Degradation of terrestrial organic matter by aquatic microbial genomes in the Amazon River

Celio Dias Santos-Junior
Institute of Science and Technology for Brain-Inspired Intelligence – ISTBI, Fudan University, Shanghai, China  https://orcid.org/0000-0002-1974-1736

Ramiro Logares (ramiro.logares@gmail.com)
Institute of Marine Sciences (ICM), CSIC, Barcelona, Catalonia, Spain  https://orcid.org/0000-0002-8213-0604

Flavio Henrique-Silva (dfhs@ufscar.br)
Molecular Biology Laboratory. Department of Genetics and Evolution – DGE, Universidade Federal de São Carlos, UFSCar, São Carlos, São Paulo, Brazil  https://orcid.org/0000-0003-3329-4597

Research Article

Keywords: Amazon River, Microbes, Population genomes, Metagenomics, Terrestrial organic matter, freshwater, TTT

DOI: https://doi.org/10.21203/rs.3.rs-32535/v2

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

Rivers connect the carbon cycle in land with that in aquatic ecosystems by transporting and transforming terrestrial organic matter (TeOM). The Amazon River receives huge loads of TeOM from the surrounding rainforest, promoting a substantial microbial heterotrophic activity and consequently, CO₂ outgassing. In the Amazon River, microbes degrade up to 55% of the lignin present in the TeOM. Yet, the main microbial genomes involved in TeOM degradation were unknown. Here, we characterize 51 Population Genomes (PGs) representing some of the most abundant microbes in the Amazon River deriving from 106 metagenomes. The 51 reconstructed PGs are among the most abundant microbes in the Amazon River, and 53% of them are not able to degrade TeOM. Among the PGs capable of degrading TeOM, 20% were exclusively cellulolytic, while the others could also oxidize lignin. The transport and consumption of lignin oxidation by-products seemed to be decoupled from the oxidation process, being apparently performed by different groups of microorganisms. By connecting the genomic features of abundant microbes in the Amazon River with the degradation machinery of TeOM, we suggest that a microbial consortium following a priming effect model could explain the quick turnover of TeOM previously observed in this ecosystem.

Introduction

Rivers connect land and ocean ecosystems, carrying about 1.9 Pg of organic carbon per year and performing carbon transformations in their course [1]. Only 50% of the carbon present in rivers as terrestrial organic matter (TeOM) is delivered to the oceans [1, 2], indicating that TeOM is actively consumed in rivers [3]. Relative organic carbon respiration rates in rivers tend to decline from headwaters to estuaries, mainly due to increased primary production [4].

The Amazon river basin is the largest freshwater basin in the world, comprising ~38% of continental South America [5]. In comparison to other large rivers, like the Mississippi river, where >50% of the particulate organic carbon comes from algae [6], algal production in the Amazon river is very low [7]. This is explained by the turbidity of the Amazon River, which favours heterotrophic microbial activity [8-11] rather than algal production, turning its waters supersaturated with CO₂ [2]. The TeOM entering this riverine ecosystem comes from the Amazonian rainforest, which is responsible for ~10% of the global primary production [12, 13]. The large amounts of TeOM present in the Amazon River generate ecological niches for microorganisms specialized in complex organic matter degradation [14].

The main components of dissolved TeOM in the Amazon river are lignin and cellulose, accounting for ~60% of all dissolved organic matter [15]. About 60% of the lignin produced in the Amazon rainforest is channelled to the river and continuously decomposed by microbes into monomers, which are subsequently reduced to low molecular weight intermediates and finally remineralized to CO₂. Lignin breakage supports 30-50% of bulk microbial respiration rates in the Amazon River [16]. This leads to CO₂ outgassing from Amazon River waters [17, 18], which releases 1.4 Tg C per year to the atmosphere [19]. In total, <5% of the lignin that the forest produces is stored within the Amazon River basin or delivered to the
ocean [16], suggesting a fast degradation of this compound by the river microbiota. The rapid degradation of recalcitrant organic compounds, such as lignin, boosted by the presence of labile compounds is called priming effect [20]. Evidence of it contributing to accelerate the degradation of TeOM in Amazon River waters was shown by incubation experiments and microbial respiration rates, suggesting that confluence river sections are hotspots of bacterial production with CO₂ levels higher than in other regions [18, 21].

Recently, the Amazon river basin non-redundant microbial gene catalogue (AMnrGC) [22] indicated a zonation in organic matter processing associated to different river sections. The AMnrGC also revealed the main biochemical machinery used by microbes to degrade plant-derived organic matter, which consisted mainly in glycosyl-hydrolases and laccases. Based on this catalogue, we proposed a priming effect model for the Amazon River [22], where two interacting populations (a lignolytic and a cellulolytic one) prime TeOM degradation. Despite the valuable insights that the AMnrGC provided, it is still necessary to link environmental genes to the genomes from where they originate in order to acquire a more holistic understanding of TeOM degradation in the Amazon River. This can help determining whether or not taxa are functionally redundant regarding TeOM degradation, and what proportion of the abundant genomes in the microbiota may carry out functions related to TeOM degradation.

The reconstruction of genomes from metagenomes has become a widely used approach during the last years [23–31], even though the genomes of the most abundant taxa are typically recovered. These so-called Population Genomes (PGs) or Metagenome-Assembled Genomes (MAGs) are crucial to provide genomic context to environmental genes. In addition, PGs may reveal metabolic adaptations or specific gene arrangements [32, 33], being also important to predict the ecological roles played by uncultured microorganisms. PGs have been retrieved from diverse environments [23–29, 34], leading to important findings, such as novel clades [23] as well as new insights on light-harvest mechanisms, nutrients uptake and nitrogen fixation in freshwater microbes [24, 30, 31]. However, to our knowledge, no previous study has extracted PGs from the Amazon River in order to connect them with major biogeochemical cycles, such as the carbon cycle.

Here, we explore 51 abundant PGs extracted from 106 metagenomes retrieved from 30 Amazon River stations in order to address the following questions: What are the PGs functional repertoires to degrade TeOM? Are the systems of lignin oxidation and hemi-/cellulose degradation decoupled? Is the biochemical machinery of lignin-oxidation coupled to the one used for processing lignin-derived aromatic monomers and dimers?

**Material And Methods**

*Metagenome curation and assembly*

We analysed 106 metagenomes from 30 stations in the Amazon River originating from previous studies [14, 35–38]. Metagenomes included Free-living (FL; 0.2 to 2-5 µm) and Particle-Attached (PA; 2 – 5 to 300
µm) size fractions, covering 5 river sections (Fig. 1a): upstream (U), downstream (D), estuary (E), plume (P) and coastal ocean (O), as we previously defined [22]. Size fractions were not compared in depth, as it was shown that there is no significant difference in their biochemical pathways of TeOM degradation [22]. Metagenomes were downloaded from the Sequence Read Archive (SRA) [projects SRP044326, SRP039390] and from the European Nucleotide Archive (ENA) [project PRJEB25171]. Reads were checked for adapters and poor quality bases (Q < 24) using Cutadapt [39]. Subsequently, metagenomes were co-assembled with MEGAHIT (v1.0) [40] using meta-large parameters. Co-assemblies were carried out in 30 groups of metagenomes that were defined according to their geographic origin (Supplementary Table S1). See more details about the co-assembly in [22]. Only contigs with > 1 Kbp in length were considered in downstream analyses.

**Binning and delineation of PGs**

Quality-filtered reads were backmapped against contigs using BWA (version 0.7.12-r1039) [41] and resulting sam-bam files were processed with SamTools (version 1.3.1) [42]. Contigs from each group were then binned with Metabat [43] (v2.12.1) using “superspecific” settings. Contig outliers in terms of Kmer and GC composition were eliminated using RefineM [23] (version 0.0.23) with default settings. Refined bins were assessed for completeness, contamination, strain heterogeneity and taxonomy (“lineage_wf” and “ssu_finder” modules) using CheckM (version 1.0.11) [44]. The 16S rRNA-gene sequences extracted from PGs were classified by mapping them against the SILVA SSU Ref NR99 database (version 123) [45, 46] using Usearch (version 9.2) [47], with an identity cut-off of 97% and a query coverage >70%. Contigs were removed from a bin if they displayed >98% of identity to a hit in the SILVA database that was incongruent with the taxonomic classification obtained via CheckM [44] “tree” mode, which uses functional-gene markers to assign taxonomy.

As previously suggested [23, 48], only bins with genome completeness >50%, contamination <10%, estimated quality ≥ 50 (quality = completeness – 5X contamination [23]) and consisting of ≤500 contigs with N50 ≥ 10 Kbp were used in downstream analyses (Supplementary Fig. S1). The Average Amino-acid Identity (AAI) between PGs was evaluated with CompareM v.0.0.13 [49] using the “aai_wf” function. PGs with AAI ≥ 99.5% were considered redundant. Pairs of redundant genomes were assessed and the most complete, with lowest contamination and strain heterogeneity, was used in downstream analyses.

**PGs similarity analysis**

To check if any of the 51 recovered PGs formed conspecific strains with previously reported genomes a similarity analysis was carried out. PGs were compared, using the Average Nucleotide Identity (ANI) as implemented in FastANI [50], against 957 PGs from the TARA Oceans expedition [51], 18 Verrucomicrobial genomes from diverse freshwater reservoirs [30], 35 PGs from lake Baikal [24], 2 PGs from freshwater *Synechococcus* [31], 3 087 uncultivated bacteria and archaea (UBA) genomes from the Genomes Taxonomy Database – GTDB [52] and a collection of 7 520 high-quality, complete, reference genomes from the National Centre for Biotechnology Information – NCBI (https://www.ncbi.nlm.nih.gov/). PGs displaying a similarity >96.5% in terms of ANI (considering an
aligned fraction (AF) >60%) with another known genome were kept, and the probability of being conspecific strains (p) was calculated [53] and reported when the difference was non-significant (p ≥ 0.9).

Estimation of PG abundance

Quality-filtered reads from 106 metagenomes were backmapped to curated PGs using BWA (v0.7.12-r1039) [41] together with sambamba [54] (version 0.6.6). After mapping, BAM files were filtered with a custom Perl script (courtesy of Amin Madoui - Genoscope, France), keeping only reads with identity >97% and coverage >80%. The abundance of each PG per metagenome was calculated as the number of reads recruited by it, divided by its size in kilobases and the metagenome size in gigabases (RPKG) [31]. The RPKG abundances were then normalized by log(RPKG+1) and used to generate a Heatmap with the R packages ggplot2 [55], gplots [56] and ColorBrewer [57].

Gene prediction and annotation

Genes were predicted using Prodigal v2.6.3 [58] in “normal” mode. Gene annotation was done using Prokka v1.11 [59]. Predicted genes annotated with Prokka were also annotated with KEGG (Release 2015-10-12) [60], COG (Release 2014) [61] and UniProtKB (Release 2016-08) [62] using Diamond v0.9.17, Blastp mode [63], with query coverage >50%, amino acid identity >45%, e-value <1e-5 and score >50. Genes were also annotated using HMMSearch v3.1b1 [64] against dbCAN (v5) [65], PFAM (v30) [66] and eggNOG (v4.5) [67] databases, using an e-value <1e-5 and mean posterior probability of aligned residues >0.75, with no domain overlaps.

Taxonomic classification

Phylogenetic trees were inferred using 43 concatenated protein marker families (Supplementary Table S2). Proteins were identified and aligned using HMMER v.3.1b1 [64]. Positions present in <50% of taxa or without a common amino acid in ≥25% of taxa were removed. Markers were present as single-copy in ~77% of PGs. The multiple-sequence alignment (MSA) included the concatenated markers from our PGs as well as orthologous dereplicated markers from GTDB [52] and RefSeq/GenBank genomes (release 76) [23]. Database markers were retrieved from the CheckM database [44] using CheckM with option “tree_qa”. Trees were inferred with FastTree v.2.1.7 [68] using the JTT+CAT model, and bootstrap support values were calculated using 100 replicates. Newick trees were visualized using Dendroscope v.3 [69].

TeOM degradation

To investigate the TeOM degradation and supplemental metabolism in the PGs, we analysed the pathways for:

i. TeOM degradation – hydrolysis of hemi-/cellulose and oxidation of lignin: laccases (PF02578), and glycosyl-hydrolases (GH) annotated using dbCAN and PFAM databases.
ii. Degradation of lignin oxidation by-products: using the same methodology and reference genes as in previous work [70].

iii. Alternative carbon source use - Tripartite tricarboxylate transporting (TTT): using the functions TctA (PF01970), TctB (PF07331) and TctC (PF03401).

iv. Carbon storage - Polyhydroxy-butryate and polyhydroxy-alkanoate (PHB/A) synthesis pathway: using the functions phaA (annotated as PF00108 and COG0183), phaB (annotated as PF00106 and COG1028), phaC (annotated as PF07167 and COG3243), phaE (PF09712) and phaR (annotated as PF05233 and COG5394).

**Data Accessibility Statement**

Metagenomes used to construct the Amazon River PGs are available at NCBI-SRA database as projects SRP044326, PRJEB25171 and SRP039390. The PGs' nucleotide sequences were deposited at the European Nucleotide Archive - ENA, in the project: PRJEB25176. The PGs' annotations are available at Zenodo (https://zenodo.org/record/1484510).

**Results**

*Population genomes (PGs) from the Amazon River*

Our dataset contained 106 Amazon River metagenomes from 30 different stations divided into 5 sections: upstream, downstream, estuary, plume and ocean (Fig. 1a). We generated 30 co-assemblies (Supplementary Table S1) and after binning, 54 high-quality PGs were selected featuring ≤ 500 contigs with half of them having >10 Kbp, completeness >50%, contamination <10%, and quality ≥ 50% (quality=completeness – 5 x contamination) [Supplementary Fig. S1]. PGs featuring >99.5% Average Amino-Acid Identity (AAI) were considered redundant and removed (Supplementary Table S3). A total of 51 non-redundant PGs were kept in our dataset, including 49% (25) high-quality and 51% (26) medium-quality PGs (Table 1 and Supplementary Table S4) according to previously used criteria [48]. PGs ranged in size between 0.5 and 7.9 Mbp, the maximum contamination was 6.8%, with ~53% of the PGs displaying a contamination <1% (Supplementary Table S4). In total, we recovered 25 PGs from the upstream section, 9 from downstream, 6 from the estuary, 9 from the plume and 2 from the ocean (Fig. 1a-b). *Proteobacteria* (39% of PGs, considering Alpha-, Beta- and Gamma- subdivisions as well as PGs classified as *Proteobacteria*), *Bacteroidetes* (15.7%) and non-classified bacteria (15.7%) predominated (Fig. 1b). In contrast, *Cyanobacteria* represented only 4% of the recovered PGs (Fig. 1b). Only two archaeal genomes were retrieved: one belonging to *Thaumarchaeota* (from the Upstream section) and another one from *Euryarchaeota* (from the Plume). PG distribution and abundance along the river was heterogeneous (Fig. 2).

We identified 10 PGs that had a high similarity (>97% ANI) and non-significant (p > 0.05) differences to other known genomes (Supplementary Table S5): *Richelia intracellularis*-A with AM_2804, *Trueperella*
pyogenes with AM_0546, Acinetobacter junii with AM_0608, Methylopumilus sp1 with AM_0507 / AM_0219, Coccinistipes sp. with AM_2208, Sphingobium sp2 with AM_1603, the taxonomically unannotated UBA11236 with AM_1606, Xanthomonas fuscans with AM_0519 as well as a few characterized species from Rokubacteria, GWA2-73-35 sp1 with AM_2207. Other PGs from the TARA-Oceans expedition [51] or freshwater environments [24, 30, 31] did not display high similarity to the Amazon PGs. Thus, ~80% of the Amazon PGs had no close genomic relative in databases or published datasets. The lowest taxonomic rank of these PGs was: Kingdom, 14% of PGs; Phylum, 14%; Class, 8%; Order, 16%; Family, 26%; and Genus 4%. The remaining 18% PGs could not be taxonomically assigned.

Cellulose and lignin oxidation

Terrestrial organic matter (TeOM) degradation is a fundamental process in the Amazon River and happens in two steps that are modulated by microbes: first, lignin oxidation mediated by laccases, and second, cellulose degradation mediated by specific glycosyl hydrolase (GH) families. Only 24 PGs out of the recovered 51 were able to degrade TeOM (Fig. 3). Laccases were present in all taxa, except Bacteroidetes. All PGs having laccases also had GHs, suggesting that the systems of hemi-/cellulose degradation and lignin oxidation are coupled. Furthermore, there were few cellulolytic PGs (~20%) that did not have lignin oxidation potential, pointing to two assemblages, one that besides being cellulolytic is also lignolytic, and another one that performs only cellulose degradation. Overall, the PGs with the highest potential for TeOM degradation were AM_0519 (Xanthomonas fuscans), AM_0876 / AM_0936 (both unclassified bacteria), and AM_1603 (Sphingobium sp2), according to our criterion of having a minimum of two protein families related to TeOM degradation, with at least two different genes.

Decoupling lignin-oxidation by-products from TeOM degradation

After lignin oxidation, small aromatic compounds are formed and need to be internalized into the cell via transmembrane transporters to complete lignin degradation. Among the PGs having transporters for lignin oxidation by-products (Supplementary Table S6) only two of them (AM_0630 and AM_0902) were also lignin oxidizers. Thus, the oxidation of lignin performed by lignolytic assemblages seems to be completed by cellulolytic microbes that degrade aromatic by-products.

PGs were analysed also for genes required to process aromatic compounds produced after lignin oxidation (Supplementary Table S7). Only two PGs (AM_0519 and AM_1603) seemed able to both degrade lignin-derived aromatic compounds and oxidize lignin. The PGs potentially able to degrade mono-/di-aryls derived from lignin did not possess genes for cellulose degradation or lignin oxidation. Therefore, there is an apparent decoupling of functions related to the oxidation of cellulose, lignin and the processing of by-products of lignin oxidation. This suggests that different microbial assemblages specialize in each step of the TeOM degradation process (that is, lignin oxidation, degradation of by-products generated by lignin oxidation, and cellulose oxidation).

Alternative carbon sources and carbon storage
TeOM degradation involves the formation of glucose (from cellulose hydrolysis) and various aromatic compounds (from oxidation of lignin and its derivatives); all viable carbon sources. Microorganisms tend to prefer specific carbon sources, like sugars, and in their absence, they metabolize other compounds, such as citrate, to obtain energy and structural carbon. Compounds that are metabolized only in the absence of the preferred carbon sources, such as glucose, are called alternative carbon sources. For an effective carbon flux in aquatic environments, transporter systems present in microbes are crucial to ensure that alternative carbon sources can be used, such as tricarboxylates, mono- and di-aryls generated during lignin oxidation. In the Amazon River, there are two main carbon contributors: The TeOM and the less complex compounds, such as humic acids and tricarboxylates. In particular, tricarboxylates are good examples of alternative carbon sources, being constituted by molecules containing three carboxyl functional groups (-COOH), e.g. citrate. Tripartite tricarboxylate transporters (TTT) use substrate binding proteins to sequestrate their ligands from the extracellular milieu and to import them into the cytoplasm (Fig. 4a).

Only seven PGs appeared to use tricarboxylates via the TTT system (Fig. 4b). The PGs containing the complete TTT system included *Alphaproteobacteria* (AM_0275) as well as *Betaproteobacteria*, mainly from the *Burkholderiales* family. One important characteristic of the TTT system is the specificity of each substrate-binding protein to a certain substrate (Fig. 4a). This promotes a high diversity of tctC genes, which were found to range from tens to hundreds across PGs (Fig. 4b). In contrast, <10 genes appeared to be needed for the membrane attached parts (tctA and tctB) of this system (Fig. 4b). PGs containing a complete TTT system seem incapable of TeOM degradation, except for AM_0630, a *Burkholderiales* member containing laccase and GH8 genes. Interestingly, all PGs containing the TTT system (except AM_0630 and AM_0233) also had the biochemical machinery to process aromatic compounds derived from lignin oxidation.

Bacteria have developed impressive mechanisms to cope with adversity. Fluctuations in the water level, change in the concentration of nutrients and seasonality, are common disturbances in the Amazon river. The production and intracellular accumulation of nutritive polymers, later used to prevent starvation during unfavourable conditions, represent an important trait in multiple microbes. Specific mechanisms, such as carbon storage, are relevant also to understand the flux of carbon in ecosystems. One of the most important carbon storage systems is the polyhydroxy-butyrate (PHB) metabolism performed by a few enzymes (Fig. 4c). PHB biosynthesis enzymes were searched in PGs to evaluate their potential to store carbon via this polymer (Fig. 4d). Almost all PGs displaying the complete PHB pathway (*phaA-C*) (Fig. 4d) included also the TTT system, except for AM_0528 and AM_1603, which were found to be TeOM degraders and did not have the TTT system. Yet, the largest number of genes related to the PHB pathway were found in the TeOM degrader PG AM_1603, a *Sphingobium*. The largest gene diversity was related to the initial steps of PHB biosynthesis (genes *phaA* and *phaB*), which are not crucial for PHB production as they perform non-specific transformations, but ensure monomer availability. However, a few gene variants encoded the last steps performed by the *phaC* gene (Fig. 4d), which is the last and crucial step for PHB formation. The gene *phaR*, a transcription regulator protein also related to the accumulation of PHB, was present in 7 out of 13 PGs presumed to produce PHB (Fig. 4d). Only AM_1111 seemed to
produce other polymers different than PHB, the polyhydroxy-alkanoate/butyrate, as it contains the \textit{phaE} gene that allows this species to produce alternative monomers (Fig. 4d).

**Discussion**

Almost half of the analysed PGs from the Amazon River seemed capable of TeOM degradation (Fig. 3). Among the protein families involved in TeOM degradation, laccases seemed to be present as single copy genes in almost all genomes, except in \textit{Bacteroidetes}. The diversity of these genes was much lower than among soil microorganisms [71], indicating that even though Amazon River PGs can potentially degrade lignin and cellulose, their capability is probably modest when compared to soil microbes. We observed a reduction of the TeOM degradation potential towards the ocean, as the number of PGs containing genes related to that function decreased in downstream and estuary sections of the river (Fig. 1b, 2 and 3). We tested whether these results could reflect a technical artefact, given that metagenomes had a heterogeneous representation in the different river sections (Fig. 2). Specifically, sequencing depth per metagenome decreased towards the ocean, while the number of libraries increased (Fig. 1a; Table S1). Results from these tests (such as comparative read back-mapping) supported a gradual reduction in PG's TeOM degradation capacity towards the ocean. This was also coherent with the low gene diversity present in the TeOM degradation machinery (mostly related to cellulose processing) observed in PGs recovered from plume and ocean zones. Comparable findings were reported in analyses of environmental genes [22]. Overall, even though the analysed PGs do not represent the entire set of genes present in the community, our results point to a selective degradation of TeOM in different river sections. This agrees with the negative correlation between TeOM degradation genes and the linear geographic distance of samples to the Amazon River source in Peru, observed in analyses of the Amazon River gene catalogue [22].

Tricarboxylates are molecules often found in TeOM and humic environments [14, 15, 72] and can be generated during lignin processing [70]. They represent alternative carbon sources and are metabolized after being transferred from the environment to the intracellular milieu. For this transfer, microbes usually recur to the TTT system, which was recently shown to be widespread in the Amazon River [8, 22] and correlated with lignin and hemicellulose degradation [22]. Only a small fraction of our PGs (13.5%) had the TTT system. Six out of 7 PGs containing the TTT system belonged to \textit{Betaproteobacteria}, mainly \textit{Burkholderiales}, agreeing with another study that suggested a predominance of this transporter system in \textit{Betaproteobacteria} [73]. The presence of tens to hundreds of gene variants of \textit{tctC} per PG (Fig. 4) concurs with earlier findings. For example, the genome of \textit{Bordetella pertussis} has 90 \textit{tctC} copies [73]. Given that each \textit{tctC} gene has a high affinity for its substrate [73–75], our results also point to multiple substrates in the Amazon River. In the analysed PGs, the decoupling between the TTT system and the TeOM degradation apparatus suggests that organisms incapable of degrading TeOM may specialize in degrading tricarboxylates. This could reflect a general differentiation in carbon use among Amazon River microbes.
The polyhydroxy-butyrates (PHB) metabolism (Fig. 4b) can be used by microbes to store excess of carbon and avoid starvation when conditions are unfavourable. We found that the complete pathway \((\text{phaA-C})\) tended to be present in organisms with the TTT system, suggesting a coupling between these systems. Most of those PGs were identified as \(\text{Betaproteobacteria}\), suggesting that this group is also important in carbon storage. The gene redundancy found in the PHB biosynthesis was much lower than that in the TTT system, and pathway limiting reactions were performed by enzymes encoded by single genes or few gene variants. This points to a potential pathway disruption in case of gene loss, and also a direction of resources inside the cell. The transcriptional repressor gene \(\text{phaR}\), which coordinates the accumulation of PHB [76], was present in more than half of the Amazon River PGs featuring the PHB pathway (Fig. 4d). This points to a microbial assemblage specialized in accumulating PHB that could represent a sink in the carbon cycle that needs to be further explored.

The \textit{priming effect} in the Amazon River indicates that the main steps of TeOM processing are correlated to different taxa [18, 19, 22]. Our work expands the comprehension of this process by proposing that there are two communities involved in TeOM degradation, one responsible for hemi-/cellulose degradation and another one responsible for lignin degradation. Our data suggest that the biochemical machinery to perform cellulose hydrolysis and lignin oxidation are usually coupled, while both being decoupled in terms of lignin-derived aromatic compounds consumption. We propose that different microbial assemblages act in synchrony to degrade TeOM (Fig. 5). This microbial consortium specialized in TeOM degradation would be composed by two communities: one that is strictly cellulolytic and another one that is also lignolytic (Fig. 5). These assemblages would work together to oxidize lignin via laccases and DYPs and expose hemi-/cellulose, which is degraded mainly by GH3 and GH1 enzymes [22]. The action of this TeOM-degrading consortium, represented by taxa harbouring genes related to lignin oxidation and hemi-/cellulose hydrolysis, provides structural carbon and energy to the entire community, including other generalist species. Given that the biochemical machinery for metabolizing lignin-oxidation by-products is decoupled from the one performing TeOM degradation, TeOM-degrading assemblages may not be able to consume these by-products, leading to their accumulation in the environment. This accumulation can be toxic, and sterically prevent cellulase reactions [77], inhibiting the TeOM-degrading consortium and decreasing its growth (Fig. 5). The finding that some PGs are able to transport and metabolize lignin oxidation by-products, being unable to oxidize lignin or hydrolyse cellulose, suggests that there is another microbial assemblage using alternative carbon sources, such as tricarboxylates, which would explain the presence of the TTT system (Fig. 5). This secondary microbial assemblage would use the by-products of lignin oxidation that inhibit the hemi-/cellulose hydrolysis allowing this process to proceed (Fig. 5). The microbial consortium using alternative carbon sources could be characterized by transporters specialized in tricarboxylates (TTT system) as well as genes related to lignin-derived aromatic compounds degradation. Interestingly, this assemblage also seems capable of intracellular carbon storage via PHB biosynthesis, representing a reversible sink in the carbon cycle.

The taxonomic composition of PGs (Fig. 1b and 2) was coherent with that observed by other authors in the Amazon River using 16S rRNA amplicons [78, 79] and reads binning [10, 11]. The Amazon River’s PGs were dominated by \(\text{Actinobacteria}\) and \(\text{Proteobacteria}\) (39% of PGs) agreeing with other studies [8, 79].
Salinity change appeared as the main factor influencing PGs distributions (Fig. 2), in agreement with previous studies indicating the importance of salinity in structuring the Amazon River microbiota [9, 37]. The upstream river section displayed the most taxonomically diverse PGs, when compared with other sections. This may be related to a deeper sequencing depth per metagenome in the upstream section, although other metagenomes from the downstream section included more libraries per section, compensating the total sequencing coverage. Yet, a shallower sequencing depth across multiple libraries could be translated into recovering more information for abundant taxa, and less information for less abundant counterparts. In turn, concentrating all the sequencing effort in fewer samples could be translated into recovering more information for less abundant taxa, that eventually could lead to high quality PGs. Given that this work uses a heterogeneous metagenome dataset, we cannot control for the different amounts of information recovered for less abundant taxa.

Even though *Synechococcus* was previously reported as an Amazon freshwater dominant phototrophic genus [78], we did not find it among our high quality PGs. One possibility is that *Synechococcus* did not assemble or did not constitute high quality bins due to its microdiversity [80, 81]. Instead, we recovered high quality PGs from *Richelia* and *Anabaena*. *Richelia* was previously identified by reads binning [82], and although it was reported to occupy preferentially the plume section [83], we found it to be abundant in the estuary and ocean sections. Furthermore, the *Anabaena*-related PG AM_0902 was found to be more abundant in downstream and plume sections. This is expected, as photosynthesis increases as the river approaches the ocean, mainly due to a decrease in particulate matter [11, 14, 82, 84].

**Declarations**

**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 (ICM) for his assistance 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 this work, specially Lidia Montiel. CDSJ received a PhD scholarship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq #141112/2016-6). FHS work was supported by Research Productivity grants from CNPq (Process #311746/2017-9). 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). RL was supported by a Ramón y Cajal fellowship (RYC-2013-12554, MINECO, Spain). This work was also supported by the projects INTERACTOMICS (CTM2015-69936-P, MINECO, Spain) and MicroEcoSystems (240904, RCN, Norway) to RL.

**Competing Interests**

Authors declare no competing interests.
References


Table 1. Population genomes (PGs) identified in this study. PGs are described in terms of the river section from where they originate, their GC% content, size in 10^6 bp (Mbp), completeness (C), contamination (Cx), quality classification and taxonomy (lowest possible rank).
<table>
<thead>
<tr>
<th>PG</th>
<th>Section</th>
<th>GC (%)</th>
<th>Size (Mbp)</th>
<th>C (%)</th>
<th>Cx</th>
<th>Classification</th>
<th>Taxonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM_0118</td>
<td>Upstream</td>
<td>37.77</td>
<td>1.35</td>
<td>81.87</td>
<td>0</td>
<td>Medium-Quality Draft</td>
<td>Streptomycetaceae</td>
</tr>
<tr>
<td>AM_0219</td>
<td>Upstream</td>
<td>36.37</td>
<td>1.05</td>
<td>76.13</td>
<td>0.71</td>
<td>Medium-Quality Draft</td>
<td>Methylopumilus</td>
</tr>
<tr>
<td>AM_0226</td>
<td>Upstream</td>
<td>63.28</td>
<td>2.99</td>
<td>98.79</td>
<td>0.34</td>
<td>High-Quality Draft</td>
<td>Rhizobiales</td>
</tr>
<tr>
<td>AM_0228</td>
<td>Upstream</td>
<td>56.89</td>
<td>2.92</td>
<td>97.56</td>
<td>2.42</td>
<td>High-Quality Draft</td>
<td>Burkholderiales</td>
</tr>
<tr>
<td>AM_0233</td>
<td>Upstream</td>
<td>53</td>
<td>2.98</td>
<td>96.26</td>
<td>0</td>
<td>High-Quality Draft</td>
<td>Proteobacteria</td>
</tr>
<tr>
<td>AM_0240</td>
<td>Upstream</td>
<td>60.07</td>
<td>3.02</td>
<td>91.32</td>
<td>0.89</td>
<td>High-Quality Draft</td>
<td>Bacteria</td>
</tr>
<tr>
<td>AM_0244</td>
<td>Upstream</td>
<td>41.41</td>
<td>2.10</td>
<td>93.23</td>
<td>0.76</td>
<td>High-Quality Draft</td>
<td>Gammaproteobacteria</td>
</tr>
<tr>
<td>AM_0256</td>
<td>Upstream</td>
<td>53.45</td>
<td>2.66</td>
<td>88.24</td>
<td>2.15</td>
<td>Medium-Quality Draft</td>
<td>Bacteria</td>
</tr>
<tr>
<td>AM_0268</td>
<td>Upstream</td>
<td>52.62</td>
<td>2.91</td>
<td>86.69</td>
<td>0.42</td>
<td>Medium-Quality Draft</td>
<td>Bacteria</td>
</tr>
<tr>
<td>AM_0275</td>
<td>Upstream</td>
<td>52.7</td>
<td>2.41</td>
<td>72.7</td>
<td>1.41</td>
<td>Medium-Quality Draft</td>
<td>Rhizobiales</td>
</tr>
<tr>
<td>AM_0466</td>
<td>Upstream</td>
<td>41.13</td>
<td>1.03</td>
<td>62.5</td>
<td>0</td>
<td>Medium-Quality Draft</td>
<td>Bacteria</td>
</tr>
<tr>
<td>AM_0507</td>
<td>Upstream</td>
<td>36.3</td>
<td>1.00</td>
<td>79.25</td>
<td>0.64</td>
<td>Medium-Quality Draft</td>
<td>Betaproteobacteria</td>
</tr>
<tr>
<td>AM_0510</td>
<td>Upstream</td>
<td>62.61</td>
<td>2.00</td>
<td>98.08</td>
<td>1.1</td>
<td>High-Quality Draft</td>
<td>Bacteria</td>
</tr>
<tr>
<td>AM_0519</td>
<td>Upstream</td>
<td>65.04</td>
<td>4.82</td>
<td>96.74</td>
<td>0.07</td>
<td>High-Quality Draft</td>
<td>Gammaproteobacteria</td>
</tr>
<tr>
<td>AM_0528</td>
<td>Upstream</td>
<td>48.07</td>
<td>2.08</td>
<td>93.43</td>
<td>0.34</td>
<td>High-Quality Draft</td>
<td>Rhizobiales</td>
</tr>
<tr>
<td>AM_0546</td>
<td>Upstream</td>
<td>58.47</td>
<td>1.47</td>
<td>70.69</td>
<td>0.07</td>
<td>Medium-Quality Draft</td>
<td>Actinobacteria</td>
</tr>
<tr>
<td>AM_0608</td>
<td>Upstream</td>
<td>38.52</td>
<td>3.23</td>
<td>99.28</td>
<td>1.07</td>
<td>High-Quality Draft</td>
<td>Moraxellaceae</td>
</tr>
<tr>
<td>AM_0615</td>
<td>Upstream</td>
<td>36.39</td>
<td>1.56</td>
<td>99.51</td>
<td>0.97</td>
<td>High-Quality Draft</td>
<td>Thaumarchaeota</td>
</tr>
<tr>
<td>AM_0616</td>
<td>Upstream</td>
<td>59.76</td>
<td>5.51</td>
<td>98.66</td>
<td>0.59</td>
<td>High-Quality Draft</td>
<td>Betaproteobacteria</td>
</tr>
<tr>
<td>AM_0619</td>
<td>Upstream</td>
<td>42.57</td>
<td>2.83</td>
<td>91.18</td>
<td>2.52</td>
<td>High-Quality Draft</td>
<td>Bacteria</td>
</tr>
<tr>
<td>AM_0621</td>
<td>Upstream</td>
<td>56.53</td>
<td>7.87</td>
<td>90.33</td>
<td>0</td>
<td>High-Quality Draft</td>
<td>Bacteria</td>
</tr>
<tr>
<td>AM_0630</td>
<td>Upstream</td>
<td>52.99</td>
<td>2.89</td>
<td>95.43</td>
<td>2.17</td>
<td>High-Quality Draft</td>
<td>Proteobacteria</td>
</tr>
<tr>
<td>AM_0643</td>
<td>Upstream</td>
<td>56.96</td>
<td>2.89</td>
<td>91.49</td>
<td>2.2</td>
<td>High-Quality Draft</td>
<td>Burkholderiales</td>
</tr>
<tr>
<td>AM_0729</td>
<td>Upstream</td>
<td>53.85</td>
<td>2.45</td>
<td>84.4</td>
<td>1.26</td>
<td>Medium-Quality Draft</td>
<td>Bacteria</td>
</tr>
<tr>
<td>AM_0764</td>
<td>Upstream</td>
<td>50.2</td>
<td>3.89</td>
<td>86.77</td>
<td>1.1</td>
<td>Medium-Quality Draft</td>
<td>Bacteria</td>
</tr>
<tr>
<td>AM_0832</td>
<td>Downstream</td>
<td>44.61</td>
<td>3.47</td>
<td>96.39</td>
<td>1.64</td>
<td>High-Quality Draft</td>
<td>Bacteroidetes</td>
</tr>
<tr>
<td>AM_0849</td>
<td>Downstream</td>
<td>65.31</td>
<td>2.11</td>
<td>86.21</td>
<td>6.78</td>
<td>Medium-Quality Draft</td>
<td>Actinobacteria</td>
</tr>
<tr>
<td>AM_0854</td>
<td>Downstream</td>
<td>52.59</td>
<td>0.53</td>
<td>73.22</td>
<td>0</td>
<td>Medium-Quality Draft</td>
<td>Bacteria</td>
</tr>
<tr>
<td>Accession</td>
<td>Location</td>
<td>Downstream</td>
<td>52.57</td>
<td>4.47</td>
<td>90.28</td>
<td>2.88</td>
<td>High-Quality Draft</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>--------</td>
<td>-------</td>
<td>--------</td>
<td>-------</td>
<td>-------------------</td>
</tr>
<tr>
<td>AM_0902</td>
<td>Downstream</td>
<td>37.07</td>
<td>4.06</td>
<td>96.49</td>
<td>0</td>
<td></td>
<td>High-Quality Draft</td>
</tr>
<tr>
<td>AM_0936</td>
<td>Downstream</td>
<td>62.13</td>
<td>2.69</td>
<td>86.63</td>
<td>4.45</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_1003</td>
<td>Downstream</td>
<td>59.74</td>
<td>6.26</td>
<td>95.93</td>
<td>0.91</td>
<td></td>
<td>High-Quality Draft</td>
</tr>
<tr>
<td>AM_1104</td>
<td>Downstream</td>
<td>59.11</td>
<td>3.00</td>
<td>78.96</td>
<td>3.08</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_1111</td>
<td>Downstream</td>
<td>38.28</td>
<td>2.82</td>
<td>90.95</td>
<td>6.68</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_1205</td>
<td>Estuary</td>
<td>47.76</td>
<td>0.71</td>
<td>65.39</td>
<td>0</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_1312</td>
<td>Estuary</td>
<td>61.53</td>
<td>2.17</td>
<td>85.71</td>
<td>1.82</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_1409</td>
<td>Estuary</td>
<td>39.04</td>
<td>0.61</td>
<td>70.98</td>
<td>1.72</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_1503</td>
<td>Estuary</td>
<td>67.21</td>
<td>2.09</td>
<td>63.57</td>
<td>0.73</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_1603</td>
<td>Estuary</td>
<td>63.07</td>
<td>4.54</td>
<td>98.42</td>
<td>1.19</td>
<td></td>
<td>High-Quality Draft</td>
</tr>
<tr>
<td>AM_1606</td>
<td>Estuary</td>
<td>55.47</td>
<td>2.24</td>
<td>96.4</td>
<td>0.61</td>
<td></td>
<td>High-Quality Draft</td>
</tr>
<tr>
<td>AM_1801</td>
<td>Plume</td>
<td>45.14</td>
<td>1.20</td>
<td>67.56</td>
<td>0</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_1811</td>
<td>Plume</td>
<td>43.37</td>
<td>2.27</td>
<td>85.47</td>
<td>3.72</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_2104</td>
<td>Plume</td>
<td>60.62</td>
<td>2.11</td>
<td>84.85</td>
<td>4.55</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_2116</td>
<td>Plume</td>
<td>46.6</td>
<td>0.69</td>
<td>52.31</td>
<td>0</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_2124</td>
<td>Plume</td>
<td>40.54</td>
<td>1.53</td>
<td>87.19</td>
<td>1.49</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_2202</td>
<td>Plume</td>
<td>47.45</td>
<td>1.80</td>
<td>97.31</td>
<td>0</td>
<td></td>
<td>High-Quality Draft</td>
</tr>
<tr>
<td>AM_2207</td>
<td>Plume</td>
<td>43.09</td>
<td>1.79</td>
<td>85.25</td>
<td>1.19</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_2208</td>
<td>Plume</td>
<td>53.01</td>
<td>1.62</td>
<td>92.99</td>
<td>0.36</td>
<td></td>
<td>High-Quality Draft</td>
</tr>
<tr>
<td>AM_2324</td>
<td>Plume</td>
<td>37.43</td>
<td>3.41</td>
<td>90.78</td>
<td>0.99</td>
<td></td>
<td>High-Quality Draft</td>
</tr>
<tr>
<td>AM_2502</td>
<td>Ocean</td>
<td>39.43</td>
<td>1.43</td>
<td>89.99</td>
<td>1.18</td>
<td></td>
<td>Medium-Quality Draft</td>
</tr>
<tr>
<td>AM_2804</td>
<td>Ocean</td>
<td>33.5</td>
<td>3.29</td>
<td>93.56</td>
<td>0.27</td>
<td></td>
<td>High-Quality Draft</td>
</tr>
</tbody>
</table>

**Figures**
Figure 1

Metagenomes and Population Genomes (PGs) from the Amazon River. Distribution of the 106 metagenomes along the five sections of the Amazon River: Upstream, Downstream, Estuary, Plume and Ocean, from where the PGs derive (a). The size of the red circles indicates the sequencing coverage per metagenome (Cov./Gbp). Number of curated PGs that were reconstructed per river section (b).

Figure 2
Abundance of PGs along Amazon River sections. PG abundances are indicated as Log(RPKG+1). Note that few PGs are ubiquitous while most of them are concentrated in specific river sections.

**Figure 3**

Presence of cellulose degradation functions in the Amazon River PGs. The y-axis shows main glucosyl hydrolase (GHs) families known to participate in cellulose degradation. The x-axis shows the PGs. The size of the bubble indicates number of genes per protein family.
Figure 4

Pathways involved in tricarboxylates usage and carbon storage. Amazon River PGs were analysed for their potential to use tricarboxylates a) and only genomes containing the complete TTT system b) are reported (IM – inner membrane; OM – outer membrane). The polyhydroxy-butyrate/alkanoate production was further investigated c) to assess the potential for carbon storage in Amazon River PGs. Those PGs with the potential to store carbon are shown in d), indicating the number of different Pha genes. NB: the enzyme phaE allows in the presence of the enzyme phaC, the biosynthesis of polyhydroxy-alkanoate/butyrate, a hybrid biopolymer (Poly3HB-co-4HB). The protein phaR regulates the accumulation of polyhydroxy-butyrate in granules inside the cell.
Figure 5

Genome-based priming effect model for the Amazon River. Green arrows indicate beneficial effects for assemblages in terms of cell growth, blue arrows represent the secretion of ectoenzymes and the degradation of TeOM, while the black arrow indicates lignin oxidation generating low molecular weight aromatic by-products. The red arrow indicates growth inhibition and the beige arrow its suppression. In this model, three microbial assemblages interact. Two of them are responsible for exposing and processing TeOM. Besides providing structural carbon and energy for cell growth, by-products of this metabolism inhibit TeOM degradation. The third assemblage prevents TeOM degradation to stop by consuming these by-products and storing the resulting carbon intracellularly.

Supplementary Files

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

- FigureS1.png
- SupplementaryTablesISMEJ.xlsx