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 not known. 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 byproducts seemed to be decoupled from the oxidation process, being apparently performed by different groups of microorganisms. Altogether, based on our findings, we suggest a new priming effect model that explains the quick turnover of TeOM as a product of the microbial consumption of lignin-derived aromatic compounds produced by lignin oxidation, reducing the inhibition of cellulose degradation and ensuring structural carbon and energy for cell growth. By connecting the genomic features of abundant microbes in the Amazon River with the degradation of recalcitrant TeOM, we contribute to increase our understanding of the rapid consumption of recalcitrant compounds 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 favors 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 channeled 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 ones 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 confluence river sections as 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, a priming effect model was proposed 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. The previous 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.

During the last 5 years, the reconstruction of partial or full genomes from metagenomes has become a widely used approach (23–31), even though in most cases, only the genomes of the most abundant taxa are recovered. These so-called population genomes (PGs) or Metagenome-Assembled Genomes (MAGs) are crucial to link environmental genes with the genomes from where they originate. 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), 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 tried to extract PGs from the Amazon River and connect them with TeOM degradation.

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?

Results

Population genomes (PGs) from the Amazon River

Our original 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 (Table S1) and after binning, 54 high-quality PGs were selected featuring ≤500 contigs with half of them >10 Kbp, completeness >50%, contamination <10%, and quality ≥50% (quality=completeness − 5 x contamination)
[Figure S1]. PGs featuring >99% of Average Amino-Acid Identity (AAI) were considered redundant and removed (Table S2). 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 Table S3) according to previously used criteria (34). 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% (Table S3). 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-, 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 (Table S4): Richelia intracellularis-A with AM_2804, Trueperella pyogenes with AM_0546, Acinetobacter junii with AM_0608, Methylotupulimus 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 (35) 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. These PGs had the lowest taxonomic rank assigned at the level of kingdom (14% of PGs), phylum (14%), class (8%), order (16%), family (26%) or 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 hydrolases (GHs) families. More than half of our PGs (~53 %) did not possess the ability to degrade TeOM, while only 24 PGs were able to degrade TeOM (Fig. 3). Laccases were present in all taxa, except Bacteroidetes. All PGs displaying laccases also displayed GHs, suggesting the systems of hemi-/cellulose degradation and lignin oxidation are coupled. Furthermore, there were few cellulolytic PGs (~20%) that did not display 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 byproducts 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 byproducts (Table S5) 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 byproducts.

PGs were analyzed also for genes required to process aromatic compounds produced after lignin oxidation (Table S6). 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 and lignin as well as functions associated to processing byproducts of lignin oxidation. The previous points to different assemblages specialized in each step of the TeOM degradation process (that is, lignin oxidation, degradation of byproducts 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 preferred carbon sources, such as glucose, are called alternative carbon sources. For an effective carbon flux in aquatic environments, transporting 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 as well as 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 portions (tctA and tctB) of this system (Fig. 4b). PGs containing a complete TTT systems seem uncappable 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 levels, 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 unfavorable conditions, represent an important trait in multiple microbes. In particular, specific mechanisms, such as carbon storage are relevant also to understand the flux of carbon inside 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 representative. The largest gene diversity was observed to be related to the initial steps of PHB biosynthesis (genes phaA and phaB), 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 was presumed to produce other polymers than PHB, the polyhydroxy-alkanoate/butyrate, as it contains the phaE gene that allows this species to produce alternative monomers (Fig. 4d).

**Discussion**

Almost half of the analyzed 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 a single copy gene in almost all genomes, except in Bacteroidetes. Overall, the diversity of these genes was much lower than among soil microorganisms (36), indicating that despite Amazon River PGs could potentially degrade lignin and cellulose, this capability is probably modest. 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 (Fig. 1b, 2 and 3). We tested whether these results could reflect a technical artifact, 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; Additional file 1, Table S1). Results from these tests (such as comparative reads mapping) supported a gradual reduction in PG’s TeOM degradation capacity towards the ocean. This was also supported by the low gene diversity present in the TeOM degradation machinery (mostly related to cellulose processing) observed in PGs recovered from plume and ocean zones. Similar finding were reported in analyses of community 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 is coherent with the negative correlation between TeOM degradation genes and the linear geographical distance of samples to the Amazon River source in Peru that was observed in analyses of the Amazon River gene catalogue (22).

Tricarboxylates are molecules often found in TeOM and humic environments (14, 15, 37, 38) and can be generated during lignin processing (39). 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 to 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 Betaproteobacteria, mainly Burkholderiales, agreeing with another study that suggested a predominance of this transporting system in Betaproteobacteria (40).
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 (40). Given that each \textit{tctC} gene has a high affinity for its substrate (40–42), our results also point to multiple substrates in the Amazon River. In the analyzed PGs, the decoupling between the TTT system and the TeOM degradation apparatus suggests that non-degrading TeOM organisms 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 future starvation. We found that the complete pathway (\textit{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 \textit{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, where 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. Interestingly, the transcriptional repressor gene \textit{phaR}, which coordinates the accumulation of PHB (43), was present in more than half of the Amazon River PGs with the PHB pathway (Fig. 4d). The previous points to a microbial assemblage specialized in accumulating PHB. This assemblage could potentially represent a carbon sink in the carbon cycle that needs to be further explored.

The idea of a \textit{priming effect} in the Amazon River, previously reported by other authors (18, 19, 22), indicates that the main steps of TeOM processing are correlated to different taxa. The present work expands the comprehension of this previous model. Specifically, we propose 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. In the light of the results shown here, 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 work together to oxidize lignin via laccases and DYPs and expose hemi-/cellulose, which is degraded mainly by GH3 and GH1 enzymes, as previously suggested (22). The action of this TeOM-degrading consortium, represented by taxa harboring 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 byproducts is decoupled from the one used to perform TeOM degradation, TeOM-degrading assemblages may not be able to consume these byproducts, leading to their accumulation in the environment. This accumulation can be toxic, and sterically prevent cellulase reactions (44), inhibiting the TeOM-degrading consortium and decreasing its cell growth (Fig. 5). The finding that some PGs are able to transport and metabolize lignin oxidation byproducts, but are unable to oxidize lignin or hydrolyze cellulose, supports the idea 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 byproducts of lignin oxidation that inhibit the hemi-/cellulose hydrolysis allowing this process to continue (Fig. 5). The microbial consortium using alternative carbon sources could be characterized by a transporting system 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 (45), miTags (46), or reads binning (10, 11). The Amazon River PGs were dominated by Actinobacteria and Proteobacteria (39% of PGs) agreeing with other studies (8, 45). 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, 47). 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 bins. 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 (46), 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 (48, 49). Instead, we recovered high quality PGs from Richelia and Anabaena. Richelia was previously identified by reads binning (50), and although it was reported to occupy preferentially the plume section (51), 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, 50, 52).

Material And Methods

Metagenome curation and assembly

We analysed 106 metagenomes from 30 stations in the Amazon River originating from previous studies (14, 47, 53–55). 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 used (22). Size fractions were not discussed in depth during this study, 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 (56). Subsequently, metagenomes were co-assembled with MEGAHIT (v1.0) (57) using meta-large parameters. Co-assemblies were carried out in 30 groups of metagenomes that were defined according to their geographic origin (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) (58) and resulting sam/bam files were processed with and SamTools (version 1.3.1) (59). Contigs from each group were then binned with Metabat (60) (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 (61) (version 1.0.11). The 16S rRNA-gene sequences extracted from PGs were classified by mapping them against the SILVA SSU Ref NR99 database (version 123) (62, 63) using Usearch (64) (version 9.2), 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 (61) “tree” mode, which uses functional-gene markers to assign taxonomy.

As previously suggested (23, 34), 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 (Figure S1). The Average Amino-acid Identity (AAI) between PGs was evaluated with CompareM (65) v.0.0.13 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 (66), against 957 PGs from the TARA Oceans expedition (35), 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 (67) 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 (68) 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) (58) together with sambamba (69) (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 per gigabases of metagenome (RPKG) (31). The RPKG abundances were then
normalized by log(RPKG+1) and used to generate a Heatmap with the R packages ggplot2 (70), gplots (71) and ColorBrewer (72).

**Gene prediction and annotation**

Genes were predicted using Prodigal (v2.6.3) (73) in “normal” mode. Gene annotation was done using Prokka (74) (v1.11). Predicted genes annotated with Prokka were also annotated with KEGG (Release 2015-10-12) (75), COG (Release 2014) (76) and UniProtKB (Release 2016-08) (77) using Diamond (Blastp mode) (78) v0.9.17, with query coverage >50%, amino acid identity >45%, e-value <1e-5 and score >50. Genes were also annotated using HMMSearch (79) v3.1b1 against dbCAN (v5) (80), PFAM (v30) (81) and eggNOG (v4.5) (82) 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 (**Table S7**). Proteins were identified and aligned using HMMER v.3.1b1 (79). 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 (67) and RefSeq/GenBank genomes (release 76) (23). Database markers were retrieved from the CheckM database (61) using CheckM with option “tree_qa”. Trees were inferred with FastTree (83) v.2.1.7 using the JTT+CAT model and bootstrap support values were calculated using 100 replicates. Newick trees were visualized using Dendroscope (v3) (84).

**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 byproducts: using the same methodology and reference genes as in previous work (39);

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

iv. Carbon storage - Polyhydroxy-butyrate 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). Those PGs containing potentially complete pathways were reported.

**Data availability**

Metagenomes used to construct the Amazon River PGs are available at NCBI-SRA: 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).

Declarations

Acknowledgements

CDJS, FHS and RL conceived the study. CDSJ and RL analyzed the data. CDSJ wrote the initial draft and all authors wrote and revised the paper. All authors approved the final manuscript. 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. C.D.S.J. was the recipient of a PhD scholarship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq #141112/2016-6). F.H.S. 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). R.L. 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 R.L.

Competing Interests

Authors declare no competing interests.

References

1. Cole JJ, Prairie YT, Caraco NF, McDowell WH, Tranvik LJ, Striegl RG, Duarte CM, Kortelainen P, Downing JA, Middelburg JJ, Melack J. 2007. Plumbing the Global Carbon Cycle: Integrating Inland Waters into the Terrestrial Carbon Budget. Ecosystems 10:172–185.
2. Mayorga E, Aufdenkampe AK, Masiello CA, Krusche AV, Hedges JL, Quay PD, Richey JE, Brown TA. 2005. Young organic matter as a source of carbon dioxide outgassing from Amazonian rivers. Nature 436:538.
3. Hernes PJ, Benner R. 2003. Photochemical and microbial degradation of dissolved lignin phenols: implications for the fate of terrigenous dissolved organic matter in marine environments. Journal of Geophysical Research: Oceans 108:3291.
4. Battin TJ, Kaplan LA, Findlay S, Hopkinson CS, Marti E, Packman AI, Newbold JD, Sabater F. 2008. Biophysical controls on organic carbon fluxes in fluvial networks. Nature Geoscience 1:95–100.
5. Mikhailov VN. 2010. Water and sediment runoff at the Amazon River mouth. Water Resources 37:145–159.
6. Kendall C, Steven RS, Kelly VJ. 2001. Carbon and nitrogen isotopic compositions of particulate organic matter in four large river systems across the Unitated States. Hydrological Processes 15:1301–1346.
7. Wissmar RC, Richey JE, Stallard RF, Edmond JM. 1981. Plankton Metabolism and Carbon Processes in the Amazon River, Its Tributaries, and Floodplain Waters, Peru-Brazil, May-June 1977. Ecology 62:1622–1633.

8. Ghai R, Rodriguez-Valera F, McMahon KD, Toyama D, Rinke R, de Oliveira TCS, Garcia JW, de Miranda FP, Henrique-Silva F. 2011. Metagenomics of the water column in the pristine upper course of the Amazon river. PLoS ONE 6:e23785.

9. Satinsky BM, Crump BC, Smith CB, Sharma S, Zielinski BL, Doherty M, Meng J, Sun S, Medeiros PM, Paul JH, Coles VJ, Yager PL, Moran MA. 2014. Microspatial gene expression patterns in the Amazon River Plume. Proceedings of the National Academy of Sciences of the United States of America 111:11085–90.

10. Doherty M, Yager PL, Moran MA, Coles VJ, Fortunato CS, Krusche A V., Medeiros PM, Payet JP, Richey JE, Satinsky BM, Sawakuchi HO, Ward ND, Crump BC. 2017. Bacterial Biogeography across the Amazon River-Ocean Continuum. Frontiers in Microbiology 8:882.

11. Satinsky BM, Smith CB, Sharma S, Ward ND, Krusche A V., Richey JE, Yager PL, Crump BC, Moran MA. 2017. Patterns of Bacterial and Archaeal Gene Expression through the Lower Amazon River. Frontiers in Marine Science 4:253.

12. Field, Behrenfeld, Randerson, Falkowski. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. Science (New York, NY) 281:237–40.

13. Malhi Y, Roberts JT, Betts RA, Killeen TJ, Li W, Nobre CA. 2008. Climate change, deforestation, and the fate of the Amazon. Science 319:169–72.

14. Satinsky BM, Fortunato CS, Doherty M, Smith CB, Sharma S, Ward ND, Krusche A V., Yager PL, Richey JE, Moran MA, Crump BC. 2015. Metagenomic and metatranscriptomic inventories of the lower Amazon River, May 2011. Microbiome 3:39.

15. Ertel JR, Hedges JI, Devol AH, Richey JE, Ribeiro M de NG. 1986. Dissolved humic substances of the Amazon River system1. Limnology and Oceanography 31:739–754.

16. Ward ND, Keil RG, Medeiros PM, Brito DC, Cunha AC, Dittmar T, Yager PL, Krusche AV, Richey JE. 2013. Degradation of terrestrially derived macromolecules in the Amazon River. Nature Geoscience 6:530–533.

17. Ward ND, Keil RG, Medeiros PM, Brito DC, Cunha AC, Dittmar T, Yager PL, Krusche A V., Richey JE. 2013. Degradation of terrestrially derived macromolecules in the Amazon River. Nature Geoscience 6:530–533.

18. Ward ND, Bianchi TS, Sawakuchi HO, Gagne-Maynard W, Cunha AC, Brito DC, Neu V, de Matos Valerio A, da Silva R, Krusche A V., Richey JE, Keil RG. 2016. The reactivity of plant-derived organic matter and the potential importance of priming effects along the lower Amazon River. Journal of Geophysical Research: Biogeosciences 121:1522–1539.

19. Sawakuchi HO, Neu V, Ward ND, Barros M de LC, Valerio AM, Gagne-Maynard W, Cunha AC, Less DFS, Diniz JEM, Brito DC, Krusche A V., Richey JE. 2017. Carbon Dioxide Emissions along the Lower Amazon River. Frontiers in Marine Science 4:76.

20. Bianchi TS. 2011. 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 108:19473–19481.
21. Farjalla VF. 2014. Are the mixing zones between aquatic ecosystems hot spots of bacterial production in the Amazon River system? Hydrobiologia 728:153–165.

22. Santos-Junior CD, Sarmento H, Miranda FP, Henrique-Silva F, Logares R. 2020. Uncovering the genomic potential of the Amazon River microbiome to degrade rainforest organic matter. PREPRINT (Version 1) available at Research Square [DOI: 10.21203/rs.3.rs-25138/v1]

23. Parks DH, Rinke C, Chuvочкина M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology 2:1533–1542.

24. Cabello-Yeves PJ, Zemskaya TI, Rosselli R, Coutinho FH, Zakharenko AS, Blinov V V., Rodriguez-Valera F. 2018. Genomes of Novel Microbial Lineages Assembled from the Sub-Ice Waters of Lake Baikal. Applied and Environmental Microbiology 84:e02132-17.

25. Tully BJ, Sachdeva R, Graham ED, Heidelberg JF. 2017. 290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology. PeerJ 5:e3558.

26. Hugerth LW, Larsson J, Alneberg J, Lindh M V., Legrand C, Pinhassi J, Andersson AF. 2015. Metagenome-assembled genomes uncover a global brackish microbiome. Genome Biology 16:279.

27. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF. 2009. Community-wide analysis of microbial genome sequence signatures. Genome Biology 10:R85.

28. Baker BJ, Lazar CS, Teske AP, Dick GJ. 2015. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. Microbiome 3:14.

29. Ghai R, Pašić L, Fernández AB, Martin-Cuadrado A-B, Mizuno CM, McMahon KD, Papke RT, Stepanauskas R, Rodriguez-Brito B, Rohwer F, Sánchez-Porro C, Ventosa A, Rodríguez-Valera F. 2011. New Abundant Microbial Groups in Aquatic Hypersaline Environments. Scientific Reports 1:135.

30. Cabello-Yeves PJ, Ghai R, Mehrshad M, Picazo A, Camacho A, Rodriguez-Valera F. 2017. Reconstruction of Diverse Verrucomicrobial Genomes from Metagenome Datasets of Freshwater Reservoirs. Frontiers in Microbiology 8:2131.

31. Cabello-Yeves PJ, Haro-Moreno JM, Martin-Cuadrado A-B, Ghai R, Picazo A, Camacho A, Rodriguez-Valera F. 2017. Novel Synechococcus Genomes Reconstructed from Freshwater Reservoirs. Frontiers in Microbiology 8:1151.

32. Muller EEL, Pinel N, Laczny CC, Hoopmann MR, Narayanasamy S, Lebrun LA, Roume H, Lin J, May P, Hicks ND, Heintz-Buschart A, Wampach L, Liu CM, Price LB, Gilleece JD, Guignard C, Schupp JM, Vlassis N, Baliga NS, Moritz RL, Keim PS, Wilmes P. 2014. Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. Nature Communications 5:5603.

33. Georges AA, El-Swais H, Craig SE, Li WK, Walsh DA. 2014. Metaproteomic analysis of a winter to spring succession in coastal northwest Atlantic Ocean microbial plankton. The ISME Journal 8:1301–1313.

34. Bowers RM, Kyrpides NC, Stepanauskas R, Hamon-Smith M, Doud D, Reddy TBK, Schulz F, Jaret J, Rivers AR, Eloé-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Bercraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooshef S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu W-T, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I,
Tyson GW, Rinke C, Kyrpides NC, Schriml L, Garrity GM, Hugenholtz P, Sutton G, Yilmaz P, Meyer F, Glöckner FO, Gilbert JA, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature Biotechnology 35:725–731.

35. Eren AM. 2017. TARA-NON-REDUNDANT-MAGs. figshare.

36. López-Mondéjar R, Zühlke D, Becher D, Riedel K, Baldrian P. 2016. Cellulose and hemicellulose decomposition by forest soil bacteria proceeds by the action of structurally variable enzymatic systems.

37. Westergaard B, Hansen HCB. 1998. Determination of anions in soil solutions by capillary zone electrophoresis. Analyst 123:721–724.

38. Gerke J. 2018. Carboxylic Acid - Key Role in Life Sciences.

39. Kamimura N, Takahashi K, Mori K, Araki T, Fujita M, Higuchi Y, Masai E. 2017. Bacterial catabolism of lignin-derived aromatics: New findings in a recent decade: Update on bacterial lignin catabolism. Environmental Microbiology Reports 9:679–705.

40. Antoine R, Jacob-Dubuisson F, Drobecq H, Willery E, Lesjean S, Locht C. 2003. Overrepresentation of a gene family encoding extracytoplasmic solute receptors in Bordetella. Journal of bacteriology 185:1470–4.

41. Antoine R, Huvent I, Chemlal K, Deray I, Raze D, Locht C, Jacob-Dubuisson F. 2005. The Periplasmic Binding Protein of a Tripartite Tricarboxylate Transporter is Involved in Signal Transduction. Journal of Molecular Biology 351:799–809.

42. Winnen B, Hvorup RN, Saier MH. 2003. The tripartite tricarboxylate transporter (TTT) family. Research in Microbiology 154:457–465.

43. Quelas JI, Mesa S, Mongiardini EJ, Jendrossek D, Lodeiro AR. 2016. Regulation of Polyhydroxybutyrate Synthesis in the Soil Bacterium Bradyrhizobium diazoefficiens. Applied and environmental microbiology 82:4299–308.

44. Qin L, Li W-C, Liu L, Zhu J-Q, Li X, Li B-Z, Yuan Y-J. 2016. Inhibition of lignin-derived phenolic compounds to cellulase. Biotechnology for Biofuels 9:70.

45. Toyama D, Santos-Júnior CD, Kishi LT, Oliveira TCS, Garcia JW, Sarmento H, Miranda FP, Henrique-Silva F. 2017. A snapshot on prokaryotic diversity of the Solimões River basin (Amazon, Brazil). Genetics and Molecular Research 16.

46. Reis MC dos, Bagatini IL, Vidal L de O, Bonnet M-P, Marques D da M, Sarmento H. 2019. Spatial heterogeneity and hydrological fluctuations drive bacterioplankton community composition in an Amazon floodplain system. PLOS ONE 14:e0220695.

47. Satinsky BM, Zielinski BL, Doherty M, Smith CB, Sharma S, Paul JH, Crump BC, Moran MA. 2014. The Amazon continuum dataset: quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010. Microbiome 2:17.

48. Haverkamp THA, Schouten D, Doeleman M, Wollenzien U, Huisman J, Stal LJ. 2009. Colorful microdiversity of Synechococcus strains (picocyanobacteria) isolated from the Baltic Sea. ISME J 3:397–408.
49. Mackey KRM, Hunter-Cevera K, Britten GL, Murphy LG, Sogin ML, Huber JA. 2017. Seasonal Succession and Spatial Patterns of Synechococcus Microdiversity in a Salt Marsh Estuary Revealed through 16S rRNA Gene Oligotyping. Front Microbiol 8.

50. Hilton JA, Satinsky BM, Doherty M, Zielinski B, Zehr JP. 2015. Metatranscriptomics of N2-fixing cyanobacteria in the Amazon River plume. The ISME journal 9:1557–69.

51. Del Vecchio R, Subramaniam A. 2004. Influence of the Amazon River on the surface optical properties of the western tropical North Atlantic Ocean. Journal of Geophysical Research 109:C11001.

52. Satinsky BM, Smith CB, Sharma S, Landa M, Medeiros PM, Coles VJ, Yager PL, Crump BC, Moran MA. 2017. Expression patterns of elemental cycling genes in the Amazon River Plume. The ISME Journal 11:1852–1864.

53. Santos-Júnior CD, Kishi LT, Toyama D, Soares-Costa A, Oliveira TCS, de Miranda FP, Henrique-Silva F. 2017. Metagenome Sequencing of Prokaryotic Microbiota Collected from Rivers in the Upper Amazon Basin. Genome Announcements 5:e01450-16.

54. Toyama D, Kishi LT, Santos-Júnior CD, Soares-Costa A, Souza De Oliveira TC, Pellon De Miranda F, Henrique-Silva F. 2016. Metagenomics Analysis of Microorganisms in Freshwater Lakes of the Amazon Basin. Genome Announc 4:1440–16.

55. Toyama D. 2016. Metagenoma da Amazônia: Busca por genes de interesse biotecnológico. Federal University of Sao Carlos.

56. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10.

57. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, Yamashita H, Lam TW. 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods.

58. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

59. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

60. Kang DD, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3.

61. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research 25:1043–55.

62. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research 41:D590-6.

63. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. 2014. The SILVA and “all-species Living Tree Project (LTP)” taxonomic frameworks. Nucleic Acids Research 42:D643-8.
64. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

65. Parks DH. 2016. CompareM. GitHub.

66. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2017. High-throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries. bioRxiv 225342.

67. Parks DH, Chuvchiona M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A proposal for a standardized bacterial taxonomy based on genome phylogeny. bioRxiv 256800.

68. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyprides NC, Pati A. 2015. Microbial species delineation using whole genome sequences. Nucleic Acids Research 43:6761–6771.

69. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31:2032–2034.

70. Wickham Hadley. 2009. Ggplot2: elegant graphics for data analysis. Springer.

71. Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz BV. 2016. gplots: Various R Programming Tools for Plotting Data. CRAN - Package gplots.

72. Neuwirth E. 2014. CRAN - Package ColorBrewer Palettes. Comprehensive R Archive Network (CRAN).

73. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser L.J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.

74. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

75. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research 45:D353–D361.

76. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

77. UniProt Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Research 43:D204–D212.

78. Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. Nature Methods 12:59–60.

79. Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Computational Biology 7:e1002195.

80. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic acids research 40:W445-51.

81. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic acids research 44:D279-85.

82. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Research 44:D286–D293.
83. Price MN, Dehal PS, Arkin AP. 2009. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Molecular Biology and Evolution 26:1641–1650.
84. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC bioinformatics 8:460.

**Tables**

*Table 1. High Quality PGs produced in this study.* PGs are described in terms of the river section from where they originate, their GC% content, size in Mbp, completeness (C), contamination (Cx), quality classification and taxonomy.
| PG     | Section | GC (%) | Size (Mbp) | C (%) | Cx  | Classification | Taxonomy         |
|--------|---------|--------|------------|-------|-----|----------------|------------------|
| AM_0118| Upstream| 37.77  | 1.35       | 81.87 | 0   | Medium-Quality | Streptomyces     |
|        |         |        |            |       |     | Draft          | eae              |
| AM_0219| Upstream| 36.37  | 1.05       | 76.13 | 0.71| Medium-Quality | Methylophilus    |
|        |         |        |            |       |     | Draft          |                  |
| AM_0226| Upstream| 63.28  | 2.99       | 98.79 | 0.34| High-Quality   | Rhizobiales      |
|        |         |        |            |       |     | Draft          |                  |
| AM_0228| Upstream| 56.89  | 2.92       | 97.56 | 2.42| High-Quality   | Burkholderiales  |
|        |         |        |            |       |     | Draft          |                  |
| AM_0233| Upstream| 53     | 2.98       | 96.26 | 0   | High-Quality   | Proteobacteria   |
|        |         |        |            |       |     | Draft          |                  |
| AM_0240| Upstream| 60.07  | 3.02       | 91.32 | 0.89| High-Quality   | Bacteria         |
|        |         |        |            |       |     | Draft          |                  |
| AM_0244| Upstream| 41.41  | 2.10       | 93.23 | 0.76| High-Quality   | Gammaproteobacteria |
|        |         |        |            |       |     | Draft          |                  |
| AM_0256| Upstream| 53.45  | 2.66       | 88.24 | 2.15| Medium-Quality | Bacteria         |
|        |         |        |            |       |     | Draft          |                  |
| AM_0268| Upstream| 52.62  | 2.91       | 86.69 | 0.42| Medium-Quality | Bacteria         |
|        |         |        |            |       |     | Draft          |                  |
| AM_0275| Upstream| 52.7   | 2.41       | 72.7  | 1.41| Medium-Quality | Rhizobiales      |
|        |         |        |            |       |     | Draft          |                  |
| AM_0466| Upstream| 41.13  | 1.03       | 62.5  | 0   | Medium-Quality | Bacteria         |
|        |         |        |            |       |     | Draft          |                  |
| AM_0507| Upstream| 36.3   | 1.00       | 79.25 | 0.64| Medium-Quality | Betaproteobacteria |
|        |         |        |            |       |     | Draft          |                   |
| AM_0510| Upstream| 62.61  | 2.00       | 98.08 | 1.1 | High-Quality   | Bacteria         |
|        |         |        |            |       |     | Draft          |                  |
| AM_0519| Upstream| 65.04  | 4.82       | 96.74 | 0.07| High-Quality   | Gammaproteobacteria |
|        |         |        |            |       |     | Draft          |                   |
| AM_0528| Upstream| 48.07  | 2.08       | 93.43 | 0.34| High-Quality   | Rhizobiales      |
| AM_0546  | Upstream  |          |       | 0.07  | Medium-Quality | Actinobacteria |
|---------|-----------|----------|-------|-------|---------------|----------------|
| AM_0608 | Upstream  |          |       | 1.07  | High-Quality   | Moraxellaceae  |
| AM_0615 | Upstream  |          |       | 0.97  | High-Quality   | Thaumarchaeota |
| AM_0616 | Upstream  |          |       | 0.59  | High-Quality   | Betaproteobacteria |
| AM_0619 | Upstream  |          |       | 2.52  | High-Quality   | Bacteria       |
| AM_0621 | Upstream  |          |       | 0     | High-Quality   | Bacteria       |
| AM_0630 | Upstream  |          |       | 2.17  | High-Quality   | Proteobacteria |
| AM_0643 | Upstream  |          |       | 2.2   | High-Quality   | Burkholderiales |
| AM_0729 | Upstream  |          |       | 1.26  | Medium-Quality | Bacteria       |
| AM_0764 | Upstream  |          |       | 1.1   | Medium-Quality | Bacteria       |
| AM_0832 | Downstream |         |       | 1.64  | High-Quality   | Bacteroidetes  |
| AM_0849 | Downstream |         |       | 6.78  | Medium-Quality | Actinobacteria |
| AM_0854 | Downstream |         |       | 0     | Medium-Quality | Bacteria       |
| AM_0876 | Downstream |         |       | 2.88  | High-Quality   | Bacteria       |
| AM_0902 | Downstream |         |       | 0     | High-Quality   | Cyanobacteria  |
| Accession | Habitat   | Downstream | Medium-Quality | Bacteria | AM_0936 | 62.13 | 2.69 | 86.63 | 4.45 | Medium-Quality | Bacteria | AM_1003 | 59.74 | 6.26 | 95.93 | 0.91 | High-Quality | Betaproteobacteria | AM_1104 | 59.11 | 3.00 | 78.96 | 3.08 | Medium-Quality | Bacteria | AM_1111 | 38.28 | 2.82 | 90.95 | 6.68 | Medium-Quality | Bacteroidetes | AM_1205 | Estuary | 47.76 | 0.71 | 65.39 | 0 | Medium-Quality | Bacteria | AM_1312 | Estuary | 61.53 | 2.17 | 85.71 | 1.82 | Medium-Quality | Bacteria | AM_1409 | Estuary | 39.04 | 0.61 | 70.98 | 1.72 | Medium-Quality | Bacteria | AM_1503 | Estuary | 67.21 | 2.09 | 63.57 | 0.73 | Medium-Quality | Bacteria | AM_1603 | Estuary | 63.07 | 4.54 | 98.42 | 1.19 | High-Quality | Sphingomonadales | AM_1606 | Estuary | 55.47 | 2.24 | 96.4 | 0.61 | High-Quality | Bacteria | AM_1801 | Plume | 45.14 | 1.20 | 67.56 | 0 | Medium-Quality | Euryarchaeota | AM_1811 | Plume | 43.37 | 2.27 | 85.47 | 3.72 | Medium-Quality | Bacteria | AM_2104 | Plume | 60.62 | 2.11 | 84.85 | 4.55 | Medium-Quality | Bacteria | AM_2116 | Plume | 46.6 | 0.69 | 52.31 | 0 | Medium-Quality | Bacteria | AM_2124 | Plume | 40.54 | 1.53 | 87.19 | 1.49 | Medium-Quality | Flavobacteriac
|                 |          |        |        | 0  |                     |             |
|----------------|----------|--------|--------|----|---------------------|-------------|
| **AM_2202**    | Plume    | 47.45  | 1.80   | 97.31 | 0                   | High-Quality |
|                 |          |        |        |    |                     | Bacteria    |
| **AM_2207**    | Plume    | 43.09  | 1.79   | 85.25 | 1.19                | Medium-Quality |
|                 |          |        |        |    |                     | Flavobacteriac |
| **AM_2208**    | Plume    | 53.01  | 1.62   | 92.99 | 0.36                | High-Quality |
|                 |          |        |        |    |                     | Bacteria    |
| **AM_2324**    | Plume    | 37.43  | 3.41   | 90.78 | 0.99                | High-Quality |
|                 |          |        |        |    |                     | Bacteroidetes |
| **AM_2502**    | Ocean    | 39.43  | 1.43   | 89.99 | 1.18                | Medium-Quality |
|                 |          |        |        |    |                     | Flavobacteriac |
| **AM_2804**    | Ocean    | 33.5   | 3.29   | 93.56 | 0.27                | High-Quality |
|                 |          |        |        |    |                     | Cyanobacteria |

**Additional Files**

**SUPPLEMENTARY TABLES**: Co-assemblies (Table S1); Average Amino-acid identity (AAI) of redundant PGs (Table S2); Main features of non-redundant PGs (Table S3); Genomes from other databases similar to those recovered from the Amazon River (Table S4); Transporters of lignin-derived aromatic compounds per PG (Table S5); Lignin-derived aromatic compounds degradation genes per PG (Table S6); and the marker protein families used for phylogenetic inference via CheckM (Table S7).

**Figure S1. Selection of high-quality Population Genomes (PGs)**. A) PGs were first selected according to their genome completeness (>50%), contamination (<10%), and quality ≥ 50 (quality=completeness − 5 x contamination). Then PGs featuring B) N50 >10 Kbp and C) <500 contigs, were kept. Subsequently, we removed three additional PGs that differed in <1% in Average Amino-Acid Identity (AAI) and that were considered redundant. The final dataset included 51 high-quality PGs.

**Figures**
Figure 1

Metagenomes and Population Genomes (PGs) from the Amazon River. A) Distribution of the 106 metagenomes along the five sections of the Amazon River: Upstream, Downstream, Estuary, Plume and Ocean, from where the PGs derive. The size of the red circles indicates the sequencing coverage per metagenome (Cov./Gbp). B) Number of curated PGs that were reconstructed per river section.
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 are reported b) (IM – inner membrane, OM – outer membrane). The polyhydroxy-butyrate/alkanoate production c) was further investigated to assess the potential for carbon storage in Amazon River PGs. Those PGs with the potential to store carbon are shown 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 byproducts. The red arrow indicates growth inhibition and the beige arrow its suppression. In this model, 3 different microbial assemblages interact. Two of them are responsible for exposing and processing TeOM. Besides providing structural carbon and energy for cell growth, byproducts of this metabolism inhibit TeOM degradation. The third assemblage prevents the TeOM degradation to stop by consuming these byproducts and storing the resulting carbon.

Supplementary Files

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

- SupplementaryTables.xlsx
- FigureS1.tiff