011;75:14–49.
50. Hutalle-Schmelzer KML, Zwirnmann E, Krüger A, Grossart H-P. Enrichment and cultivation of pelagic bacteria from a humic lake using phenol and humic matter additions. FEMS Microbiol Ecol. 2010;72:58–73.
51. Logares R, Brate J, Bertilsson S, Clasen JL, Shalchian-Tabrizi K, Rengefors K. Infrequent marine–freshwater transitions in the microbial world. Trends Microbiol. 2009;17:414–22.
52. Herlemann DPR, Manecki M, Meeske C, Pollehne F, Labrenz M, Schulz-Bull D, et al. Uncoupling of bacterial and terrigenous dissolved organic matter dynamics in decomposition experiments. PLOS ONE. Public Library of Science; 2014;9:e93945.
53. Sipler RE, Kellogg CTE, Connelly TL, Roberts QN, Yager PL, Bronk DA. Microbial community response to terrestrially derived dissolved organic matter in the coastal arctic. Front Microbiol. 2017;8:1018.
54. Qin L, Li W-C, Liu L, Zhu J-Q, Li X, Li B-Z, et al. Inhibition of lignin-derived phenolic compounds to cellulase. Biotechnol Biofuels. 2016;9:70.
55. Monlau F, Sambusiti C, Barakat A, Quemeneur M, Trably E, Steyer JP, et al. Do furanic and phenolic compounds of lignocellulosic and algae biomass hydrolyzate inhibit anaerobic mixed cultures? A comprehensive review. Biotechnol Adv. 2014;32:934–51.
56. Xue S, Jones AD, Sousa L, Piotrowski J, Jin M, Sarks C, et al. Water-soluble phenolic compounds produced from extractive ammonia pretreatment exerted binary inhibitory effects on yeast fermentation using synthetic hydrolysate. PLOS ONE. 2018;13:e0194012.
57. Aston JE, Apel WA, Lee BD, Thompson DN, Lacey JA, Newby DT, et al. Degradation of phenolic compounds by the lignocellulose deconstructing thermoacidophilic bacterium Alicyclobacillus Acidocaldarius. J Ind Microbiol Biotechnol. 2016;43:13–23.
58. Farjalla VF. Are the mixing zones between aquatic ecosystems hot spots of bacterial production in the Amazon River system? Hydrobiologia. 2014;728:153–65.
59. Laraque A, Guyot JL, Filizola N. Mixing processes in the Amazon River at the confluences of the Negro and Solimões Rivers, Encontro das Águas, Manaus. Brazil. Hydrol Process. 2009;23:3131–40.
60. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10.
61. Andrews S. Babraham Bioinformatics - FastQC a quality control tool for high throughput sequence data [Internet]. 2017 [cited 2017 Nov 8]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
62. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, et al. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. 2016;102:3–11.
63. Vollmers J, Wiegand S, Kaster A-K. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! Rodriguez-Valera F, editor. PLOS ONE. 2017;12:e0169662.
64. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.
65. Staley C, Gould TJ, Wang P, Phillips J, Cotner JB, Sadowsky MJ. Bacterial community structure is indicative of chemical inputs in the Upper Mississippi River. Front Microbiol. 2014;5:524.
66. Van Rossum T, Peabody MA, Uyaguari-Diaz MI, Cronin KI, Chan M, Slobodan JR, et al. Year-long metagenomic study of river microbiomes across land use and water quality. Front Microbiol. 2015;6:1405.
67. Meyer KM, Klein AM, Rodrigues JLM, Nüsslein K, Tringe SG, Mirza BS, et al. Conversion of Amazon rainforest to agriculture alters community traits of methane-cycling organisms. Mol Ecol. 2017;26:1547–56.
68. Benoit G, Peterlongo P, Mariadassou M, Drezen E, Schbath S, Lavenier D, et al. Multiple comparative metagenomics using multiset k-mer counting. PeerJ Comput Sci. 2016;2:e94.
69. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci. 2003;14:927–30.
70. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
71. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
72. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.
73. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet TIG. 2000;16:276–7.
74. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
75. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
76. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods. 2012;10:71–3.
77. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40:D109–14.
78. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003;4:41.
79. Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, et al. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. Nucleic Acids Res. 2011;39:D546–51.
80. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43:D204–12.
81. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2014;12:59–60.
82. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45:D353–61.
83. Eddy SR. Accelerated Profile HMM Searches. Pearson WR, editor. PLoS Comput Biol. 2011;7:e1002195.
84. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2012;40:W445–51.
85. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44:D279–85.
86. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44:D286–93.
87. Kanehisa M, Sato Y. KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci Publ Protein Soc. 2020;29:28–35.
88. Ye Y, Doak TG. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. Ouzounis CA, editor. PLoS Comput Biol. 2009;5:e1000465.
89. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. Nature Publishing Group. 2017;35:1026–8.
90. Brumm PJ. Bacterial genomes: what they teach us about cellulose degradation. Biofuels. 2013;4:669–81.
91. López-Mondéjar R, Zühlke D, Becher D, Riedel K, Baldrian P. Cellulose and hemicellulose decomposition by forest soil bacteria proceeds by the action of structurally variable enzymatic systems. Sci Rep. 2016;6.

92. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics. 2010;26:1608–15.
93. Wei T, Simko V. R package "corrplot": Visualization of a Correlation Matrix [Internet]; 2017 [cited 2017 Nov 7]. Available from: https://github.com/taiyun/corrplot.
94. Neuwirth E. CRAN - Package ColorBrewer Palettes [Internet]. Comprehensive R Archive Network (CRAN); 2014 [cited 2017 Nov 7]. Available from: https://cran.r-project.org/web/packages/RColorBrewer/index.html.
95. Douglas Nychka, Reinhard Furrer, John Paige, Stephan Sain. fields: Tools for spatial data [Internet]. Boulder, CO, USA: University Corporation for Atmospheric Research; 2017 [cited 2017 Nov 7]. Available from: www.image.ucar.edu/ nychka/Fields.

Publisher's Note