# Uncovering the gene machinery of the Amazon River microbiome to degrade rainforest organic matter

**Célio Dias Santos Júnior**
Universidade Federal de Sao Carlos
**Hugo Sarmento**
Universidade Federal de Sao Carlos
**Fernando Pellon de Miranda**
Petrobras
**Flávio Henrique-Silva**
Universidade Federal de Sao Carlos
**Ramiro Logares** ( Ramiro.Logares@gmail.com )
Institut de Ciencies del Mar https://orcid.org/0000-0002-8213-0604

**Research**

**Keywords:** Amazon River, freshwater bacteria, biodiversity, metagenomics, lignin degradation, cellulose degradation, priming effect, gene catalogue

**Posted Date:** November 14th, 2019

**DOI:** https://doi.org/10.21203/rs.2.17206/v1

# 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 analyzing 106 metagenomes from 30 stations 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 sampled environments (rivers and rainforest soil). Analyses of TeOM-degradation genes revealed that lignin degradation pathways correlated to tricarboxylates and hemicellulose processing, pointing to a higher lignin degradation coupled to the consumption of labile compounds. We propose a model on how the degradation of recalcitrant TeOM modulated by labile compounds (i.e. priming effect) may operate in the Amazon River waters.

Conclusions: Our work contributes to expand significantly our comprehension of the world's largest river microbiome and its role in TeOM degradation. Furthermore, the AMnrGC represents an important resource for future works exploring the links between TeOM and its degradation by aquatic microbiotas in tropical ecosystems.


# Background

Continental waters play a major biogeochemical role by linking terrestrial and marine ecosystems [1]. In particular, rainforest rivers typically receive large amounts of terrestrially-derived organic matter (TeOM), which may then reach the ocean. TeOM is typically 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, ~40 million genes have been reported for the global ocean microbiome [7] and ~160 million genes for the global topsoil microbiome [8]. So far, there is no comprehensive gene catalogue for rivers, which hinders our comprehension of the genomic machinery that degrade almost half of the 1.9 Pg C of recalcitrant TeOM that are discharged into rivers every year [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 [11,12]. The Amazon River basin comprises almost 38% of continental South America [13] and its discharge accounts for 18% of the world's inland-water inputs to the oceans [14]. Despite its relevance for global scale processes, there is a limited understanding of the Amazon River microbiome, as well as the microbiomes from other large tropical rivers.

The large amounts of organic and inorganic particulate material [15] turns the Amazon River into a turbid system. High turbidity reduces light penetration and, consequently, the Amazon River has very low rates of algal production [16], meaning that TeOM is the major carbon source for microbial growth [17]. 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 [18], almost equivalent to the amount of carbon sequestered by the forest[11,12]. 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 [19,20]. 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 [20].

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 [19–23], the functions associated to their degradation were expected to be widespread in the Amazon microbiome. Instead, these functions exhibited very low abundances [24–26], 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) catalyze the hydrolysis of glycoside linkages, while polysaccharide esterases support the action of GHs over hemicelluloses, and polysaccharide lyases promote depolymerization [27,28]. In contrast, lignin is more resistant to degradation [29], since its role is preventing microbial enzymes from degrading labile cell-wall polysaccharides [30]. 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 [31]. 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 [21,22]. Lignin degradation tends to occur in oxic waters of the Amazon River, using the hydrogen peroxide produced by the metabolism of cellulose and hemicellulose [32].

Here, we produced the first gene catalogue of the world's largest rainforest river by analyzing 106 metagenomes (~500 $x10^9$ base pairs), originating from 30 stations covering a total of ~ 2,106 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 to metabolize 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 associated to TeOM degradation? Do TeOM degradation genes and functions display a spatial distribution pattern? 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 in groups that shared the same geographic origin (*Figure 1a; Supplementary Tables 1 and 2, Additional file 1).*. 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. About 52% of the AMnrGC genes were annotated with at least one database (*Figure 1b),*, 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 accumulation curves, which tended towards saturation (*Figure 1c).*.

# The Amazon River microbiome differed from other microbiomes

We compared the metagenomic information contained in the Amazon River microbiome with that of Amazon rainforest soil and other available rivers (Canada watersheds and Mississippi river) using k-mers (*Supplementary Table 3, Additional file 1).*. The k-mer comparison of these microbiomes indicated that they are different in terms of genomic information content (*Figure 1d),* forming groups of heterogenous composition (significant β dispersion [that is, average distance of samples to the group centroid] - PERMUTEST, $F = 25.7$, $p < 0.001$). In particular, the genomic information content of Amazon River samples was markedly different to the other microbiomes (PERMANOVA, $R^2 = 0.10$, $p = 9.99x10^{-5}$; ANOSIM, $R = 0.27$, $p < 0.001$), which suggests that this basin, or tropical rainforest rivers in general, 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) were significantly different between them (PERMANOVA test, $F = 1.52$, $p < 9.9e{-}5$; *Figure 2a),*, indicating that they represent different gene assemblages. Each of these groups was homogenous, as there was a non-significant β dispersion ($F = 2.3$, $p = 0.06$) among metagenomic samples in each group (*Figure 2b).*.

# 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; *Figure 1b).*. Genes annotated exclusively through Hidden Markov Models (HMM) represented 13.3% of the AMnrGC (*Figure 1b).*. 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" (*Figure 1b).*.

The previous highlights 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 in the AMnrGC, with only 0.3% and 0.6% of the genes having eukaryotic or viral origin, respectively (*Figure 1b).*.

# Core metabolisms

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; *Figure 3).*. Genes with unknown function represented 21.4% of the COG annotated proteins.

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 the general carbon metabolism, being predominantly associated to 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).

# TeOM degradation machinery

A total of 6,516 genes from the AMnrGC were identified as taking part in the TeOM degradation machinery of 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 (2,324 genes) and tricarboxylate transport (3,884 genes) [*Figure 4].*. The large

gene diversity associated to the metabolism of lignin-derived compounds and the transport of tricarboxylates likely reflects the molecular diversity of the compounds generated during the lignin oxidation process that are present in the Amazon River waters as humic and fulvic acids.

# 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 involved in the oxidation of lignin and degradation of cellulose and hemicellulose were investigated (*Figure 4)*.. We found that lignin oxidation in the Amazon River seems to be mainly mediated by dye-decolorizing peroxidases (DyPs), being predominantly associated to freshwater areas. Only laccases and peroxidases were found in the Amazon River microbiome, no other families involved in lignin oxidation, like phenolic acid decarboxylase or glyoxal oxidase, were found. In turn, hemicellulose degradation seems to be performed mainly by glycosyl hydrolase GH10 in all river sections (*Figure 4b)*.. We observed a similar ubiquitous dominance of glycosyl hydrolase GH3 in cellulose degradation across river sections (*Figure 4a)*..

# Degradation of lignin-derived aromatic compounds

Following the initial degradation of lignin, plenty of 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 O-demethylation/C1 metabolism and ring cleavage pathways to form pyruvate or oxaloacetate, which can be incorporated to the TCA cycle 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 complete degradation pathway of lignin-derived compounds (*Figure 4d)* included 772 and 449 genes belonging to funneling pathways of diaryls and monoaryls, respectively. Examination of the pathways starting with vanilate and syringate revealed 346 genes responsible for the O-demethylation and C1-metabolism steps, while 713 genes seemed 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, 2) O-demethylation/C1 metabolism and 3) ring cleavage.

We evaluated whether genes associated to TeOM degradation had a spatial distribution pattern along the river course. For this, we used the linear geographic distance of samples to the Amazon River source in Peru. Geographic distance was negatively correlated with the number of genes associated to lignin oxidation, hemicellulose degradation, ring cleavage pathway, tripartite tricarboxylate transporting and the

AAHS transporters (*Figure 5).*. This is coherent with a trend displayed by gene-function along the river, which points to cellulose and hemicellulose degradation being replaced by lignin oxidation in brackish waters. A potential reduction of the microbial gene repertoire related to lignin processing as the river approaches the ocean points to the aging of TeOM during its flow through the river.

The gene machinery associated to the processing of lignin-derived aromatic compounds was positively correlated to lignin oxidation genes along the river course (*Figure 5)„* suggesting a co-processing of lignin and its byproducts. Lignin oxidation and hemicellulose degradation pathways were positively correlated (*Figure 5)„* supporting the idea that monomers of hemicellulose, mainly carbohydrates, could be priming lignin oxidation. In terms of genes, cellulose degradation was not correlated with lignin oxidation, but had a weak positive correlation to hemicellulose degradation (*Figure 5)„* suggesting a coupling between both pathways.

We did not find correlations between the genes associated to the different types of funneling pathways (FP Dimers and FP Monomers) and the linear geographic distance along the river course (*Figure 5).*. This indicates that the degradation of lignin-derived aromatic compounds was not restricted to any river section. Moreover, the number of genes related to hemi-/cellulose degradation was positively correlated to lignin-derived aromatic compounds degradation pathways, revealing a potential co-metabolism of lignin-derived compounds and hemi-/cellulose degradation, instead of lignin-oxidation.

# Tripartite tricarboxylate transporting system and the processing of allochthonous organic carbon in Amazon River

Lignin-derived aromatic compounds need to be transported from the extracellular environment to the cytoplasm prior to their degradation. Transporters that could be associated to lignin degradation (MFS transporter, AAHS family and ABC transporters) were found in the AMnrGC, while transporters from the MHS family, ITS superfamily and TRAP were not. MFS transporters were not correlated to any of the other examined pathways. AAHS transporters were negatively correlated to geographic distance, while the other transporter families did not show any type of correlation with distance (*Figure 5).* Furthermore, AAHS and ABC transporters showed positive correlations to the funneling pathway of monoaryls, suggesting their transport by those transporter families. ABC transporters were positively correlated to O-demethylation and C1 metabolism, while AAHS and ABC transporters were correlated to the ring cleavage pathway. This suggests that ABC and AAHS transporters are relevant for the metabolism of monoaryls derived from lignin oxidation.

The tripartite tricarboxylate transporting (TTT) system was correlated to the processing of allochthonous organic material in the Amazon River. Three proteins compose this system, where one is responsible of capturing substrates in the extracellular space and bringing them to the transporting channel made by the other two proteins, which recognize the substrate binding protein and internalize the substrate. There was

a large gene diversity associated to the substrate binding proteins, since each protein is specific to one or a few substrates. Furthermore, the number of genes in the TTT system displayed a negative correlation with linear geographic distance, suggesting its predominance in freshwaters sections (*Figure 5)*..

The TTT system was positively correlated to AAHS and ABC transporters (*Figure 5)* pointing to functional complementarity, as the TTT would transport substrates not transported by the other transporter families. Furthermore, the TTT transporters showed a positive correlation with lignin oxidation and hemicellulose degradation, suggesting either the transport of the products of those processes or a dependence of these processes in the compounds transported by the TTT.

# Discussion

The AMnrGC allows to expand significantly our comprehension of the world's largest river microbiome. 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. Furthermore, half of the ~3.7 million genes in the AMnrGC had no orthologs, suggesting gene novelty. Altogether, this could reflect a distinct microbiota in the Amazon River and perhaps local adaptive evolution, although more samples from other rivers are necessary to test these hypotheses.

Analyses of COG functions pointed to a number of core functional genes along the Amazon River course, which was supported by the similar distribution of COG super-classes along the different river sections (*Figure 3)*.. In particular, COG functions within the superclass "Metabolism" were the most abundant in the AMnrGC, as well as in the upper Mississippi River [33]. Salinity is known to affect microbial spatiotemporal distribution, and jumps across the salinity barrier are rare evolutionary events [34]. We observed a subset of gene functions present in both fresh- and brackish water sections, pointing to core genes, while other gene functions displayed a predominance in either of these sections, pointing to non-core genes structured by salinity. The plume section displayed a higher gene diversity than the ocean, probably reflecting the coalescence of freshwater and marine microbial communities and their gene repertoires [35].

Core functions included a general carbohydrate metabolism and several transporter systems, mainly ABC transporters. Our results suggest a sophisticated machinery to process TeOM in the Amazon River, where TeOM degradation appears related to acetogenic and methanotrophic pathways. This agrees with previous findings indicating a high expression of C1 metabolism genes (i.e. methane monooxygenase - *mmoB* and formaldehyde activating enzyme - *fae)* [24]. The non-core pathways suggest 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).

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]. 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 [38]. Recent findings also

suggest a TTT system related to the transport of TeOM degradation byproducts [39,40]. 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 suggest that the TTT transporters could be linked to lignin oxidation and hemicellulose degradation, supporting their role in TeOM degradation.

Based in our findings, we propose a model of the potential priming effect acting in ligno-cellulose complexes in the Amazon River (*Figure 6).*. In this model, there are two different communities co-existing in a consortium: one responsible for hemi-/cellulose degradation and another one responsible for lignin degradation. The first community releases extracellular enzymes (mainly glucosyl hydrolases from families GH3 and GH10), whose reaction produces carbohydrates. These sugars can provide structural carbon and energy for the hemi-/cellulose degrader community as well as for the lignin degrader community. The lignolytic community can also use the cellulolytic byproducts to growth, promoting an oxidative metabolism. This oxidative metabolism triggers the production and secretion of reactive oxygen species (ROS) [*Figure 6].*. ROS are then used by DyPs and laccases secreted by lignolytic communities to oxidize lignin, exposing more hemi-/cellulose to cellulolytic communities and re-starting the cycle (*Figure 6).*. Another important role of lignolytic communities is the degradation of lignin-derived aromatic compounds generated by the lignin oxidation. If those compounds are not degraded, they could inhibit cellulolytic enzymes and microbial growth [41–44], preventing TeOM degradation. This cycle may be considered as a priming effect, where both communities benefit from each other.

# Conclusions

Our work represents a first effort to link carbon fixation occurring in the Amazon rainforest with the degradation of a substantial fraction of the terrestrially-fixed carbon in the Amazon River. Our results point to multiple metabolic mechanisms, mainly prokaryotic, that are likely key in the degradation of the large amounts of TeOM that are discharged every day into the Amazon River. Last but not least, the generated AMnrGC represents a resource that can be used to investigate diverse ecological questions in an understudied ecosystem such as the Amazon River as well as to explore biotechnological applications. Future studies using metatranscriptomics need to determine the expression of the genes reported in this study in order to advance our understanding of the TeOM degradation in the Amazon River.

# Methods

We analyzed 106 metagenomes [45–48] from 30 stations distributed along the Amazon River basin, with an average coverage of $5.0x10^9$ base pairs per metagenome (*Supplementary Table 1, Additional file 1).*. The stations 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 ~2,106 km and were divided into 5 sections (*Figure 1a; Supplementary Table 1, 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 [49,50]), 3) *Estuary section* (part of the river that meets the Atlantic Ocean) and 4) *Plume section* (the area where the Ocean is influenced by the Amazon River inputs).

Samples were taken as previously indicated [45–48]. 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 [45–48]. 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, Additional file 1.*

# Metagenome analysis

*Illumina* adapters and poor-quality bases were removed from metagenomes using Cutadapt [51]. Only reads longer than 80 bp, containing bases with Q ≥ 24, were kept. The quality of the reads was checked with FASTQC [52]. Reads from metagenomes belonging to the same station were assembled together using MEGAHIT (v1.0) [53], with the meta-large presets. Only contigs > 1 Kbp were considered, as recommended by previous work [54]. Assembly quality was assessed with QUAST [55]. Metagenome assembly yielded 2,747,383 contigs ≥1,000 base pairs, in a total assembly length of ~ $5.5x10^9$ base pairs (see *Supplementary Table 2, Additional file 1).*.

# Analysis of k-mer diversity over different river zones

A k-mer diversity analysis was used to compare the genetic information along the Amazon River microbiome against that in other microbiomes from Amazon forest soil or temperate rivers (*Supplementary Table 3, Additional file 1).*.Specifically, the Amazon River metagenomes (106) were compared against 37 metagenomes from the Mississippi River [56], 91 metagenomes from three watersheds in Canada [57], and 7 metagenomes from the Amazon forest soil [58]. The rationale to include soil metagenomes was to check whether genomic information in the river could derive from soil microbiotas. K-mer comparisons were run with SIMKA (version 1.4) [59] 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 beta-dispersion in the groups, and permutational multivariate analysis of variance (PERMANOVA/ANOSIM) was used to test the groups' difference. Both analyzes were performed using the R package Vegan [60].

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

Genes were predicted using Prodigal (version 2.6.3) [61]. 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) [62,63] 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 the Infoseq, EMBOSS package (version 6.6.0.0) [64].

# Gene abundance estimation

The quality-checked sequencing reads were mapped against our non-redundant gene catalogue using BWA (version 0.7.12-r1039) [65] and SamTools (version 1.3.1) [66]. Gene abundances were estimated using the software eXpress (version 1.5.1) [67], with no bias correction, as counts per million (CPM). We used a CPM $\geq 1.00$ 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) [68], COG (Release 2014) [69], CAMERA Prokaryotic Proteins Database (Release 2014) [70] and UniProtKB (Release 2016–08) [71] via the Blastp algorithm implemented in Diamond (v.0.9.22) [72], with a query coverage $\geq 50\%$, identity $\geq 45\%$, e-value $\leq 1e^{-5}$ and score $\geq 50$. KO-pathway mapping was performed using KEGG mapper [73]. HMMSearch (version 3.1b1) [74] was used to search proteins against dbCAN (version 5) [75], PFAM (version 30) [76] and eggNOG (version 4.5) [77] databases, using an e-value $\leq 1e^{-5}$, and posterior probability of aligned residues $\geq 0.9$, and no domain overlapping. Accumulation curves were obtained using random progressive nested comparisons with 100 pseudo-replicates for genes and PFAM predictions.

# Gene taxonomy assignment

Gene-taxonomy was assigned considering the best hits (score, e-value and identity; see above) using KEGG (Release 2015–10–12) [68], UniProtKB (Release 2016–08) [71]and CAMERA Prokaryotic Proteins Database (Release 2014) [70]. Taxonomic last common ancestors (LCA) were determined from TaxIDs (NCBI) associated to UniRef100 and KO entries. Information from the CAMERA database was also used to retrieve taxonomy (NCBI TaxID). Proteins were annotated as ‘unassigned’ if their taxonomic signatures were mixed, containing representatives from several domains of life, or if they only had the function

assigned without taxonomic information. Reference sequences with hits to poorly annotated sequences from other metagenomes were referred as “Metagenomic”.

# TeOM degradation machinery

To investigate TeOM degradation, we grouped samples by river section and assessed their gene content. Genes were then searched against reference sequences and protein families involved in TeOM degradation (see *Supplementary Table 4, Additional file 1).*. In particular, lignin degradation starts with extracellular polymer oxidation followed by internalization and metabolism of the produced monomers or dimers by bacteria. 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) [72], with query coverage ≥ 50%, protein identity ≥ 40% and e-value ≤ $1e^{-5}$ as recommended by Kamimura et al. [36], 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) [78] and cellulose-binding motifs (CBM1, CBM2, CBM3, CBM6, CBM8, CBM30 and CBM44) [78,79] were searched in PFAM annotations. In addition, the most common hemicellulolytic families (GH2, GH10, GH11, GH16, GH26, GH30, GH31, GH39, GH42, GH43 and GH53) [79] were searched in the PFAM database. Lytic polysaccharide monooxygenases (LPMO) [79] 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 diversity of transporter systems [37]. The typical transporters associated to lignin degradation (MFS transporter, AAHS family, ABC transporters, MHS family, ITS superfamily and TRAP transporter) were searched with Diamond (v.0.9.22) [72], using query coverage ≥ 50%, protein identity ≥ 40% and e-value ≤ $1e^{-5}$ and a reference dataset [36].

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) [36]. Recently, several substrate binding proteins (TctC) belonging to the tripartite tricarboxylate transporter (TTT) system were associated to the transport of TeOM degradation byproducts, like adipate [39] and terephthalate [40]. 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 above-mentioned strategy were submitted to PSORT v.3.0 [80], 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 total gene counts per metagenome. Subsequently, correlograms were produced using Pearson’s correlation coefficients with the R packages Corrplot [81] and RColorBrewer [82]. 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 count along the Amazon River course. Geographic distance was calculated with the R package Fields [83].

# 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

DyPs - Dye-decolorizing peroxidases

FP - Funneling pathway

GHn - Glucosyl-hydrolase family “n”

KEGG - Kyoto Encyclopedia of Genes and Genomes

LCA - Taxonomic last common ancestor

LPMO - Lytic polysaccharide monooxygenases

MFS - Major facilitator superfamily of transporters

PFAM - Protein Family

Pg C—Peta ($10^{15}$) grams of carbon

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

## Declarations

# Ethics approval and consent to participate

## Not applicable

# Consent for publication

## Not applicable

# Availability of data and materials

Metagenomes used to construct the Amazon River gene catalogue (AMnrGC) are publicly available (See *Supplementary Table 1, Additional file 1)* from the following SRA projects: SRP044326, PRJEB25171 and SRP039390). Other publicly available metagenomes used in the k-mer diversity comparisons are detailed in *Supplementary Table 3, Additional file 1 (*Amazon forest [PRJNA336764, PRJNA336766, PRJNA337825, PRJNA336700, PRJNA336765], Mississippi River [SRP018728] and Canada watersheds [PRJNA287840]). The AMnrGC and all the associated files are publicly available in a permanent Zenodo repository (10.5281/zenodo.1484503).

# Competing interests

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

# Funding

C. D. S. J. was supported by a PhD scholarship from Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil (CNPq #141112/2016–6). F. H. S. and H.S work was supported by Research Productivity grants from CNPq (Process # 311746/2017–9 and #309514/2017–7, respectively). R. L.

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).

# Authors' contributions

CDSJ, FHS & RL designed the study. CDSJ compiled and curated the data and performed bioinformatic analysis. CDSJ, FHS, HS & RL interpreted the results. FHS, RL, FPM and HS supervised and administered the project, providing funding. The original draft was written by CDSJ. All co-authors contributed substantially to manuscript revisions.

# 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 for his orientation with bioinformatics analyses and support. 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.

## 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;10:172–185.
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:S3–14.
3. Guenet B, Danger M, Abbadie L, Lacroix G. Priming effect: bridging the gap between terrestrial and aquatic ecology. Ecology 2010; 91:2850–2861.
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;2:1367–1373.
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;348:1261359–1261359.
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. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. Primary production of the biosphere: integrating terrestrial and oceanic components. Science 1998; 281: 237–40.
12. 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.
13. Mikhailov VN. Water and sediment runoff at the Amazon River mouth. Water Resour. 2010;37:145–159.
14. 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.
15. Sioli H. The Amazon and its main affluents: Hydrography, morphology of the river courses, and river types. In: Sioli H editor. The Amazon. Limnology and Landscape Ecology of a Mighty Tropical River and Its Basin. Dordrecht: Springer; 1984. p. 127–165.
16. 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.
17. 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.
18. Richey JE, Melack JM, Aufdenkampe AK, Ballester VM, Hess LL. Outgassing from Amazonian rivers and wetlands as a large tropical source of atmospheric CO2. Nature. 2002;416:617–20.
19. 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;6:530–533.
20. 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;121:1522–1539.

21. Ertel JR, Hedges JI, Devol AH, Richey JE, Ribeiro M de NG. Dissolved humic substances of the Amazon River system. Limnol Oceanogr 1986;31: 739–754.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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;11:1852–1864.
27. Payne CM, Knott BC, Mayes HB, Hansson H, Himmel ME, Sandgren M, et al. Fungal Cellulases. Chem Rev. 2015;115:1308–1448.
28. van den Brink J, de Vries RP. Fungal enzyme sets for plant polysaccharide degradation. Appl Microbiol Biotechnol. 2011;91:1477–1492.
29. Kögel-Knabner I. The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter. Soil Biol Biochem. 2002;34:139–162.
30. Pauly M, Keegstra K. Cell-wall carbohydrates and their modification as a resource for biofuels. Plant J. 2008;54:559–568.
31. 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;29:108–119.
32. Sanchez C. Lignocellulosic residues: Biodegradation and bioconversion by fungi. Biotechnol Adv. 2009;27:185–194.
33. 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;5:524.
34. Logares R, Brate J, Bertilsson S, Clasen JL, Shalchian-Tabrizi K, Rengefors K. Infrequent marine–freshwater transitions in the microbial world. Trends Microbiol. 2009;17:414–422.
35. Rillig MC, Antonovics J, Caruso T, Lehmann A, Powell JR, Veresoglou SD, et al. Interchange of entire communities: microbial community coalescence. Trends Ecol Evol. 2015;30:470–6.
36. 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.

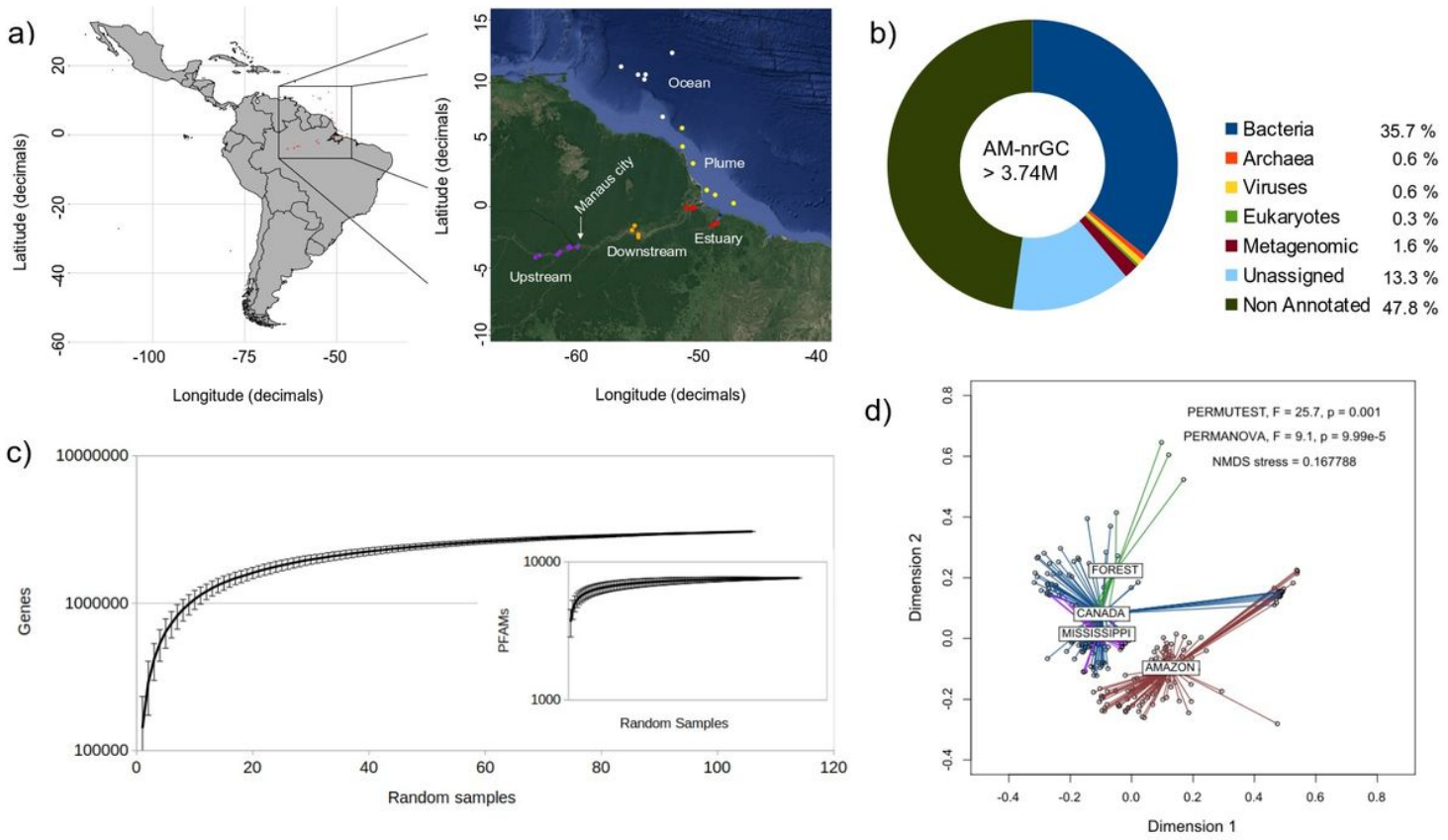
37. Poretsky RS, Sun S, Mou X, Moran MA. Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. Environ Microbiol. 2010;12:616–27.
38. Ghai R, Rodriguez-Valera F, McMahon KD, Toyama D, Rinke R, de Oliveira TCS, et al. Metagenomics of the water column in the pristine upper course of the Amazon river. PLoS ONE. 2011;6:e23785.
39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. 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.
45. 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.
46. 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.
47. Santos-Júnior CD, Toyama D, Oliveira TCS, Miranda FP, Henrique-Silva F. Flood Season Microbiota from the Amazon Basin Lakes: Analysis with Metagenome Sequencing. Microbiol. Res. Announc. 2019;8:e00229–19.

1. 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.
2. Farjalla VF. Are the mixing zones between aquatic ecosystems hot spots of bacterial production in the Amazon River system? Hydrobiologia. 2014;728:153–165.
3. 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–

3140.
4. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10.
5. Andrews S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. 2017. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
6. 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. Methods. 2016;102:3–11.
7. Vollmers J, Wiegand S, Kaster A-K. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters!. PLOS ONE. 2017;12:e0169662.
8. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–1075.
9. 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.
10. 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.
11. 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.
12. 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.
13. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci. 2003;14:927–930.
14. 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.
15. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–3152.
16. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–1659.
17. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet TIG. 2000;16:276–7.
18. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–1760.
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–2079.
20. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods. 2012;10:71–73.

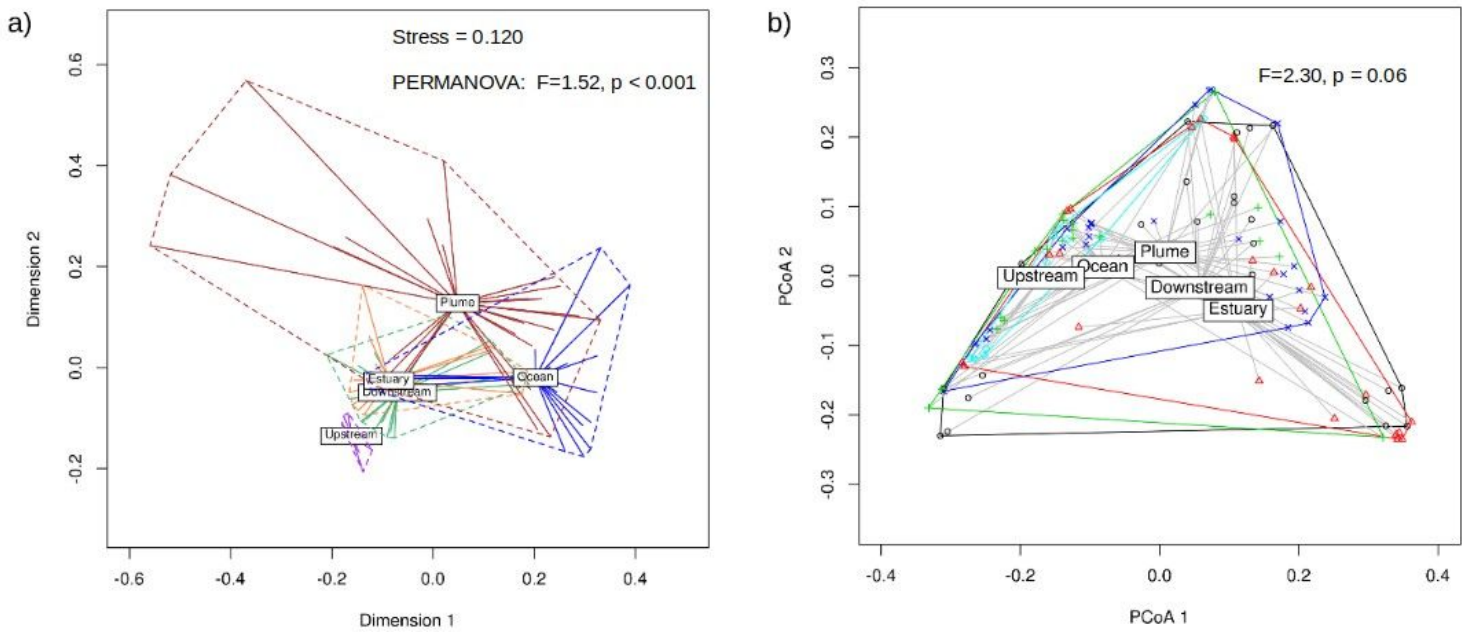
21. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40:D109–14.
22. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003;4:41.
23. 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.
24. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43:D204–D212.
25. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2014;12:59–60.
26. 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–D361.
27. Eddy SR. Accelerated Profile HMM Searches. Pearson WR, editor. PLoS Comput Biol. 2011;7:e1002195.
28. 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.
29. 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.
30. 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–D293.
31. Brumm PJ. Bacterial genomes: what they teach us about cellulose degradation. Biofuels. 2013;4:669–81.
32. 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.
33. 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.
34. Wei T, Simko V. R package 'corrplot': Visualization of a Correlation Matrix. https://github.com/taiyun/corrplot.
35. Neuwirth E.  Package ColorBrewer Palettes. 2014.
36. Douglas Nychka, Reinhard Furrer, John Paige, Stephan Sain. fields: Tools for spatial data. Boulder, CO, USA: University Corporation for Atmospheric Research; 2017. Available from: www.image.ucar.edu/ nychka/Fields

# Figures



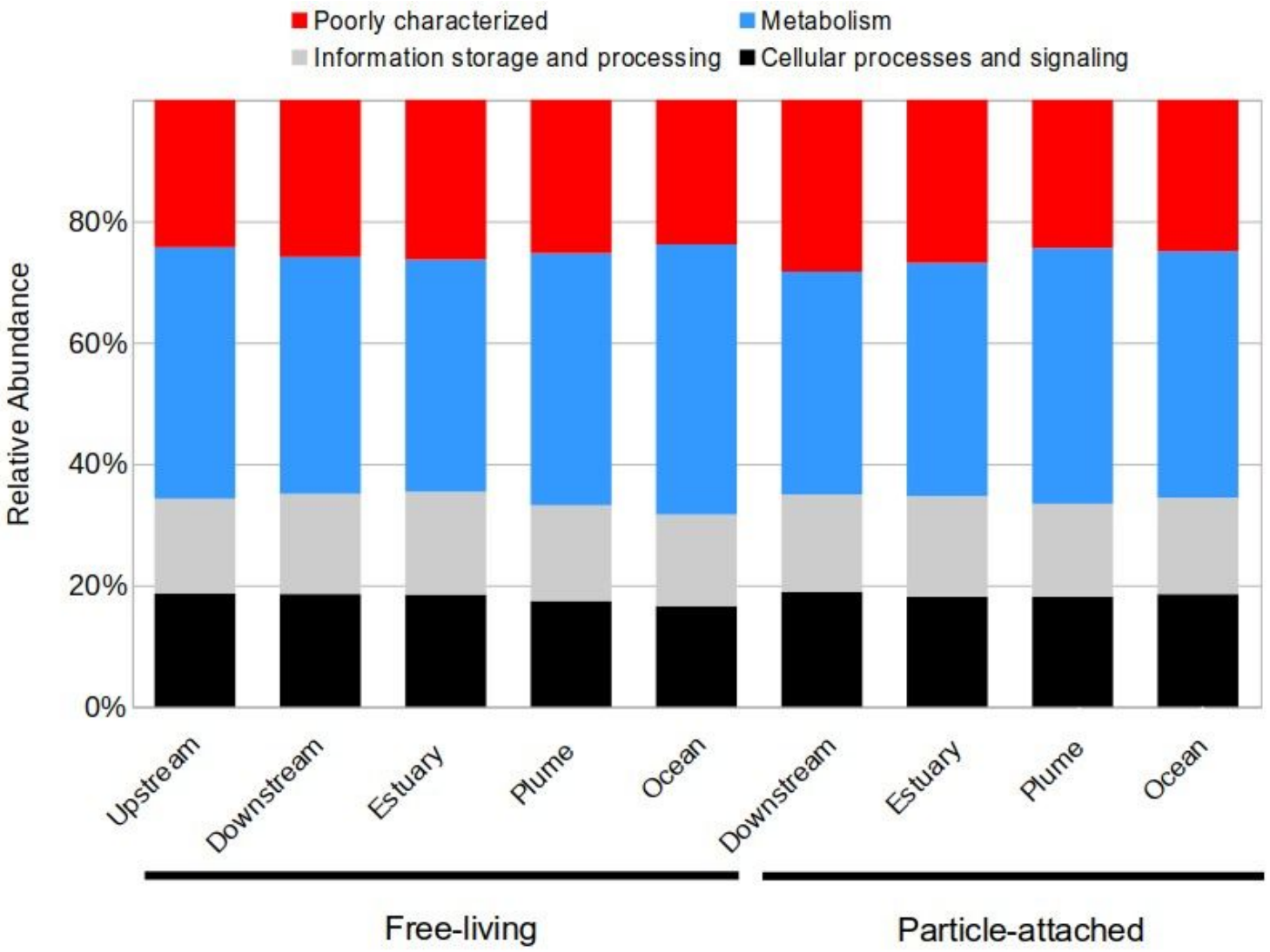

## Figure 1

. 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 ~ 3.74 million genes in the AMnrGC. “Unassigned” genes were not assigned taxonomy, but they were functionally assigned, being different from “non-annotated” genes, which do not have any ortholog. Those genes displaying orthology to poorly characterized genes found in metagenomes were referred as “Metagenomic”. c) Accumulation curves of non-redundant genes and PFAM families (inset) 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 (MISSISSIPI)].



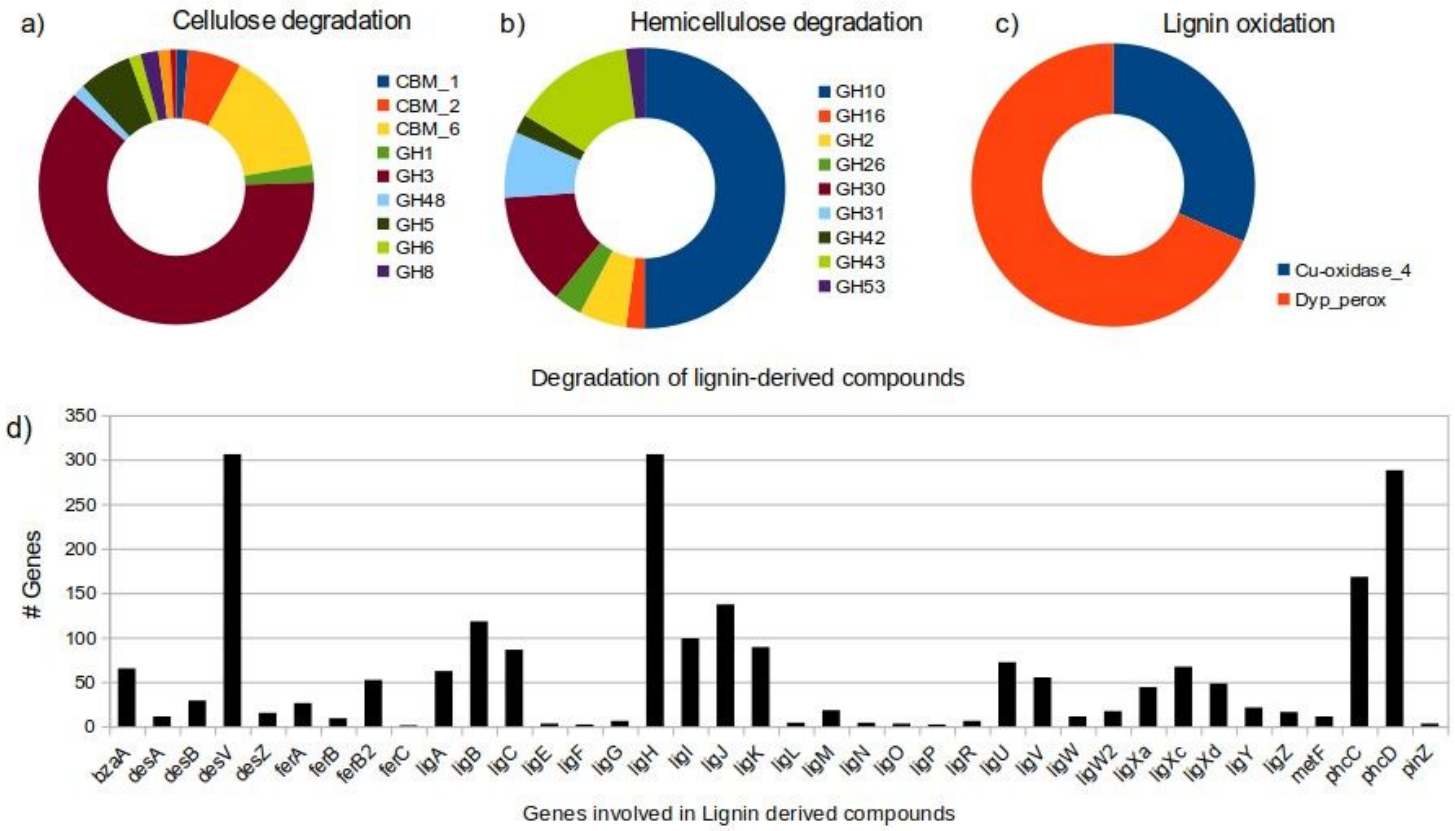

## Figure 2

Metagenomic composition of the five 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) NMDS groups were statistically different [PERMANOVA, F=1.52, p-value = 9.99e-5], b) while the composition inside groups was homogeneous [β-dispersion; PERMUTEST, F=2.30, p-value = 0.06].




## Figure 3

Functional composition across size fractions and sections of the Amazon River. Gene functions grouped into COG super classes are shown per river section and microbial lifestyle (particle-attached vs. free-living). Functions related to the metabolism super class were more represented in free-living that in particle-attached communities ($p < 0.05$, Mann-Whitney U Test). In fresh- and brackish water, all COG classes were differentially distributed, with higher gene diversities observed in freshwaters ($p < 0.01$, Mann-Whitney U Test). The Upstream river section is not shown in the particle-associated fraction, since it was not sampled.




## Figure 4

Main enzyme families involved in TeOM degradation in the AMnrGC. a) Cellulose and b) hemi-cellulose degradation, c) lignin oxidation, and d) lignin-derived compounds degradation shown as number of detected genes per protein family. Abbreviations indicate protein families.

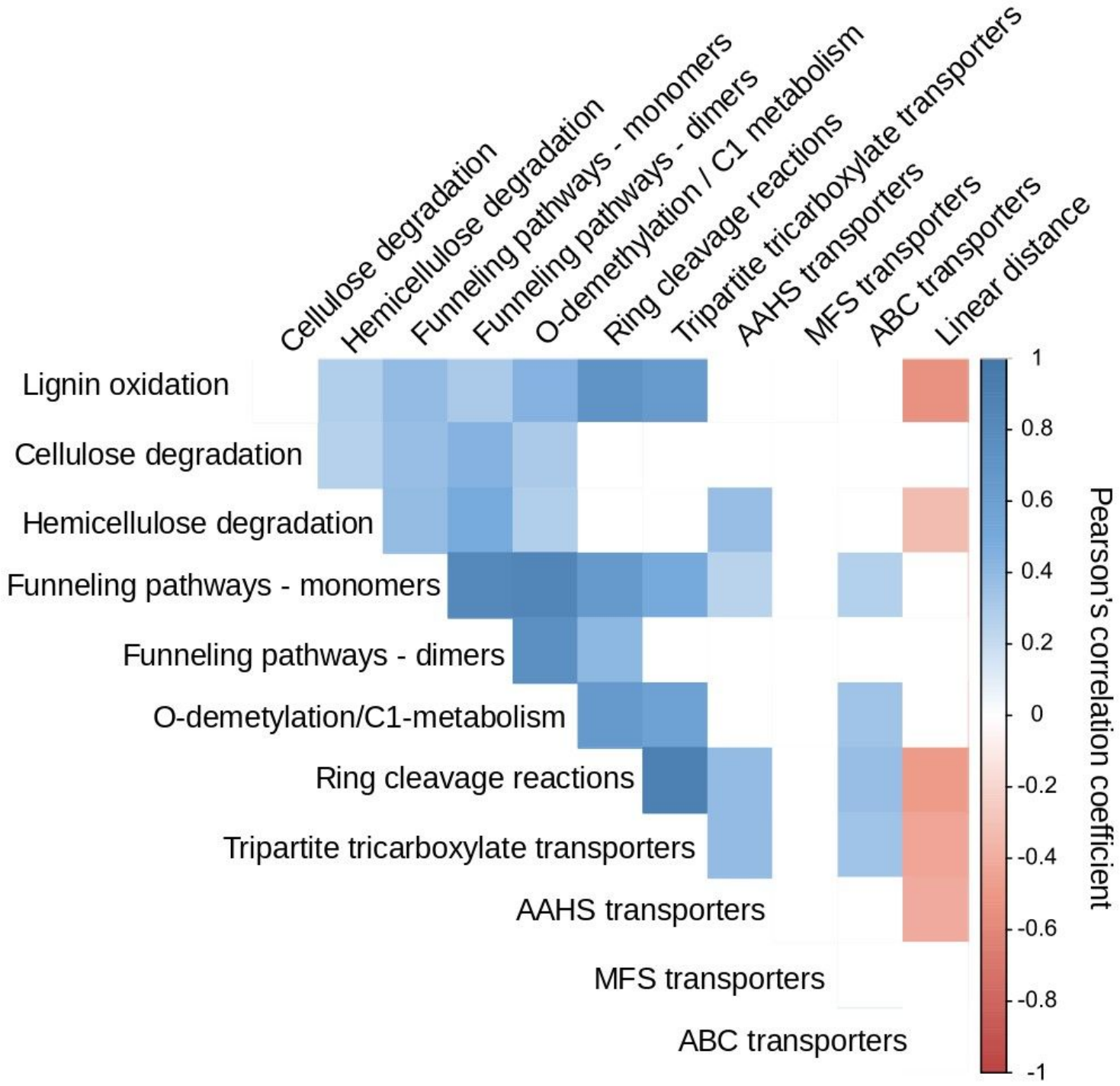
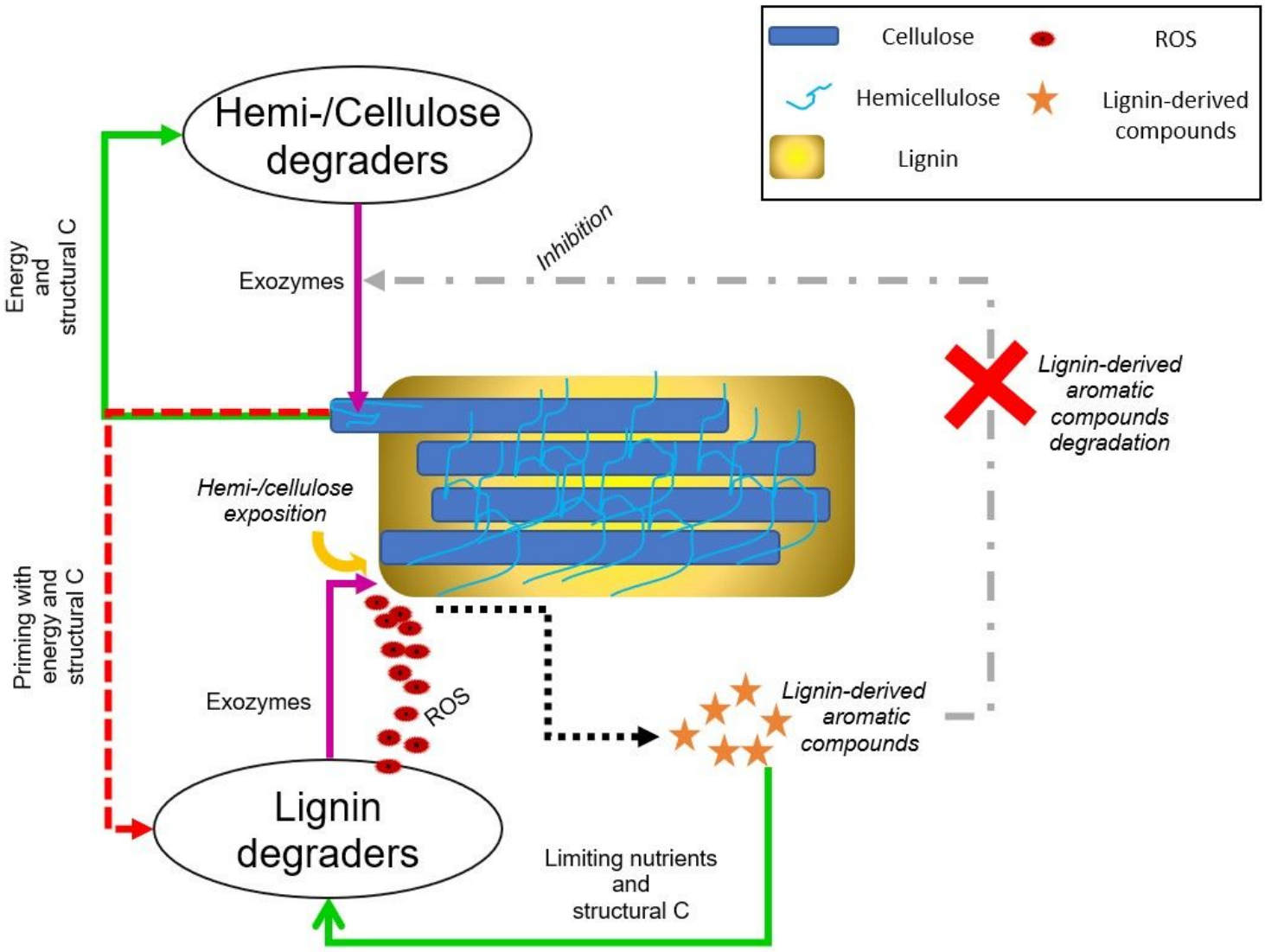



**Figure 5**

Correlations among genes associated to the processing of TeOM and geographic distance in the Amazon River. Correlations between the number of genes associated to lignin oxidation (Lignin oxidation), cellulose and hemicellulose deconstruction (Cellulose and Hemicellulose degradation, respectively), transporting systems (AAHS, MFS, ABC and TTT), lignin-derived aromatic compounds processing pathways (Ring cleavage reactions, O-demethylation/C1-metabolism and Funneling pathways - dimers / monomers), and linear geographic distance using the river source as a starting point (Linear distance).

Color indicates correlation strength (Pearson's correlation coefficient). Only significant correlations ($p < 0.01$) are shown.




## Figure 6

Priming effect model of microbial TeOM degradation in the Amazon River. The cellulolytic communities degrade hemi-/cellulose through secretion of glucosyl hydrolases (mainly GH3/GH10), which release sugars to the environment. These sugars can promote growth in the cellulolytic and lignolytic communities, and during this process, the oxidative metabolism produces reactive oxygen species (ROS). ROS activate the exoenzymes (mainly through 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. [Legend: green arrows – feedback; red dashed arrow – priming effect; black dashed arrow – products; magenta arrows – release of exoenzymes over a substrate; gray arrow – inhibition that cellulolytic organisms suffer from byproducts of lignin oxidation]

# Supplementary Files

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

- Additionalfile1.xlsx