Grevesse *et al. Microbiome* (2022) 10:237
https://doi.org/10.1186/s40168-022-01417-6

Microbiome

RESEARCH

Open Access

Degradation pathways for organic matter of terrestrial origin are widespread and expressed in Arctic Ocean microbiomes

Thomas Grevesse[1], Céline Guéguen[2], Vera E. Onana[1] and David A. Walsh[1*]

Abstract

Background: The Arctic Ocean receives massive freshwater input and a correspondingly large amount of humic-rich organic matter of terrestrial origin. Global warming, permafrost melt, and a changing hydrological cycle will contribute to an intensification of terrestrial organic matter release to the Arctic Ocean. Although considered recalcitrant to degradation due to complex aromatic structures, humic substances can serve as substrate for microbial growth in terrestrial environments. However, the capacity of marine microbiomes to process aromatic-rich humic substances, and how this processing may contribute to carbon and nutrient cycling in a changing Arctic Ocean, is relatively unexplored. Here, we used a combination of metagenomics and metatranscriptomics to assess the prevalence and diversity of metabolic pathways and bacterial taxa involved in aromatic compound degradation in the salinity-stratified summer waters of the Canada Basin in the western Arctic Ocean.

Results: Community-scale meta-omics profiling revealed that 22 complete pathways for processing aromatic compounds were present and expressed in the Canada Basin, including those for aromatic ring fission and upstream funneling pathways to access diverse aromatic compounds of terrestrial origin. A phylogenetically diverse set of functional marker genes and transcripts were associated with fluorescent dissolved organic matter, a component of which is of terrestrial origin. Pathways were common throughout global ocean microbiomes but were more abundant in the Canada Basin. Genome-resolved analyses identified 12 clades of *Alphaproteobacteria*, including *Rhodospirillales*, as central contributors to aromatic compound processing. These genomes were mostly restricted in their biogeographical distribution to the Arctic Ocean and were enriched in aromatic compound processing genes compared to their closest relatives from other oceans.

Conclusion: Overall, the detection of a phylogenetically diverse set of genes and transcripts implicated in aromatic compound processing supports the view that Arctic Ocean microbiomes have the capacity to metabolize humic substances of terrestrial origin. In addition, the demonstration that bacterial genomes replete with aromatic compound degradation genes exhibit a limited distribution outside of the Arctic Ocean suggests that processing humic substances is an adaptive trait of the Arctic Ocean microbiome. Future increases in terrestrial organic matter input to the Arctic Ocean may increase the prominence of aromatic compound processing bacteria and their contribution to Arctic carbon and nutrient cycles.

*Correspondence: david.walsh@concordia.ca

[1] Department of Biology, Concordia University, 7141 Sherbrooke St. West, Montreal, QC H4B 1R6, Canada
Full list of author information is available at the end of the article

Introduction

Humic substances (HS) are a heterogeneous mixture of organic compounds resulting from biochemical transformations of dead plants and microbes. HS are ubiquitous in both terrestrial and aquatic systems and

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

constitute the largest fraction of organic matter (OM) in terrestrial ecosystems [1, 2], reaching 60–80% in soils [3] and 50–80% in freshwaters [4]. The fraction of HS is relatively high (20–60%) in shelves, coastal zones, and estuaries [5] due to the input of terrestrial OM (tOM) with freshwater runoff and exchange with sediments. HS constitute a smaller fraction of dissolved OM (DOM) in the open ocean (0.7–2.4%) [6]. The lesser amount of HS in the DOM of open oceans indicates that HS is removed by ocean microbiomes and additional non-biological processes [7, 8].

The Arctic Ocean receives a disproportionately high input of freshwater (10% of total global freshwater input for 1.3% of total ocean volume) and a correspondingly high tOM input (10% of ocean total tOM input) [9]. Rivers annually discharge 25–36 Tg of dissolved organic carbon and 12 Tg of particulate organic carbon to the Arctic Ocean [10, 11]. Climate change is strongly influencing the Arctic region, which in turn is influencing Arctic hydrology and organic matter dynamics [12, 13]. More specifically, permafrost thawing [14], combined with intensifying river runoff [15], coastal erosion [16], and groundwater input [17], is driving an increase in the amount of humic-rich DOM input into the Arctic Ocean. The humic-rich DOM consequently contributes significantly to the carbon pool of the Arctic Ocean DOM compared to other oceans [18] and potentially represents a significant and increasing growth resource for the Arctic Ocean microbiome.

The origins and distributions of tOM in the Canada Basin in the western Arctic Ocean have been extensively studied, making it a useful system for investigating interactions between tOM and ocean microbiomes. In spring and summer, humic-rich OM is transported by riverine inputs to the surface mixed layer of the Arctic Ocean shelves [19, 20]. In shelf waters, tOM is partially photodegraded [21], while some flocculates upon mixing with salt water and sinks to the sediments along with particulate OM [15]. In fall and winter, the tOM remaining in the surface layer sinks with the dense brine expelled during ice formation. This brine flows along Chukchi Sea and Beaufort Sea shelves, exchanging organic matter with bottom sediments, ultimately accumulating in the deeper and more saline water of Pacific Ocean origin [16, 22]. The interactions with shelf sediments and pore waters constitute a substantial source of tOM which may have been reprocessed by sediment microbiomes [23]. It has been estimated that 11–44% of Arctic Ocean sediment OM is of terrestrial origin [24]. As a consequence of the OM dynamics, the Canada Basin is characterized by a strong and distinctive signal of humic-rich DOM that extends from the subsurface water to a depth of ~300 m [25, 26].

HS are heterogeneous supramolecular assemblies formed by microbial and physicochemical transformations [27] of organic matter. In terrestrial systems, HS originate from vascular plant residues (lignin and other biopolymers) and other organic detritus [28], giving rise to HS rich in aromatic moieties. In contrast, HS produced in marine environments have a strong aliphatic and branched structure [29]. In the Arctic Ocean however, HS are aromatic rich due to their terrestrial and sediment origin [30]. HS usually show a high degree of recalcitrance that is dependent on their physicochemical interactions with the environment [31]. In soils, sorption of HS to mineral particles drives a physical separation of HS from microbes and their enzymes, preventing fast degradation of HS [32, 33]. Numerous studies have therefore demonstrated that HS freed from their soil environment can be used to support microbial growth [7, 34, 35].

The capacity of microbiomes to couple HS transformation to growth relies on the ability to degrade aromatic compounds from HS. The degradation of aromatic compounds follows two main steps. Funneling pathways transform (e.g., via oxidation, decarboxylation, and/or demethylation) larger and more substituted aromatic compounds to a small set of key aromatic compounds (e.g., gentisate, catechol, protocatechuate), which then undergo an aromatic ring-fission step followed by further processing to generate central carbon metabolism intermediates. In humic-rich environments such as soils, microbiomes use a wide variety of funneling pathways to access the diverse set of lignin-derived aromatic compounds (e.g., vanillate, syringate, benzoate, and their derivatives) that have been incorporated in HS [36].

In soils, fungi degrade most of the humic substances [37]. In the ocean, bacteria are considered the main actors in OM degradation [38], even if recent studies have highlighted an important role for fungi [39, 40], for example, in processing OM in marine snow [41] or by parasitizing phytoplankton [42]. Certain bacteria inhabiting humic-rich environments can grow on HS as sole carbon and energy sources and are therefore able to access aromatic compounds within HS [5, 34, 43]. Transcriptomic analysis from the humic acid-degrading bacterium *Pseudomonas* sp. isolated from subarctic tundra soils showed that genes involved in the funneling and ring-opening steps of aromatic compound degradation pathways were upregulated when fed with humic acids compared to glucose [44]. Recently, it was shown that *Chloroflexi* genomes from the Canada Basin encoded a diverse set of genes associated with aromatic compound degradation [45]. The *Chloroflexi* populations appeared to be endemic to the Arctic Ocean and were associated with the humic-rich fluorescence DOM maximum

(FDOMmax). These observations suggest the disproportionately high fraction and diversity of aromatic-rich HS in the Arctic Ocean DOM compared to other oceans may select for a diverse HS-degrading microbiome.

The genomic diversity and metabolic pathways in the Arctic Ocean microbiomes can provide important insights regarding the fate of HS and its impact on Arctic Ocean biogeochemical cycles. However, outside of perhaps the *Chloroflexi*, we know very little about how phylogenetically widespread HS degradation is in the Arctic Ocean microbiomes, nor the diversity of metabolic pathways employed by the Arctic Ocean microbiomes to process HS. We hypothesized that the capacity for aromatic compound degradation was linked to the distribution of humic-rich tOM and enhanced in the Arctic Ocean compared to other oceans with a lesser amount of HS. Finally, we hypothesized that the vast amount of HS in the Arctic Ocean may have played a role as an ecological pressure for the adaptive evolution of the taxa most implicated in aromatic compound degradation.

Results

Environmental context

We surveyed the microbiomes along a latitudinal transect (73–81° N) of the salinity-stratified waters of the Canada Basin using a combination of metagenomics and metatranscriptomics (Fig. 1 a–b, Table S1). The sampling design targeted distinct water column features, including the relatively fresh surface mixed layer (surface; 5 m and 20 m), the subsurface chlorophyll maximum (SCM; 55–95 m), the FDOMmax associated with colder Pacific-origin water (32.3 and 33.1 PSU, 90–250 m), the warmer Atlantic-origin water (Tmax and AW; 360–1000 m depth), and Arctic bottom water (~3800 m). The warmer Atlantic-origin water and Arctic bottom water are herein collectively referred to as deep waters.

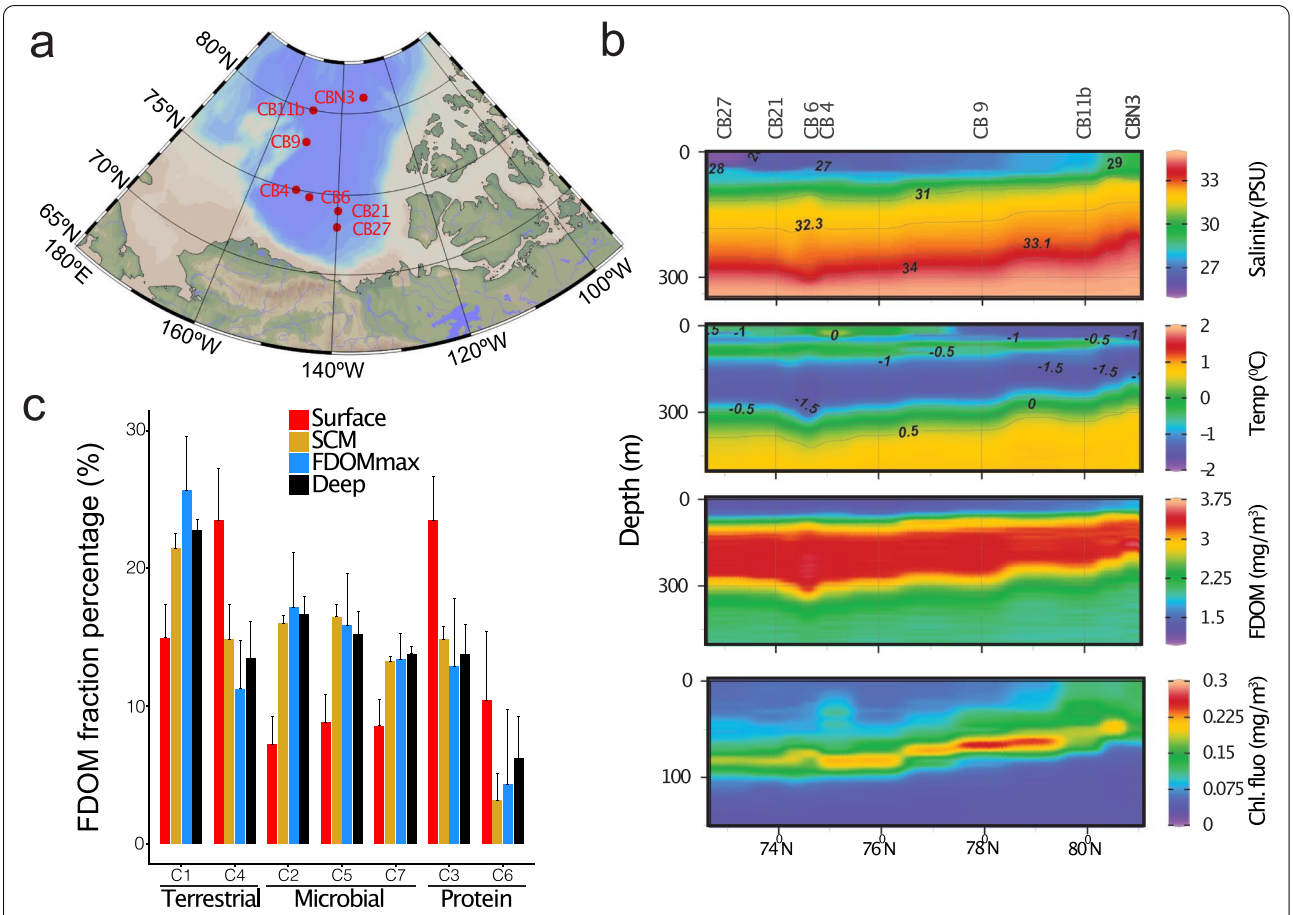

Fig. 1 Spatial biogeochemistry of the Canada Basin. **a** Map of the 7 stations sampled in this study. **b** Depth profile of salinity (PSU), temperature (°C), fluorescent dissolved organic matter (FDOM, mg/m^3), and chlorophyll fluorescence (mg/m^3) at the 8 stations sampled in this study. **c** Percentage of the 7 FDOM fractions identified using excitation emission matrix fluorescence spectroscopy combined with parallel factor analysis. Samples are grouped in 4 samples water features: surface, subsurface chlorophyll maximum (SCM), fluorescent dissolved organic matter maximum (FDOMmax), and deep waters

We sought to determine the distribution and composition of OM in the Canada Basin, with a focus on the distribution of tOM. Optical properties of the OM, such as fluorescence, have previously been used to assess the composition of OM in the ocean and differentiate between terrestrial and marine OM sources [46, 47]. We used excitation emission matrix (EEC) fluorescence spectroscopy combined with parallel factor analysis (PARAFAC) to determine the distribution of fluorescent DOM components. In the Canada Basin, seven components (C1-C7) were identified, as previously defined in DeFrancesco and Guéguen [26]. These components corresponded to terrestrially derived humic-like DOM (C1 and C4), amino acid or protein material (C3 and C6), or microbially derived humic-like DOM (C2, C5, and C7) (Fig. 1c). The aromatic-rich C1 was the most abundant component within the FDOMmax samples (25–27%) but also in the whole water column below the surface (20–22% in the SCM and 21–23% in the deep), verifying that a significant fraction of OM is of terrestrial origin. Of the terrestrial components, C4 was the dominant component in the surface (19–30%). The reduced contribution of C1 in the surface is because C1 is more red shifted than C4 indicating a stronger aromatic character and thus enhanced photosensitivity. Overall, these results indicate a strong contribution of a photostable fraction from terrestrial origin in the FDOM of the surface and an aromatic-rich fraction from terrestrial origin in the FDOM of the whole water column below the surface.

Vertical partitioning of metabolic features in metagenomes and metatranscriptomes

We investigated the abundance and distribution of aromatic compound degradation pathways in the Canada Basin microbiomes in relation to tOM availability. To investigate how metabolic pathways were distributed across the Canada Basin, we first performed nonmetric multidimensional scaling (NMDS) analysis on the abundance of enzyme-encoding genes (genes assigned to enzyme commission (EC) numbers) annotated from metagenome or metatranscriptome assemblies. NMDS ordination showed that metagenomes (stress = 0.11) were partitioned into four clusters consisting of samples collected from either the surface, the SCM, FDOMmax, or deep water (Fig. 2a). A similar pattern was observed in the NMDS ordination of metatranscriptomes (stress = 0.0503), although the variation between samples from within the same water column feature was higher than observed in metagenomes (Fig. 2b). In addition, there was less separation between the samples from the FDOMmax and from deeper Atlantic-origin waters in the ordination based on metatranscriptomes compared to metagenomes.

We next determined which enzymatic reactions differentiated the metagenomes across the stratified water column using nonnegative matrix factorization (NMF), which is a tool for extracting meaningful features from high dimensional data [48]. In our analyses, NMF decomposes the matrix of EC number abundances into two matrices. Matrix 1 presents a reduced number of elements that describe the overall similarities of the metagenomes based on EC number composition, while matrix 2 presents the weighted contribution of individual EC numbers on each of the elements in matrix 1. We determined that a decomposition with four elements best represented the overall enzyme composition of metagenomes (Fig. S1). The four elements, herein referred to as sub-metagenomes (Fig. S2), represented the same patterns as observed in the NMDS ordination, corresponding to the surface, SCM, FDOMmax, and deep waters (Fig. 2a).

We then assessed which EC numbers were strongly associated with each of the four sub-metagenomes by calculating an EC index value. This EC index value quantifies the tendency of an EC number to be specific to a single sub-metagenome (EC index values range between −1 and 1). The distribution of EC indices was plotted for each of the four sub-metagenomes. Overall, the means of the EC indices associated with aromatic compound degradation and other metabolic pathways in the four water column features were significantly different (PERMANOVA, $F = 89.8$, $p < 0.0001$). Each sub-metagenome has a collection of EC numbers with relatively high indices (> 0.5) (Fig. 2c). However, the most striking observation was that EC numbers involved in aromatic compound degradation were predominantly associated with the FDOMmax sub-metagenome, as demonstrated by the higher index values for EC numbers from aromatic compound degradation pathways than from other metabolic pathways in the FDOMmax (Student *t*-test, $t = 13.26$, $p < 0.0001$). The EC indices for aromatic compound degradation genes were smaller than the EC indices associated with other metabolic processes in the surface (Student *t*-test, $t = 8.89$, $p < 0.0001$) and not significantly different for the SCM (Student *t*-test, $t = 0.369$, $p = 0.414$) and the deep (Student *t*-test, $t = 0.56$, $p = 0.545$) sub-metagenomes.

We performed a similar NMF analysis on EC numbers in the metatranscriptomes (Fig. S1). Similar with the NMF analysis of metagenomes, decomposition resulted in four elements, herein referred to as sub-metatranscriptomes (Fig. S1), corresponding to the surface, SCM, FDOMmax, and deep waters (Fig. S3). For the sub-metatranscriptomes, the means of the EC indices from aromatic compound degradation and from other

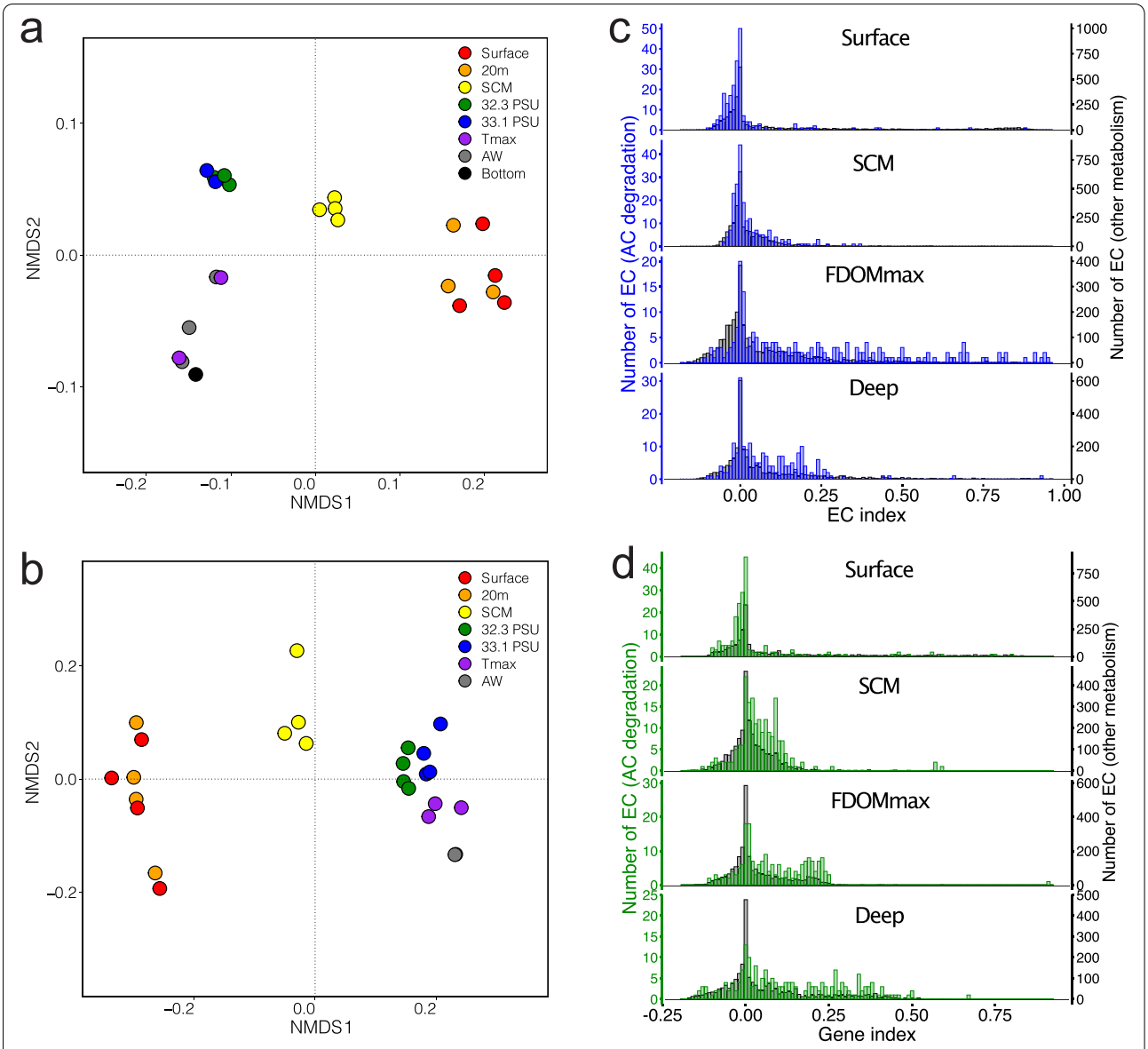

Fig. 2 Identification of the microbial enzymatic reactions associated with the 4 water column features. Nonmetric multidimensional scaling of the matrices of enzymatic reaction abundances from **a** metagenomes and **b** metatranscriptomes. Distribution of gene indices calculated for the enzymatic reactions from **c** metagenomes and **d** metatranscriptomes in the 4 water features. Blue/green, distribution for enzymatic reactions involved in the degradation of aromatic compounds from metagenomes (blue) and metatranscriptomes (green); gray, distribution for all enzymatic reactions not involved in the degradation of aromatic compounds

metabolic pathways in the four water column features were significantly different (PERMANOVA, $F = 121$, $p < 0.0001$). For the sub-metatranscriptomes, however, we observed higher indices for the EC numbers from aromatic compound degradation than for EC numbers from other metabolic processes in the SCM (Student t-test, $t = 0.0218$, $p < 0.0001$), the FDOMmax (Student t-test, $t = 0.0444$, $p < 0.0001$), and the deep waters (Student t-test, $t = 0.0611$, $p < 0.0001$) (Fig. 2d).

Aromatic compound degradation genes in global ocean metagenomes

As the humic-rich OM input to the Arctic Ocean is disproportionately high compared to other oceans, we investigated if genes associated with aromatic compound degradation were more abundant in the Canada Basin metagenomes compared to other oceanic metagenomes. As terrestrial OM is a significant contributor of HS to the Arctic Ocean, we restricted our analysis to genes involved

in processing aromatic compounds of terrestrial origin. We focused the analysis on a set of 46 pathways previously implicated in degrading aromatic compounds from lignin (Fig. 3a, Table S2). We compared the relative abundance of aromatic compound degradation genes between metagenomes of the Canada Basin water features (surface, SCM, FDOMmax, deep) and metagenomes from both the surface and subsurface waters (SCM, mesopelagic, bathypelagic) of the Atlantic, Pacific, Indian, and Southern Oceans as well as the Mediterranean Sea and Red Sea (Fig. 3b). The Canada Basin FDOMmax metagenomes contained the highest fraction of aromatic compound degradation genes (3.4%). Aromatic compound degradation genes were identified in other oceanic metagenomes (1.5–2.5% of total protein-coding genes) and the relative abundance of aromatic compound degradation genes increased with water depth. Overall, the mean percentage of aromatic compound degradation genes in the water column features of the Canada Basin were significantly different than in other oceans (PERMANOVA, $F = 27.8$, $p < 0.0001$). Specifically, the relative abundance was consistently higher (1.3–1.7 fold) in the microbiomes of the Canada Basin upper water column features compared to microbiomes from other

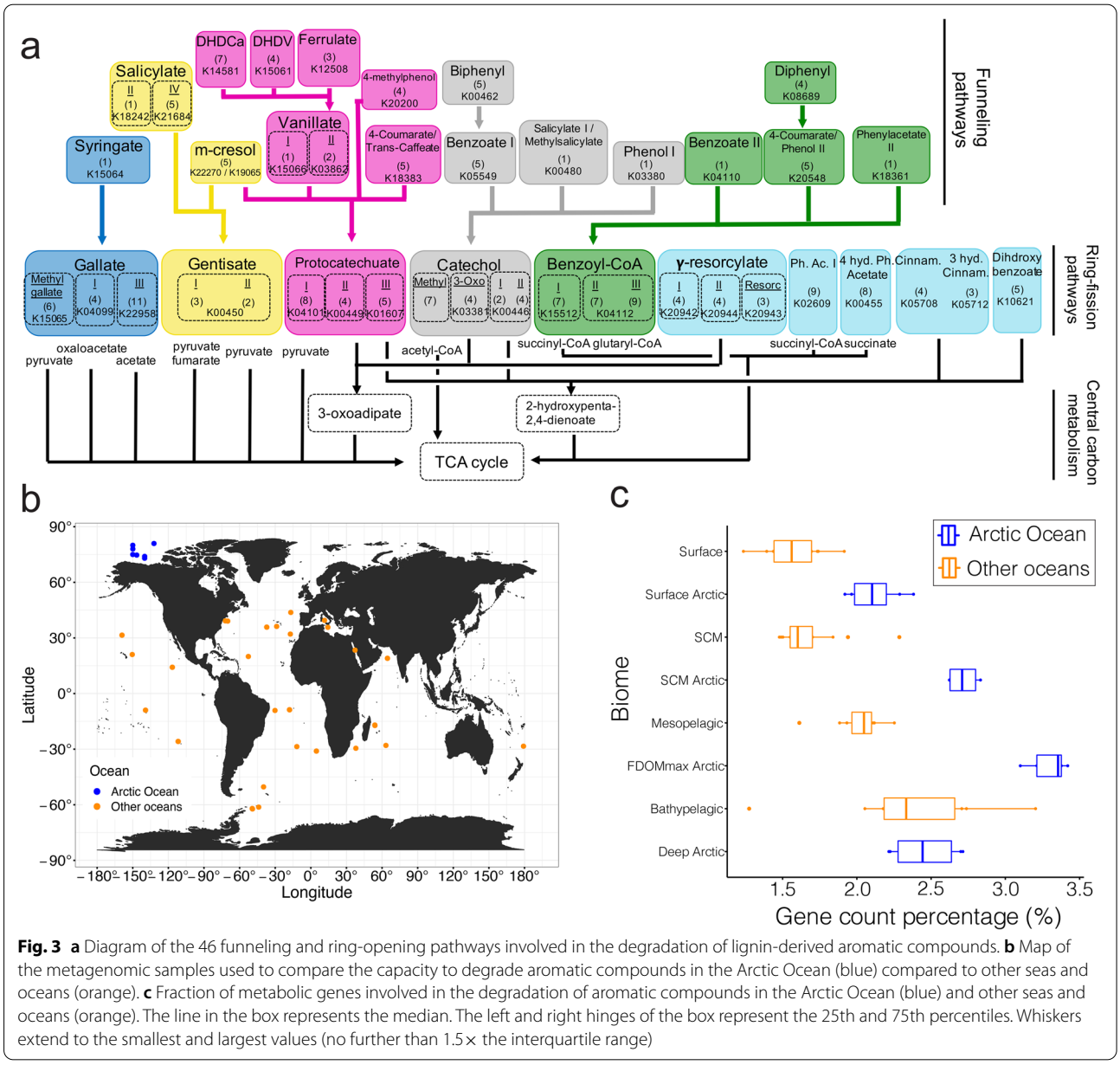

Fig. 3 **a** Diagram of the 46 funneling and ring-opening pathways involved in the degradation of lignin-derived aromatic compounds. **b** Map of the metagenomic samples used to compare the capacity to degrade aromatic compounds in the Arctic Ocean (blue) compared to other seas and oceans (orange). **c** Fraction of metabolic genes involved in the degradation of aromatic compounds in the Arctic Ocean (blue) and other seas and oceans (orange). The line in the box represents the median. The left and right hinges of the box represent the 25th and 75th percentiles. Whiskers extend to the smallest and largest values (no further than 1.5× the interquartile range)

oceanic zones (Student *t*-test: surface/surface, $t = 0.58$, $p = 0.0013$; SCM/SCM, $t = 0.92$, $p = 0.0001$; FDOMmax/mesopelagic, $t = 0.81$, $p = 0.0001$) (Fig. 3c). However, we did not observe significant differences between the percentage of aromatic compound degradation genes of Arctic deepwater microbiomes and the microbiomes of other oceans deep waters (Student *t*-test, $t = 0.20$, $p = 0.490$) (Fig. 3c).

Distribution of aromatic compound degradation genes and pathways in metagenomes and metatranscriptomes

To elucidate the diversity of aromatic compounds that the Arctic Ocean microbiomes can access as growth substrates, we assessed the diversity and the completeness of aromatic compound degradation pathways in Canada Basin metagenomes. We found evidence for the presence of 44 of the 46 aromatic compound degradation pathways in the metagenomes (Fig. S4). A complete set of genes were identified for over half of these pathways in the metagenomes, irrespective of the water column feature (Fig. S4). Evidence for the 44 pathways was also identified in the metatranscriptomes, including expression of the full complement of genes for 22 pathways (Fig. S4).

To measure the distribution of the aromatic compound degradation pathways through the water column, we used a selection of 39 unique marker genes for the 46 aromatic compound degradation pathways (Table S2). To provide a measure of pathway abundance and expression, we summed the depth of coverage of each marker gene or transcript and corrected for differences in metagenome sequencing effort (Fig. 4). Out of the 39 unique marker genes, 32 were detected in the Canada Basin metagenomes (Fig. 4a, Fig. S5). Generally, the most abundant genes were also most abundant in the metatranscriptomes (Fig. 4 a–b, Fig. S5, Fig. S6). Most of the marker genes were most abundant in the FDOMmax yet were most highly expressed in deep waters (Fig. 4 a–b).

Vanillate monooxygenase (K03862) was the most abundant marker gene within all water column features (20 copies/10^6 reads in the FDOMmax) for pathways degrading aromatic compounds from terrestrial sources, while 3-O-methylgallate 3,4-dioxygenase (K15065) showed a lower abundance (8 copies/10^6 reads in the FDOMmax). While vanillate monooxygenase was more abundant in the metatranscriptomes of the FDOMmax, 3-O-methylgallate 3,4-dioxygenase was more abundant in the metatranscriptomes of the deep layers. Both ring fission protocatechuate dioxygenases (K00449 and K04101) were abundant in metagenomes (8 and 6 copies/10^6 reads in the FDOMmax) and metatranscriptomes (2.5 and 7.5 copies/10^6 reads in the FDOMmax). Overall, these results show that the Canada Basin microbiomes can fully transform aromatic compounds from terrestrial sources into central carbon metabolism intermediates, with an enhanced capacity in the FDOMmax.

A number of aromatic compounds (e.g., salicylate, 3-hydroxycinnamate, or benzoate) can originate from lignin as well as other sources such as marine phytoplankton. Within the pathways involved in the degradation of aromatic compounds from possible marine or terrestrial origin, benzoate CoA-ligase (K04110), salicylate monooxygenase (K00480), and 3-hydroxycinnamate hydroxylase (K05712) were the most abundant in metagenomes (20, 24, and 17 copies/10^6 reads, respectively) but not in metatranscriptomes (7.5, 5, and 2 copies/10^6 reads). The most common funneling pathway for benzoate was through benzoyl-CoA as evidenced by the lower abundance of genes (5 copies/10^6 reads) and transcripts (2 copies/10^6 reads) encoding benzoate 1,2-dioxygenase (K05549) compared to benzoate CoA-ligase. Accordingly, the ring-fission benzoyl-CoA 2,3-epoxidase (K15512) was significantly more abundant (22 copies/10^6 reads) than the ring-fission marker gene catechol 1,2-dioxygenase (K03381) and catechol 2,3-dioxygenase (K00446) (3 and 7 copies/10^6 reads in the deep and SCM, respectively). However, both benzoyl-CoA 2,3-epoxidase and catechol 2,3-dioxygenase were among the most abundant genes in the metatranscriptomes (20 copies/10^6 reads) but with maximum abundance in the SCM and the deep, respectively (Fig. 4b). Of the ring-fission pathway marker genes, gentisate 1,2-dioxygenase (K00450) was one of the most abundant in metagenomes (15 copies/10^6 reads in the FDOMmax) and metatranscriptomes (20 copies/10^6 reads in the deep).

Taxonomic identity of aromatic compound degradation genes and their distribution across the microbiomes

We estimated the fraction of bacterial genomes harboring each marker gene by comparing the total number of gene variants for select aromatic compound degradation pathway markers to the number of the single-copy universally distributed *recA* genes (Fig. 4c, Fig. S7). The estimated fraction of bacterial genomes with aromatic compound degradation genes increased with depth, reaching a maximum in the FDOMmax (8–75%, Fig. 4c) and then decreased in the deep water (5–25%). The genes present in the highest fraction of bacterial genomes were involved in the degradation of benzoate through benzoyl-CoA (50% for K04110 and 45% for K15521 in the FDOMmax), gentisate (65% for K00450 in the FDOMmax), vanillate (75% for K03862 in the FDOMmax), and salicylate and 3-hydroxycinnamate (45% for K00480 and 40% for K05712 in the SCM) (Fig. S7). These numbers may be overestimated as they assume only a single gene copy per genome, whereas multiple paralogs may be present in a single genome (continued below).

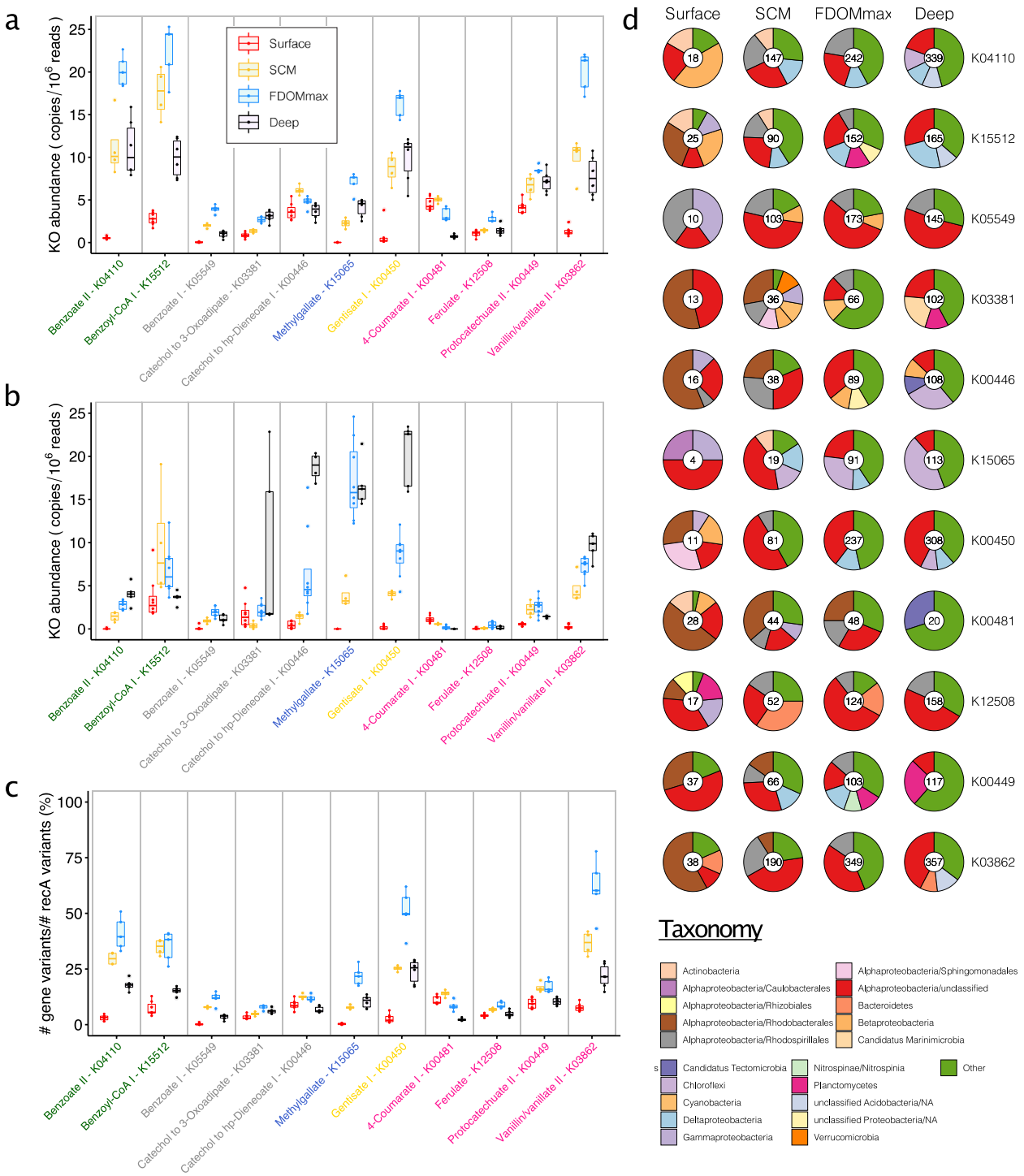

Fig. 4 Distribution and taxonomy of marker genes involved in the degradation of aromatic compounds in the Arctic Ocean. **a** Normalized abundance of genes annotated with KEGG orthology numbers (KOs) marker of aromatic degradation pathways in the 4 different water features. **b** Normalized abundance of transcripts annotated with KOs marker of aromatic degradation pathways in the 4 different water features. **c** Estimated fraction of the microbiomes harboring genes annotated with KOs marker of aromatic degradation pathways in the 4 different water column features. The line in boxes represents the medians. The bottom and upper hinges of the box represent the 25th and 75th percentiles. Whiskers extend to the smallest and largest values (no further than 1.5× the interquartile range). **d** Taxonomy of genes annotated with KOs marker of aromatic degradation pathways in the 4 different water column features. Numbers within pie charts represent the number of gene clusters (clustered at 95% identity) identified in each water column feature. Red, surface samples; yellow, SCM samples; blue, FDOMmax samples; black, deep samples

Taxonomic analysis of aromatic compound degradation marker genes revealed that the number of gene clusters generally increased continuously with depth (Fig. 4d). Surface gene clusters were predominantly affiliated with *Rhodobacterales* (more than 50% of the gene clusters for K03381, K00446, K00481, and K03862) and unclassified *Alphaproteobacteria* (up to 50% for K15065 gene clusters), with a significant contribution from *Gammaproteobacteria* for K05549 (40%) and K15065 (25%) (Fig. 4d). In the SCM and FDOMmax, unclassified *Alphaproteobacteria* dominated the taxonomic affiliations of aromatic compound degradation gene clusters (10–55%), and *Rhodospirillales* contributed significantly to all gene clusters (10–30%), except for genes involved in the degradation of methyl gallate (K15065), which was primarily encoded by *Chloroflexi* (30% in the FDOMmax) (Fig. 4d). We generally observed more gene clusters in the deep than in the FDOMmax (Fig. 4d), while these genes were present in a smaller fraction of the deep communities than the FDOMmax communities (Fig. 4c), suggesting a broader phylogenetic diversity of aromatic compound degradation genes in the deep than in the FDOMmax. This is supported by the large contribution of other taxa (taxa contributing individually to < 5%) in the deep microbiomes. The contribution of taxa such as *Rhodospirillales* and *Rhodobacterales* may be underestimated due to the large fraction of *Alphaproteobacteria* genes that could only be assigned at the class level.

Aromatic compound processing pathways captured in metagenome-assembled genomes

We reconstructed metagenome-assembled genomes (MAGs) from our metagenomic data. We performed metagenomic binning of each of the 22 metagenome assemblies individually to reconstruct a total of 1772 MAGs. After filtering for genomes with greater than 30% completeness and less than 10% contamination, 823 genomes remained (Fig. S8). Thirty-one of the 32 marker genes involved in aromatic compound degradation pathways were identified (only dihydroxyphenylacetate 2,3-dioxygenase — K00455 was not detected) across 59% (482 of 823) of the MAGs (Figs. S9 and S10). The highest percentage of MAGs harboring aromatic compound degradation genes was in the FDOMmax (64%) and SCM (67%) and the lowest percentage in the surface (47%) and deep waters (54%). In general, the taxonomic diversity of MAGs increased with depth. Marker genes were identified in a broad taxonomic diversity of MAGs, including *Alphaproteobacteria*, *Gammaproteobacteria*, and *Dehalococcoidia* (Fig. S9). *Alphaproteobacteria* were common in the SCM and FDOMmax, while *Gammaproteobacteria* were common in the surface.

To investigate the ecology of bacterial taxa most implicated in the degradation of aromatic compounds, we further examined MAGs with complete or near-complete aromatic compound degradation pathways. We selected 46 MAGs most enriched in near-complete aromatic compound degradation pathways (see "Methods," Fig. S10, Table S3). Of the 46 MAGs, 24 were recovered from metagenomes originating from the FDOMmax and 16 from the SCM layers. Thirty-eight of the MAGs were assigned to *Alphaproteobacteria* (Fig. S11), 3 MAGs belonged to the *Dehalococcoida*, 4 MAGs to the *Gammaproteobacteria*, and one MAG to the class *Binatia*.

Given the large representation of *Alphaproteobacteria* in the MAGs most implicated in aromatic compounds degradation, we investigated the evolutionary origins and phylogenetic relationships between our 38 *Alphaproteobacteria* MAGs and reference genomes available from the Genome Taxonomy Database (GTDB) [49] (Table S4). Based on an average nucleotide identity threshold of 95%, our 38 *Alphaproteobacteria* MAGs belonged to 16 genomospecies, which were phylogenetically associated with 12 clades (Fig. 5a, Fig. S12). The clades were within the *Rhodobacterales*, *Thallassobaculales*, *Rhodospirillales*, *Defluviicoccales* and five GTDB orders of uncultured *Alphaproteobacteria* (UBA8366, UBA6615, GCA2731375, UBA2966, UBA828). Each clade was comprised of Canada Basin MAGs as well as a basal branch consisting of genomes of marine origin. These results demonstrate that the *Alphaproteobacteria* MAGs were phylogenetically distinct but shared recent common ancestry with genomes from other oceanic zones.

To investigate the distribution of the 12 clades represented by the *Alphaproteobacteria* MAGs across the Arctic water column features, we performed fragment recruitment of both the metagenomic and metatranscriptomic reads against the MAGs representing the 16 genomospecies (Fig. 5b). Overall, the distribution in metagenomes and metatranscriptomes was similar, and all the MAGs were most abundant and active either in the FDOMmax or the SCM (Fig. 5b). We identified four general patterns of distribution across water column features consisting of (1) restriction to the FDOMmax (Defluvii-CB9_331, UBA2966); (2) common to the SCM and FDOMmax (UBA8366, UBA6615 clade, *Thalassobaculales* clade, *Defluviiccocales* genomes CB21_SCM.1 and CB4_SCM.1); (3) common in the FDOMmax and deeper waters (UBA828 clade genomes, GCA 2732375 genomes, and *Rhodospirillales* genomes); and (4) restricted to the SCM (Rhodosp-CB4_SCM and Rhodobact-CB21_SCM). These results show that MAGs with near-complete aromatic compound degradation pathways are strongly associated with and active in HS-rich regions of the water column.

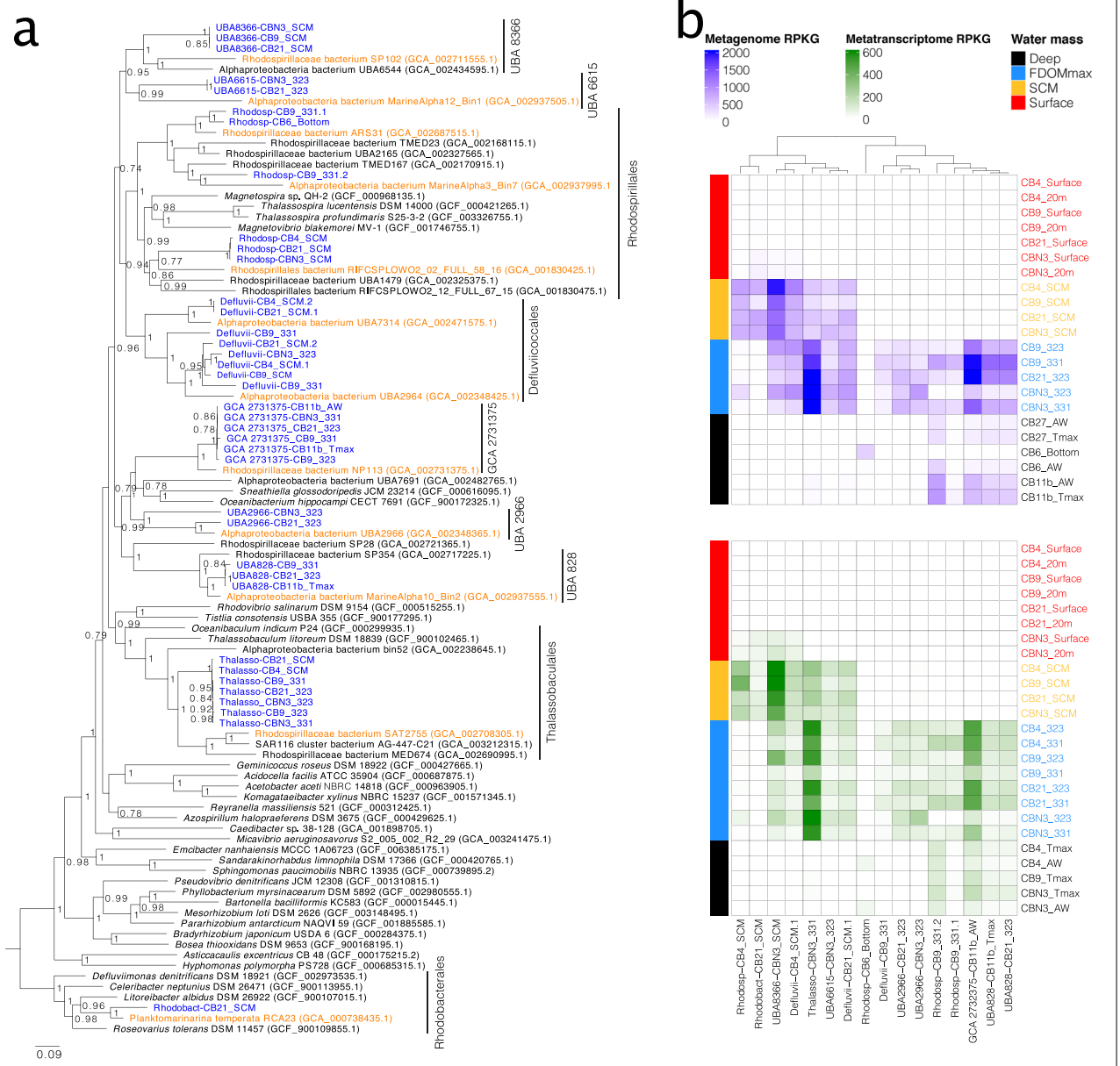

Fig. 5 **a** Phylogenetic tree of the *Alphaproteobacteria* MAGs identified as most implicated in the degradation of aromatic compounds in the Arctic Ocean. Blue, MAGs reconstructed in this study; orange, genomes from other studies identified as the closest relatives to the MAGs reconstructed in this study; black, other genomes. **b** Geographic distribution of the *Alphaproteobacteria* MAGs identified as most implicated in the degradation of aromatic compounds within the Canada Basin. Top panel, mapping of the metagenomes on the MAGs. Bottom panel, mapping of the metatranscriptomes on the MAGs

Global ocean distribution of *Alphaproteobacteria* MAGs from Canada Basin

We sought to determine if the MAGs most implicated in aromatic compound degradation were more broadly distributed beyond the Arctic Ocean. We therefore investigated the distribution of the *Alphaproteobacteria* MAGs (Table S5) by fragment recruitment against a set of 151 metagenomes broadly representative of the global ocean microbiome (Fig. 6). Of the 16 representative MAGs, two were commonly detected outside of the Arctic Ocean. Rhodosp-CB9_331.2 was identified in the mesopelagic metagenomes from all oceanic regions, but not the Mediterranean Sea or Red Sea. The *Rhodobacterales* MAG (Rhodobact-CB21_SCM)

was also identified in surface water metagenomes, most notably from the Southern Ocean. Although several other MAGs were detected at low frequency in the Southern Ocean (e.g., Rhodosp-CB6-bottom and Defluvii-CB9-331), the majority (12 MAGs, 75% of the MAGs) were not detected outside of the Arctic Ocean. Likewise, the vast majority of the most closely related marine reference genomes were not detected in Canada Basin metagenomes. The exceptions were *Planktomarinamarina* (*Rhodobacterales*) and the reference genome within UBA828.

Enhanced aromatic compound degradation capacity in *Alphaproteobacteria* MAGs restricted to the Canada Basin

We hypothesized that an enhanced capability for aromatic compound degradation may be implicated in the evolutionary adaptation of the Arctic Ocean populations. We therefore compared the abundance and diversity of aromatic compound degradation genes between the Arctic MAGs and the set of their closest relatives. Between 1.5 and 4% of the genes with an EC annotation was annotated with an EC from lignin-derived aromatic

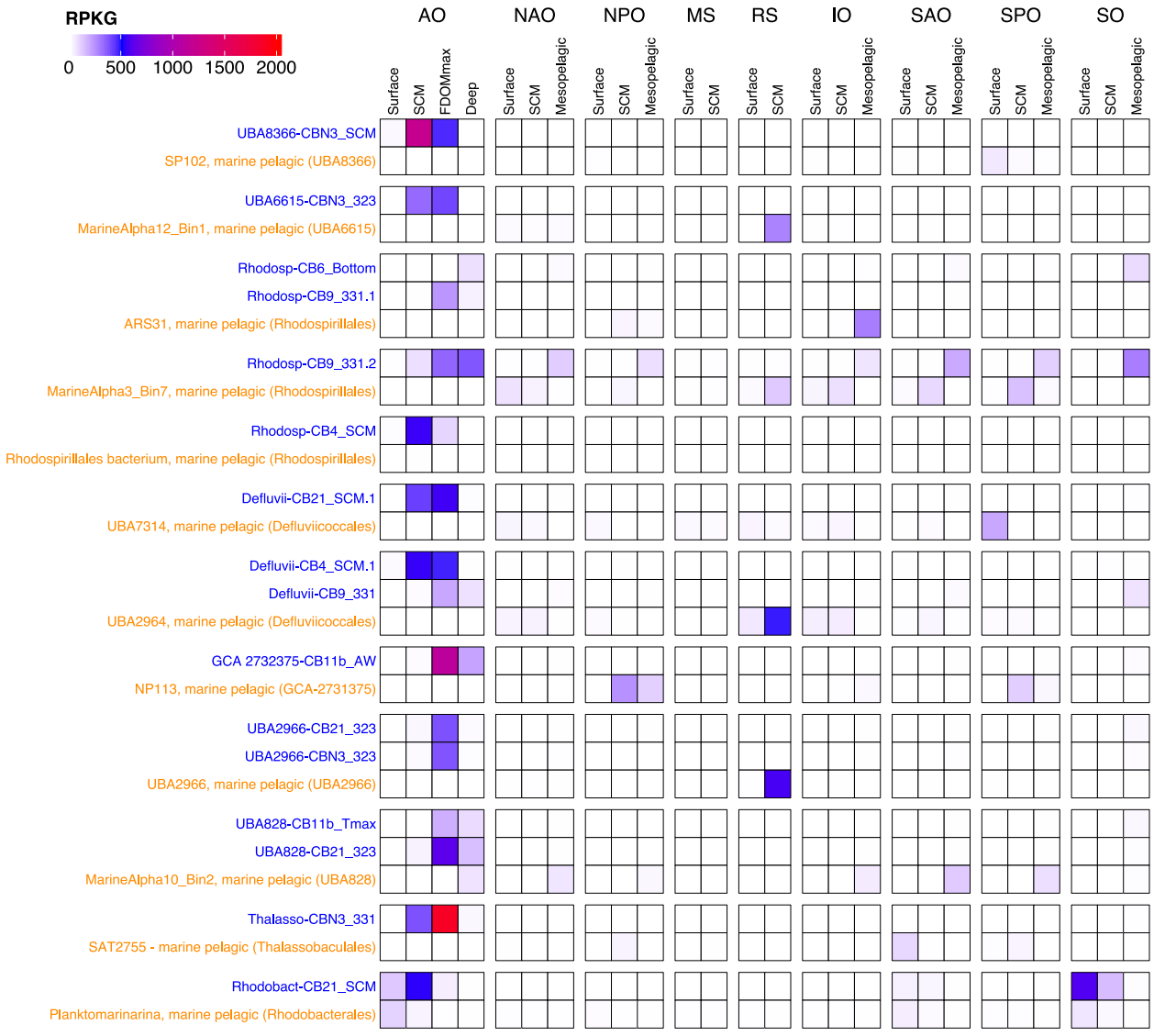

Fig. 6 Fragment recruitment of global oceans metagenomes on the *Alphaproteobacteria* MAGs identified as most implicated in the degradation of aromatic compounds. Blue, MAGs identified in this study and representative of the 16 genomospecies; orange, genomes from other studies identified as the closest relatives to the MAGs reconstructed in this study. AO, Arctic Ocean; IO, Indian Ocean; MS, Mediterranean Sea; NAO, North Atlantic Ocean; NPO, North Pacific Ocean; RS, Red Sea; SAO, Southern Atlantic Ocean; SO, Southern Ocean; SPO, Southern Pacific Ocean

compound degradation pathways for both the Arctic MAGs and genomes from other oceans (Fig. 7). However, out of the 16 Arctic Ocean MAGs, 10 possessed a higher diversity of marker genes, and 14 possessed a higher number of marker gene copies than their sister taxa from other oceans (Fig. 7). This is despite the difference in genome completeness estimated for the reference MAGs (64–98%) compared to the Arctic Ocean MAGs (range of 42–95% completeness). Some marker genes were found exclusively in our Arctic *Alphaproteobacteria* MAGs, including those for the degradation of catechol to 3-oxoadipate, phenol 1, dehydrodivanillate, protocatechuate 1, cinnamate, and resorcinol (Fig. 7). Other marker genes were found exclusively in the reference MAGs, such as genes for the degradation of benzoyl-CoA II, gallate I and III, m-cresol, and *γ*-resorcylate.

Every single Arctic MAG contained at least one and up to seven marker genes that were not found in its closest related genome from other oceans (Fig. 7). These included gentisate 1,2-dioxygenase (K00450) in Rhodosp-CB4_SCM, Defluvii-CB9_331, UBA2966-CBN3_323, Thalasso-CBN3_331, and Rhodobact-CB21_SCM, as well as vanillate O-demethylase (K03862) in Rhodosp-CB6_Bottom and Rhodosp-CB9_331.1 or benzoate CoA-ligase (K04110) in UBA6615-CBN3_323 and Rhodosp-CB9_331.2. These genes were the most abundant in the metagenomes and metatranscriptomes and most common throughout the Arctic Ocean microbiomes (Fig. 4 a–c). The three MAGs with the highest number of aromatic compound degradation gene copies (Defluvii-CB21_SCM.1, Rhodosp-CB9_331.1, and Defluvii-CB9_331) all possessed more copies of the vanillate O-demethylase (K03862) gene than their sister genomes. In addition, all of their sister genomes lacked the protocatechuate ring-opening genes (K04101 or K00449). A vast majority of the marker genes were also expressed in our Arctic Ocean MAGs (Fig. 7, right panel), and we found only gentisate 1,2-dioxygenase (K00450) slightly expressed in the MarineAlpha10_Bin2 of the reference genomes.

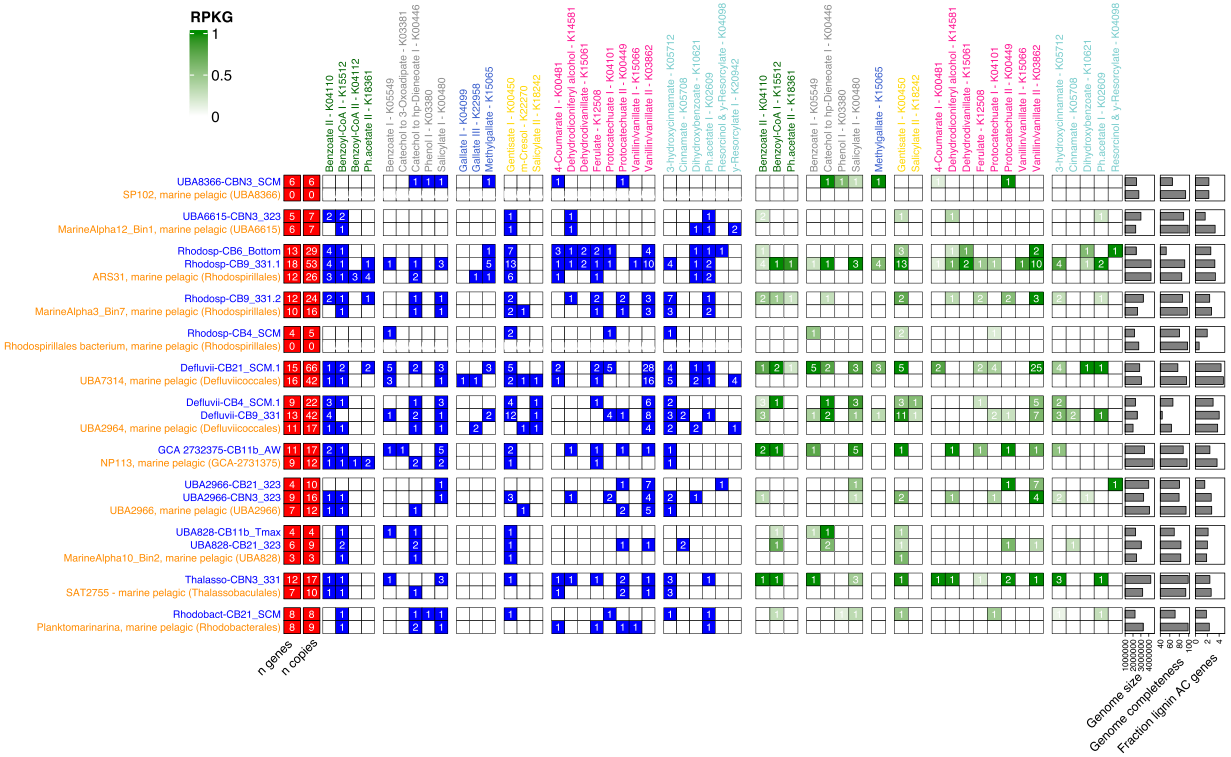

Fig. 7 Comparative genomics of the Arctic Ocean *Alphaproteobacteria* MAGs identified as most implicated in the degradation of aromatic compounds (blue) and their closest relative genomes (orange) from other studies. Left panel: graphical table of the presence (blue cells) and absence (white cells) of genes annotated with KOs marker of aromatic compound degradation pathways. Number in the cells represents the number of genes annotated with a KO. Right panel: graphical table of the mean expression (green cells) and the absence of expression (white cells) of genes annotated with KOs marker of aromatic compound degradation pathways. Number in the cells represent the number of genes expressed

Discussion

Aromatic compound degradation capacity of the Arctic Ocean microbiomes reflects the ability to degrade humic-rich DOM from terrestrial and sediment origin

The degradation of aromatic compounds emerged as a central metabolism of the Canada Basin microbiomes, which is consistent with a capacity to degrade the humic-rich DOM in the Arctic Ocean. We found that the aromatic compound degradation genes contributed more to the total metabolic capacity of the Arctic Ocean microbiome compared to other oceans, paralleling the higher concentrations of HS in the Arctic Ocean [18]. This enhanced aromatic compound degradation capacity was associated with the humic-rich tOM layer of the Canada Basin (FDOMmax). Individually, most of the aromatic compound degradation genes were more abundant in the FDOMmax (Fig. 3a), where humic-rich DOM concentrations are maximal, as evidenced by the distribution of the FDOM C1 fraction (Fig. 1c). The lower abundance of aromatic compound degradation genes in the surface (Fig. 3a) may be explained by a preference of the surface microbiomes to process the nonaromatic tOM fraction as sunlit waters are generally poor in aromatic compounds. This is supported by the higher contribution of the non-photoreactive (non-aromatic) FDOM C4 fraction and very low contribution of the aromatic C1 fraction in the surface (Fig. 1c).

We found that the distribution and diversity of aromatic compound degradation genes in the Canada Basin matched the diversity of aromatic compounds expected from the Arctic Ocean watershed organic matter. Lignin is a polymer of three aromatic monolignols, which form the H, G, and S unit (defined by 0, 1, and 2 methoxy group on the aromatic ring, respectively) when cross-linked. Coniferous trees dominate the boreal forest of the Arctic Ocean watershed and are characterized by a high G/S ratio in their lignin polymer [50]. vanillin (G-unit) is therefore expected in the tOM of the riverine input to the Arctic Ocean. This has been shown in the Mackenzie River, the major river draining to the Canada Basin; the OM of suspended sediments in the Mackenzie was dominated by 1-methoxy aromatic compounds (G), among which vanillin and vanillate contributed the most [51]. The Mackenzie River also contained a significant contribution of benzoate derivatives (H) and smaller amounts of syringate derivatives (S). Our results showed that the genes involved in the degradation of vanillate, benzoate, and methylgallate (syringate derivatives) were among the most abundant and expressed genes (Fig. 3 a–b) in the Canada Basin microbiomes, paralleling the aromatic compounds composition of the McKenzie River DOM input.

Sediments also contribute to the humic-rich OM reaching the Canada Basin, by exchanging OM with brine sinking along the shelf [16]. The OM on the MacKenzie shelf, bordering the Canada Basin, contained higher levels of vanillyl, syringyl, and cinnamyl phenols than any other North American Arctic shelf [52]. The high abundance and expression of genes involved in the degradation of vanillate, syringate, and 3-hydroxycinnamate (Fig. 3 a–b) suggest that the Canada Basin microbiomes access the variety of aromatic compounds from shelf sediments sources. This demonstrates that the Canada Basin microbiomes can access humic-rich DOM from terrestrial and sediment origin as growth substrates.

Rhodospirillales are implicated as aromatic compound degraders in the Arctic Ocean

We showed that a few clades of *Alphaproteobacteria* were strongly implicated in aromatic compound degradation (Figs. 4, S8–S9). Previous work in the Arctic Ocean showed that *Gammaproteobacteria* are associated with humic-rich DOM degradation. For example, in an experiment adding humic-rich OM derived from a thermokarst to coastal Arctic water (Chukchi sea) microbiomes, *Gammaproteobacteria* taxa *Colwelliaceae* (order *Alteromonadales*) rapidly dominated the microbial communities [53]. Similarly, *Alteromonadales* (genera *Glaciecola*, SAR92 clade) were associated with humic-rich riverine-derived OM consumption in an Arctic fjord [54]. These studies may seem incongruent with our results. However, these studies focused on the microbiomes of the surface waters only. In our study, we observed a significant *Gammaproteobacteria* signal in the taxonomy of aromatic compound degradation genes within the surface samples (Figs. 3, S7), which supports earlier studies. In addition, the tOM used in previous experiments contained many other compounds than aromatic compounds, including more labile protein-like compounds. The *Gammaproteobacteria* in the surface could then be adapted to a fast consumption of pulses of labile OM, while the *Rhodospirillales* taxa of our study were more adapted to a slow degradation of more refractory and steadier amount of aromatic compound-rich humic substances in the FDOMmax. This would be in line with the high variability in seasonal conditions and DOM concentrations in the surface waters of the Arctic Ocean compared to more stable DOM concentrations and conditions in the FDOMmax throughout the year [55, 56].

All but one of the MAGs most implicated in aromatic compound degradation belonged to closely related *Alphaproteobacteria* clades, based on the GTDB. Based on

NCBI taxonomy, all but one of these MAGs belonged to the *Rhodosprillales* order. Here, we used the NCBI taxonomy to be able to relate our findings to previous reports in the literature. We concluded that the capacity to degrade aromatic compounds is phylogenetically concentrated in *Rhodosprillales* within the Arctic Ocean. In the global ocean, a few studies previously identified *Rhodospirillales* taxa in aromatic compound- and HS-degrading consortia. A *Rhodospirillales* strain (*Thalassospira profundimaris*) was one of six taxa isolated from the East China Sea surface microbiomes enriched with vanillic acid [57]. This strain was able to grow on benzoic acid, 4-hydroxybenzoic acid, and to a lesser extent on syringate and ferulate. However, this *Rhodospirillales* strain is a member of a different clade (also including *Magnetospira* and *Magnetovibrio*) than the *Rhodospirillales* genomes we report in our study (Fig. 4a). *Rhodospirillales* were also identified in flow-through experiments in which marine microbiomes were exposed to riverine HS as a sole carbon source [43]. *Rhodospirillales* represented 6% of the taxa identified and were only reported in the low salinity experiment (14 PSU). Taxa were found in the *Thalassospira* (4 taxa) and *Thalassobaculales* (1 taxa) clades. We also reported *Thalassobaculales* genomes within the taxa implicated in aromatic compound degradation. However, the genomes we reported were located in the water column at salinities > 30 PSU. The studies reporting *Rhodospirillales* focused only on the surface water microbiomes, while we investigated the whole water column. The focus on surface waters in other studies is usually based on the assumption that aromatic compounds and HS have a terrestrial origin and are transported to the ocean with freshwater input and therefore concentrated in the surface layers. The focus on surface water microbiomes may therefore explain why *Rhodospirillales* have not yet been reported as most implicated in aromatic compounds and HS degradation within the ocean. Based on the distribution of HS in the Arctic Ocean, we investigated the whole water column and specifically the FDOMmax, which allowed us to identify *Rhodospirillales* as strongly implicated in aromatic compounds degradation. Further work within other oceans, as well as experimental work using HS as sole carbon sources in the microbiomes of the FDOMmax, will be necessary to fully elucidate the role of *Rhodospirillales* in the degradation of HS in the Arctic Ocean and the global ocean.

Evolutionary adaptation of *Rhodospirillales* in the Arctic Ocean

The phylogenetic divergence of our *Rhodospirillales* MAGs from relatives in other oceanic regions and their restricted distribution to the Arctic Ocean suggests the Arctic populations are evolutionarily adapted to life in the Arctic Ocean. The high number of aromatic compound degradation gene copies in the MAGs compared to their closest relatives in other oceans suggest that the capacity to use aromatic compounds as a growth substrate played a role in their evolutionary adaptation. The disproportionately high amount of HS in the FDOMmax may then act as a selective pressure on these MAGs. Previous studies have demonstrated evolutionary adaptation of microbial MAGs restricted to the Canada Basin: a new *Methylophilaceae* clade [58] as well as SAR11 [59] and *Chloroflexi* ecotypes [45]. The *Methylophilaceae* clade evolved via a freshwater to marine transition, highlighting the importance of the terrestrial-marine interface in shaping the Canada Basin microbiomes. However, HS does not appear as the main selective pressure for *Methylophilaceae* as their distribution is restricted to the surface. The Arctic SAR11 and *Chloroflexi* clades were restricted to the SCM and FDOMmax, where humic-rich DOM is enriched within the Canada Basin. The *Chloroflexi* Arctic ecotype was replete with aromatic compound degradation genes, some of these acquired by lateral gene transfer from terrestrial taxa. This *Chloroflexi* ecotype was found within the water masses rich in HS (FDOMmax), similarly to the *Rhodospirillales* MAGs of our study. The preference for humic-rich water masses coupled to an enhanced capacity to degrade aromatic compounds in *Chloroflexi* and *Rhodospirillales* suggests that the ability to use aromatic compounds as growth substrate provides an evolutionary advantage in the humic-rich environment of the Canada Basin FDOMmax.

Conclusion

The dissolved organic matter of the Arctic Ocean is characterized by a disproportionately high contribution of HS compared to other oceans. With the increasing terrestrial input of humic-rich OM to the Arctic Ocean as a result of escalating permafrost thawing and river runoff, it is predicted that the contribution of HS to the Arctic Ocean OM will increase [26]. The fate of this carbon is important to consider with respect to changing biogeochemical cycles of the Arctic Ocean. In this study, we showed that the metabolic pathways involved in the degradation of HS were widespread, abundant, and expressed in the microbiomes of the Canada Basin. The capacity to degrade humic-rich OM in the Arctic Ocean microbiomes was enhanced compared to the microbiomes of the global ocean in the upper water column. The diversity and distribution of the aromatic compound degradation machinery revealed that the Arctic Ocean microbiomes were equipped to use OM from terrestrial sources as growth substrates. We identified that the aromatic compound degradation capacity was concentrated phylogenetically in *Rhodospirillales*. The

phylogeny, comparative genomics, and biogeographic distribution of these *Rhodospirillales* suggest an evolutionary adaptation driven by the disproportionately high amount of HS in the Arctic Ocean. Overall, this study demonstrates that the Arctic Ocean microbiomes are capable of processing OM of terrestrial origin. Our study predicts that OM of terrestrial origin can be remineralized in the Arctic Ocean, and that *Rhodospirillales* will gain importance as tOM inputs continue to increase in the Arctic Ocean.

Methods

Sampling, DNA, and RNA extraction

Samples were collected in September 2017 during the Joint Ocean Ice Study cruise to the Canada Basin. We analyzed 22 metagenomes and 25 metatranscriptomes generated from samples collected across the water column of the Canada Basin. Eight specific water masses were sampled: the surface mixed layer (surface: 5 m and 20 m depth) characterized by fresher water due to riverine input and ice melt; the subsurface chlorophyll maximum (SCM), in the halocline (FDOMmax at salinity of 32.3 and 33.1 PSU, referred as 32.3 and 33.1); and deeper water from Atlantic origin at the temperature maximum (referred as Tmax), 1000 m depth (Atlantic water, further referred as AW), and 10 or 100 m above the bottom (further referred as bottom).

We filtered 14 L of seawater for DNA samples and 7 L of seawater for RNA samples sequentially through a 3 µm pore size polycarbonate track etch membrane filter (AMD manufacturing, ON, Canada) and a 0.22 µm pore size Sterivex filter (Millipore, MA, USA). Filters were stored in RNALater (Thermo Fisher, MA, USA) and kept frozen at −80 °C until processing in the lab. DNA was extracted following the method described in [60]. Briefly, the preservation solution was expelled and replaced by a SDS solution (0.1 M Tris-HCl pH 7.5, 5% glycerol, 10 mM EDTA, 1% sodium dodecyl sulfate) and incubated at room temperature for 10 min and then at 95 °C for 15 min. The cell lysate was then centrifuged at 3270 × g. Proteins were removed by precipitation with MCP solution (Lucigen, WI, USA), and the supernatant was collected after centrifugation at 17,000 × g for 10 min at 4 °C. DNA was precipitated with 0.95 volume of isopropanol and rinsed twice with 750 µL ethanol before being air dried. The DNA was resuspended in 25 µL of low TE buffer and pH 8 (10 mM Tris-HCl, 0.1mM EDTA) and stored at −80 °C.

The RNA extraction procedure was adapted from the mirVana RNA extraction kit (Thermo Fisher, MA, USA). RNAlater was expelled from the Sterivex and replaced by 1.5 mL of lysis buffer, and Sterivex was vortexed. A total of 150 µL of miRNA homogenate were added, and the Sterivex was vortexed and incubated on ice for 10 min. The cell lysate was expelled from the Sterivex, 0.9× the volume of acid-phenol-chloroform was added, and the solution was vortexed for 30–60 s. The mix was centrifuged at 10,000 × g for 5 min, and the top aqueous phase gently removed and transferred to a fresh tube. A total of 1.25 volume of 100% ethanol was added to the aqueous phase and vortexed to mix. The mix was filtered through mirVana Filter Cartridges by centrifugating at 10,000 × g for 10 s, and the flow through discarded. The RNA was rinsed with 700 µL of Wash Solution 1 and then with 500 µL Wash Solution 2/3 by centrifugating at 10,000 × g for 10 s. RNA was then eluted with 50 µL of Elution Solution (0.1 M EDTA) warmed at 95 °C. 700 µL of RTL buffer, and 500 µL 100% ethanol was added to the RNA suspension, and the suspension was centrifuged for 15 s at 10,000 × g on a RNeasy MinElute column. RNA was washed first with RPE buffer by centrifuging 500 µL for 15 s at 10,000 × g and then 80% ethanol for 2 min at 10,000 × g. The empty column was then centrifuged at 12,000 × g for 5 min to discard the excess liquid. The RNA was finally eluted by centrifugation of 28 µL and then 10 µL of RNase-free water for 1 min at 12,000 × g and stored at −80 °C.

Dissolved organic matter samples collection and analysis of fluorescence measurements

DOM samples were collected in Niskin bottles mounted on a conductivity-temperature-depth rosette profiler and immediately filtered using pre-combusted 0.3 µm glass fiber filters (GF75, Advantec) into pre-combusted amber glass vials. Fluorescence spectra were measured using a FluoroMax 4 Jobin Yvon fluorometer [26]. Parallel factor analysis was applied to decompose the fluorescence signal into their main components following the procedures outlined in Murphy et al. [61] The PARAFAC model validated 7 components including 5 humic-like (C1-C2, C4-C5, and C7) and 2 protein-like components (C3 and C6) in 4483 samples collected from surface to 10 m above bottom sediment in the Canada Basin between 2007 and 2017 [26].

Metagenomic sequencing, assembly, and annotation

Sequencing, assembly, and annotation were performed by the Joint Genome Institute (CA, USA). Each individual metagenome and metatranscriptome were sequenced on the Illumina NovaSeq platform, generating paired-end reads of 2 × 150 bp for all libraries. Single assemblies were created for each individual sample using SPAdes [62] with k-mer sizes of 33, 55, 77, 99, and 127 bp. Gene prediction and annotation were performed using the DOE-Joint Genome Institute Integrated Microbial Genomes Annotation Pipeline v.4.16.5 [63].

Building of EC and KO abundance matrices

The number of copies of genes annotated to each Enzyme Commission (EC) number or KEGG Orthology (KO) number in a metagenome or metatranscriptome was calculated by summing the depth of coverage of all genes or transcripts annotated with this EC or KO. To obtain the final EC and KO abundances (number of gene or transcript copies/10^6 reads for this EC or KO), the total number of copies was then normalized by the library size (number of reads) with the TMM method [64], using the calcNormFactors function of the edgeR package in *R* [65]. Genes were assigned to a total of 3102 EC numbers and 12,018 KO identifiers for the metagenomes and 2830 EC numbers and 10,556 KO identifiers for the metatranscriptomes. Before multivariate analysis, EC and KO abundances were transformed with a Hellinger transformation (decostand function from the R vegan package [66]).

Multivariate analysis

Nonmetric multidimensional scaling (NMDS) was performed on the EC number abundance matrices with the metaMDS function from the *R vegan* package, using two dimensions and the Bray-Curtis dissimilarity metric. Nonnegative matrix factorization (NMF) was performed with the nmf function from the NMF package in R [67]. NMF decomposes the abundance matrix into two matrices: a coefficient matrix that describes the overall structure of the abundance matrix with a limited number of descriptors (called sub-metagenomes and sub-metatranscriptomes in this study, their number being the rank) and a basis matrix that provides the weights of each original descriptors (EC number) on the new descriptors (sub-metagenomes, sub-metatranscriptomes). The advantage of NMF is to directly link the overall structure of the abundance matrix to the individual elements (EC number) driving this structure. We first performed the NMF analysis with rank values ranging from 3 to 7, 100 runs, and various algorithms ("nsnmf," "Brunet," "KL"). We obtained the optimal results for the nsNMF algorithm, random seed of the factorized matrices, and a rank value of 4. We performed the final analysis with 200 runs, rank of 4, random seed, and nsNMF algorithm.

Calculation of gene and pathway indices

The indices were calculated for EC number annotated from metagenomes and metatranscriptomes by combining two methods described in Jiang et al. [68] and Kim et al. [69]. We first used the EC number abundance matrices (annotated from metagenomes and metatranscriptomes) and the coefficient matrices (SMG/SMT × samples) to calculate both the spearman correlation coefficient and the multidimensional projection between all pairs of EC number annotated from metagenomes (ECMG) and SMG as well as all pairs of EC number annotated from metatranscriptomes (ECMT) and SMT. The spearman correlation coefficient between an ECMG/ECMT (i) and a SMG/SMT (k) $\rho i,k$ was calculated using the abundance profile of a ECMG/ECMT and a SMG/SMT along all the samples. The multidimensional projection between an ECMG/ECMT and a SMG/SMT was calculated as the cosine of the angle between the vectors represented by an ECMG/ECMT abundance in the samples space and the vector represented by SMG/SMT in the samples space. The abundance profiles of ECMG/ECMT and SMG/SMT were first normalized, and the multidimensional projection was calculated as follows:

$$Cos\theta_{i,k} = \sum_{j=1}^{n} a_{i,j} \times S_{k,j}$$

where $Cos\theta i,k$ is the multidimensional projection between the ECMG/ECMT i and the SMG/SMT k, n is the number of samples, ai,j is the normalized abundance of the ECMG/ECMT i in the sample j, and sk,j is the normalized abundance of the SMG/SMT k in the sample j. We then used the basis matrix to calculate the score of each ECMG/ECMT as follows:

$$Score(i) = 1 + \frac{1}{\log_2(q)} \sum_{k=1}^{q} p(i,k)\log_2(p(i,k))$$

where i is the ECMG/ECMT, q is the number of SMG/SMT (4 in our study), k is the SMG/SMT, and $p(i,k)$ is the probability of finding the ECMG/ECMT i in the SMG/SMT k.

We calculated the final EC index (annotated from metagenome and metatranscriptome) on each SMG/SMT by multiplying the spearman correlation coefficient, cos theta, and the EC score:

$$I_{i,k} = \rho_{i,k} \times Cos\theta_{i,k} \times Score\ (i)$$

This allowed us to calculate an index for each pair of ECMG/ECMT and SMG/SMT.

Aromatic compound degradation gene and pathway selection

To select the metabolic pathways and enzymes involved in the degradation of lignin-derived aromatic compounds, we first surveyed the literature for the various lignin breakdown compounds reported to be degraded by bacteria [36, 70, 71]. We then retrieved the MetaCyc (https://metacyc.org) pathways that were involved

in the degradation of these compounds, as well as the EC numbers involved in these pathways. For analysis of marker genes for each pathway, we first selected one key reaction in each pathway. We then retrieved genes annotated with KO numbers corresponding to the EC numbers associated with the key reactions. We first chose reactions involved in aromatic ring-opening, aromatic ring-oxidation, and aromatic ring-reduction steps. If the EC numbers of these reactions were not specific to a pathway, we chose reactions involved in the addition of CoA to the aromatic ring or reactions involved in oxidoreduction steps of the side chains of the aromatic ring. If 2 aromatic compounds degradation pathways possessed the same EC number associated with the selected key reaction, and if this EC number was not found on any other Metacyc pathways, we used the EC as marker for only one of the two pathways. We could not retrieve any marker EC number for several pathways.

Calculation of the fraction of genes involved in the degradation of aromatic compounds within the pool of metabolic genes

To calculate the percentage of the gene pool associated with the degradation of aromatic compounds in each sample, we summed the total number of genes copies annotated with EC number within our selection of aromatic compound degradation pathways. We then divided this number by the total number of gene copies annotated with EC numbers.

Statistical analyses

To compare the means of EC indices distribution as well as the percentage of aromatic compounds degradation genes in metagenomes, we first performed a permutational ANOVA (PERMANOVA) using the perm.anova function of the RVAideMemoire package in R. When the *p*-value of the PERMANOVA test was less than 0.05, we performed pairwise comparison between groups with Student *t*-test, using the PERMANOVA residuals variance as the variance for the Student *t*-tests. Two groups were considered different if their *p*-value was less than 0.05.

Calculation of aromatic compound degradation pathways completeness

Pathway completeness in a sample was calculated by dividing the number of EC numbers belonging to this specific pathway and present in a sample by the total number of EC number of this specific pathway. The marker genes abundance was obtained using the normalized abundance of a specific KO as calculated to obtain the KO abundance matrices (see above).

Estimation of the fraction of taxa harboring aromatic compound degradation genes

The estimated percentage of genomes harboring marker genes in a sample was calculated by dividing the total number of gene variants annotated with the KO of interest by the total number of gene variants annotated with the KO corresponding to the *recA* gene (K03553).

Taxonomic assignment of aromatic compound degradation genes

Genes annotated with KO identified as marker for selected lignin-derived aromatic compound degradation pathways were grouped by water column feature and dereplicated with CD-hit (v.4.6) [72] at 95% identity. The dereplicated set of genes was searched against the NCBI nr protein database (downloaded 21 August 27) using DIAMOND (v.0.9.30.131) [73]. To assign a taxonomic identity to these genes, the DIAMOND (v.0.9.30.131) output was imported in MEGAN [73] using the January 2021 mapping file ("megan-map-Jan2021.db"). The lowest common ancestor parameters were set at minimum *e*-value of 1×10^{-20} and at top percent of 1%. The file containing taxonomic identity of the genes was then exported from MEGAN and processed with a custom-made *R* script.

Metagenome binning

Metagenomic binning was performed on each individual assembly with Metabat2 (v.2.12.1) [74] using scaffold longer than 2500 bp. Contamination and completeness of the metagenome-assembled genomes (MAGs) were estimated with CheckM (v.1.0.7) [75]. MAGs greater than 30% completeness and less than 10% contamination were selected for further analysis. Phylogenetic placement of MAGs was performed based on the concatenation of 120 conserved genes for bacteria and 122 conserved gene for archaea using the Genome Database Taxonomy Database toolkit (GTDB-Tk — v.1.3.0) [49, 76].

Metabolic reconstruction and MAG selection based on the capacity to degrade aromatic compounds

To select MAGs enriched in aromatic compound degradation capacity, we selected all the genes annotated with EC numbers belonging to pathways involved in the degradation of aromatic compounds within the MAGs. As ring-fission pathways can be involved in the degradation of non-lignin aromatic compounds, we used only the funneling pathways to select for MAGs enriched in the degradation of lignin aromatic compounds. For each MAG, we calculated the completeness of all funneling aromatic compound degradation pathways. We considered a pathway complete if a MAG contained genes annotated with all the EC number of this pathway. The

pathway completeness percentage was obtained by dividing the number of EC numbers involved in a pathway within a MAG by the number of reactions of this pathway. This number was normalized by the MAG completeness. For each MAG, we then calculated the median of all pathway completeness. Based on the distribution of the medians of all MAGs, we selected 4% as the median threshold above which a MAG was selected as having a high capacity to degrade aromatic compounds. We obtained a total of 46 MAGs. We calculated the average nucleotide identity (ANI) between these 46 MAGs using fastANI (v.1.3) [77] and grouped the MAGs with an *ANI* > 95% as the same genomospecies, obtaining a total of 22 genomospecies.

Phylogenetic analyses of MAGs

To reconstruct the phylogeny of the 38 Alphaproteobacteria MAGs (16 genomospecies), we first manually investigated their phylogenetic placement within the GTDB. For each of our selected MAGs, we picked the most closely related genomes from the GTDB, as well as genomes representative of distinct families. We then reconstructed a phylogeny with our selected MAGs and the selected genomes from GTDB using concatenation of 120 conserved genes and FastTree [49].

Metagenome and metatranscriptome fragment recruitment

In order to evaluate the abundance and overall expression of our selected MAGs across the samples, we mapped the reads of the metagenomes and metatranscriptomes to our selected MAGs using bbmap (v.35) and a minimum sequence identity of 98%. We then calculated final RPKG values (reads per MAG kilobase pairs per metagenome giga base pairs), by dividing the total number of reads mapped to each MAG, by the size of the MAG (kbp) and the size of the metagenome/metatranscriptome (Gbp).

In order to evaluate the distribution of selected Alphaprotebacteria MAGs and their most closely related reference genomes across oceans, we performed fragment recruitment as in Kraemer et al. [78]. The 16 MAGs representative of the genomospecies enriched with the capacity to degrade aromatic lignin moieties, and 12 closely related reference genomes were searched against the metagenomic dataset using blastall (v.2.2.25) (*e*-value = 0.00001). The recruited reads were extracted from the metagenomes and searched against a database consisting of the concatenation of all 28 genomes (16 Arctic Alphaproteobacteria MAGs and 12 reference genomes) using blastall (v.2.2.25). We selected the best hit and filtered for a minimum of 100 bp alignment and 98% sequence identity. We then calculated the RPKG values by normalizing the number of reads recruited by kilobase of genome and gigabase of metagenome.

Annotation of publicly available genomes

Gene sequences were retrieved from publicly available genomes at NCBI (Table S4) and translated to proteins. Ribosomal RNA genes were predicted in Infernal v. 1.1.2 [79] against Rfam v. 14.2 [80]. Gene functions were annotated in KofamScan using default settings and a bitscore-to-threshold ratio of 0.7 [81].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-022-01417-6.

Additional file 1: Figure S1. Estimation of the rank for the NMF analysis of the EC abundance matrices annotated from metagenomes (top panels) and metatranscriptomes (bottom panels). Left panels: evolution of various parameters as a function of the rank used in the NMF analysis. Cophenetic correlation represents the correlation between the sample distances from the consensus matrix and the cophenetic distance between these samples when they are clustered. The rss is the residual sum of squares between the original EC abundance matrix and its estimate using the NMF algorithm. The dispersion is defined as $1\text{-rss}/\Sigma_{i,j} (V_{i,j})^2$ ($V_{i,j}$ are the entries of the EC abundance matrix) and estimates the fraction of variance of the EC abundance matrix explained by the NMF results. Residuals is the sum of residuals between the original EC abundance matrix and the matrix estimated using the NMF. Right panels: consensus matrices based on clustering the coefficient matrices at each of the 100 runs of the NMF analysis. The heatmap represents the fraction of times 2 samples fall in the same clusters out of 100 runs. **Figure S2.** Heatmaps of the basis matrix (left) and coefficient matrix (right) obtained after running an NMF analysis on the EC abundance matrix annotated from metagenomes and using a rank value of 4. **Figure S3.** Heatmaps of the basis matrix (left) and coefficient matrix (right) obtained after running an NMF analysis on the EC abundance matrix annotated from metatranscriptomes and using a rank value of 4. **Figure S4.** Lignin-derived aromatic compound degradation pathways completeness in the 4 water column features (surface, SCM, FDOMmax, and deep water) for metagenomes (left) and metatranscriptomes (right). **Figure S5.** Normalized abundance per water column feature of KO number markers of aromatic compounds degradation pathways annotated from metagenomes. **Figure S6.** Normalized abundance per water column feature of KO numbers markers of aromatic compounds degradation pathways annotated from metatranscriptomes. **Figure S7.** Estimated fraction of the microbiome harboring genes annotated with KO marker of aromatic compounds. **Figure S8.** Completeness and contamination of the set of 1772 MAGs reconstructed from 22 individual metagenomes. The vertical and horizontal dotted lines represent 10% contamination and 30% completeness respectively. Selected MAGs are the 46 MAGs that have been selected as most implicated in lignin-derived aromatic compound degradation based on the number and completeness of their aromatic compound degradation pathways. **Figure S9.** Taxonomic identity of the MAGs harboring genes annotated with KO marker of aromatic compounds degradation pathways. The taxonomy is displayed at the class level. **Figure S10.** Phylogenetic tree reconstructed from the concatenation of 120 conserved genes for the bacterial genomes of the MAGs dataset. The heatmap represents the completeness of the lignin-derived aromatic compounds degradation funneling pathways. Red stars correspond to the MAGs selected based on the amount and completeness of pathways they harbor. **Figure S11.** Taxonomic identity of the MAGs most implicated in the degradation of lignin-derived aromatic compounds. Taxonomy is displayed at the order level and based on the tree placement of the MAGs, using the GTDB. **Figure S12.** Heatmap of

the average nucleotide identity for the 46 MAGs most implicated in the degradation of lignin-derived aromatic compounds.

Additional file 2: Table S1. Metadata and physicochemical properties of the samples collected for this study. **Table S2.** List of metabolic pathways (obtained from metacyc) involved in the degradation of lignin-derived aromatic compounds and their EC numbers. KO numbers, protein names and enzyme functions information is included for reactions identified as markers for these pathways. **Table S3.** Properties of the 46 MAGs most implicated in the degradation of aromatic compounds. Genome size, completeness, contamination, N50 values, number of genes and GC content were obtained using *CheckM*. The GTDB taxonomy was obtained using the nearest placement in a phylogenetic tree with the *GTDBTk*. The NCBI taxonomy was obtained using the nearest placement in a phylogenetic tree with *CheckM*.**Table S4.** Information and sources of genomes from other studies used in this study. **Table S5.** Properties of the 16 *Alphaproteobacteria* MAGs most implicated in the degradation of aromatic compounds and representative of the genomospecies, as well as the 12 publicly available genomes identified as their closest relatives. **Table S6.** Data availability.

Acknowledgements
The data were collected aboard the CCGS Louis S. St-Laurent in collaboration with researchers from Fisheries and Oceans Canada at the Institute of Ocean Sciences and Woods Hole Oceanographic Institution's Beaufort Gyre Exploration Program and are available at http://www.whoi.edu/beaufortgyre. We would like to thank both the captain and crew of the CCGS Louis S. St-Laurent and the scientific teams aboard. We also thank Québec-Océan for their scientific and financial support.

Authors' contributions
DAW designed the study. TG generated the metagenomic and metatranscriptomc data and performed the bioinformatic analyses. TG and DAW wrote the manuscript. VO performed the functional annotation of the genomes. The authors read and approved the final manuscript.

Funding
The work was conducted in collaboration with "Facilities Integrating Collaborations for User Science" (FICUS) program between the Joint Genome Institute (JGI) and the Environmental Molecular Sciences Laboratory (EMSL). Funding from the Canadian Natural Science and Engineering Research Council (NSERC) Discovery grants (DW and CG) and the Canada Research Chair Program (DW) is acknowledged.

Availability of data and materials
The metagenomic data generated in this study are available in the Integrated Microbial Genomes database at the Joint Genome Institute at https://img.jgi.doe.gov, GOLD Project ID: Gs0134626. Metagenome-assembled genome projects will be deposited at ENA under the accession number PRJEB57575. For the purpose of review, the MAG datasets (.fa, .faa, and .gff files) supporting the conclusions of this article are included as supporting documents.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
I confirm that I understand ***Microbiome*** is an open-access journal that levies an article processing charge per articles accepted for publication. By submitting my article I agree to pay this charge in full if my article is accepted for publication.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Biology, Concordia University, 7141 Sherbrooke St. West, Montreal, QC H4B 1R6, Canada. [2]Department of Chemistry, Sherbrooke University, 2500 Blvd de l'Université, Sherbrooke, QC J1K 2R1, Canada.

Received: 15 February 2022 Accepted: 8 November 2022
Published online: 24 December 2022

References

1. Dou S, et al. Are humic substances soil microbial residues or unique synthesized compounds? A perspective on their distinctiveness. Pedosphere. 2020;30:159–67.
2. Kida M, et al. Contribution of humic substances to dissolved organic matter optical properties and iron mobilization. Aquat Sci. 2018;80:1–11.
3. Park HJ, et al. Temporal changes in soil bacterial diversity and humic substances degradation in subarctic tundra soil. Microb Ecol. 2015;69:668–75.
4. Kisand V, Gebhardt S, Rullkötter J, Simon M. Significant bacterial transformation of riverine humic matter detected by pyrolysis GC-MS in serial chemostat experiments. Mar Chem. 2013;149:23–31.
5. Esham EC, Ye W, Moran MA. Identification and characterization of humic substances-degrading bacterial isolates from an estuarine environment. FEMS Microbiol Ecol. 2000;34:103–11.
6. Opsahl S, Benner R. Distribution and cycling of terrigenous dissolved organic matter in the ocean. Nature. 1997;386:480–2.
7. Kisand V, Rocker D, Simon M. Significant decomposition of riverine humic-rich DOC by marine but not estuarine bacteria assessed in sequential chemostat experiments. Aquat Microb Ecol. 2008;53:151–60.
8. Rocker D, et al. Differential decomposition of humic acids by marine and estuarine bacterial communities at varying salinities. Biogeochemistry. 2012;111:331–46.
9. Shen Y, Benner R, Robbins LL, Wynn JG. Sources, distributions, and dynamics of dissolved organic matter in the Canada and Makarov basins. Front Mar Sci. 2016;3:198.
10. Holmes RM, et al. Seasonal and annual fluxes of nutrients and organic matter from large rivers to the arctic ocean and surrounding seas. Estuaries Coasts. 2012;35:369–82.
11. Raymond PA, et al. Flux and age of dissolved organic carbon exported to the Arctic Ocean: a carbon isotopic study of the five largest arctic rivers. Global Biogeochem Cycles. 2007;21:1–9.
12. Macdonald RW, Kuzyk ZA, Johannessen SC. It is not just about the ice: a geochemical perspective on the changing Arctic Ocean. J. Environ Stud Sci. 2015;5:288–301.
13. Duarte CM, Lenton TM, Wadhams P, Wassmann P. Abrupt climate change in the Arctic. Nat Clim Chang. 2012;2:60–2.
14. Frey KE, McClelland JW. Impacts of permafrost degradation on arctic river biogeochemistry. Hydrol Process. 2009;23:169–82.
15. Mann PJ, et al. Pan-Arctic trends in terrestrial dissolved organic matter from optical measurements. Front Earth Sci. 2016;4:1–18.
16. Anderson LG, Macdonald RW. Observing the Arctic Ocean carbon cycle in a changing environment. Polar Res. 2015;34:26891.
17. Connolly CT, Cardenas MB, Burkart GA, Spencer RGM, McClelland JW. Groundwater as a major source of dissolved organic matter to Arctic coastal waters. Nat Commun. 2020;11:1–8.
18. Gonçalves-Araujo R, et al. Using fluorescent dissolved organic matter to trace and distinguish the origin of Arctic surface waters. Sci Rep. 2016;6:1–12.
19. Muscolo A, Sidari M, Nardi S. Humic substance: relationship between structure and activity. deeper information suggests univocal findings. J Geochemical Explor. 2013;129:57–63.
20. Polyakov V, Orlova K, Abakumov E. Soils of the Lena river delta, Yakutia, Russia: diversity, characteristics and humic acids molecular composition. Polarforschung. 2018;88:135–50.
21. Chupakova AA, Chupakov AV, Neverova NV, Shirokova LS, Pokrovsky OS. Photodegradation of river dissolved organic matter and trace metals in the largest European Arctic estuary. Sci Total Environ. 2018;622-623:1343–52.
22. Jung J, et al. Tracing riverine dissolved organic carbon and its transport to the halocline layer in the Chukchi Sea (western Arctic Ocean) using humic-like fluorescence fingerprinting. Sci Total Environ. 2021;772:145542.
23. Chen M, et al. Production of fluorescent dissolved organic matter in Arctic Ocean sediments. Sci Rep. 2016;6:39213.
24. Belicka LL, Harvey HR. The sequestration of terrestrial organic carbon in Arctic Ocean sediments: a comparison of methods and implications for regional carbon budgets. Geochim Cosmochim Acta. 2009;73:6231–48.

25. Gao Z, Guéguen C. Distribution of thiol, humic substances and colored dissolved organic matter during the 2015 Canadian Arctic GEOTRACES cruises. Mar Chem. 2018;203:1–9.
26. Gueguen C, DeFrancesco C. Long-term trends in dissolved organic matter composition and its relation to sea ice in the Canada Basin, Arctic Ocean (2007–2017). J. Geophys Res Ocean. 2021;126:e2020JC016578.
27. Sutton R, Sposito G. Molecular structure in soil humic substances: the new view. Environ Sci Technol. 2005;39:9009–15.
28. Gerke J. Concepts and misconceptions of humic substances as the stable part of soil organic matter: a review. Agronomy. 2018;8:76.
29. Esteves VI, Otero M, Duarte AC. Comparative characterization of humic substances from the open ocean, estuarine water and fresh water. Org Geochem. 2009;40:942–50.
30. Dittmar T, Kattner G. The biogeochemistry of the river and shelf ecosystem of the Arctic Ocean: a review. Mar Chem. 2003;83:103–20.
31. Hedges JI, et al. The molecularly-uncharacterized component of nonliving organic matter in natural environments. Org Geochem. 2000;31:945–58.
32. Schmidt MWI, et al. Persistence of soil organic matter as an ecosystem property. Nature. 2011;478:49–56.
33. Ekschmitt K, et al. Soil-carbon preservation through habitat constraints and biological limitations on decomposer activity. J Plant Nutr Soil Sci. 2008;171:27–35.
34. Kim D, Park HJ, Nam S, Kim SC, Lee H. Humic substances degradation by a microbial consortium enriched from subarctic tundra soil. Korean J Microbiol. 2019;55:367–76.
35. Hertkorn N, Claus H, Schmitt-Kopplin P, Perdue EM, Filip Z. Utilization and transformation of aquatic humic substances by autochthonous microorganisms. Environ Sci Technol. 2002;36:4334–45.
36. Brink DP, Ravi K, Lidén G, Gorwa-Grauslund MF. Mapping the diversity of microbial lignin catabolism: experiences from the eLignin database. Appl Microbiol Biotechnol. 2019;103:3979–4002.
37. Janusz G, et al. Lignin degradation: microorganisms, enzymes involved, genomes analysis and evolution. FEMS Microbiol Rev. 2017;41:941–62.
38. Herndl GJ, et al. Regulation of aquatic microbial processes: The 'microbial loop' of the sunlit surface waters and the dark ocean dissected. Aquat Microb Ecol. 2008;53:59–68.
39. Comeau AM, Vincent WF, Bernier L, Lovejoy C. Novel chytrid lineages dominate fungal sequences in diverse marine and freshwater habitats. Sci Rep. 2016;6:1–6.
40. Gladfelter AS, James TY, Amend AS. Marine fungi. Curr Biol. 2019;29:R191–5.
41. Bochdansky AB, Clouse MA, Herndl GJ. Eukaryotic microbes, principally fungi and labyrinthulomycetes, dominate biomass on bathypelagic marine snow. ISME J. 2017;11:362–73.
42. Ilicic D, Grossart HP. Basal parasitic fungi in marine food webs—a mystery yet to unravel. J Fungi. 2022;8:114.
43. Rocker D, Brinkhoff T, Grüner N, Dogs M, Simon M. Composition of humic acid- degrading estuarine and marine bacterial communities. FEMS Microbiol Ecol. 2012;80:45–63.
44. Kim D, Park HJ, Sul WJ, Park H. Transcriptome analysis of Pseudomonas sp. from subarctic tundra soil: pathway description and gene discovery for humic acids degradation. Folia Microbiol (Praha). 2018;63:315–23.
45. Colatriano D, et al. Genomic evidence for the degradation of terrestrial organic matter by pelagic Arctic Ocean Chloroflexi bacteria. Commun Biol. 2018;1:90.
46. Murphy KR, Stedmon CA, Waite TD, Ruiz GM. Distinguishing between terrestrial and autochthonous organic matter sources in marine environments using fluorescence spectroscopy. Mar Chem. 2008;108:40–58.
47. Guéguen C, Kowalczuk P. Colored dissolved organic matter in frontal zones. In: Chemical Oceanography of Coastal Zones; 2013. https://doi.org/10.1007/698.
48. Seung HS, Lee DD. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401:788–91.
49. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics. 2020;36:1925–7.
50. Amon RMW, et al. Dissolved organic matter sources in large Arctic rivers. Geochim Cosmochim Acta. 2012;94:217–37.
51. Goñi MA, Yunker MB, MacDonald RW, Eglinton TI. Distribution and sources of organic biomarkers in arctic sediments from the Mackenzie River and Beaufort shelf. Mar Chem. 2000;71:23–51.
52. Goñi MA, et al. Distribution and sources of organic matter in surface marine sediments across the North American Arctic margin. J Geophys Res Ocean. 2013;118:4017–35.
53. Sipler RE, et al. Microbial community response to terrestrially derived dissolved organic matter in the coastal Arctic. Front Microbiol. 2017;8:1–19.
54. Lund Paulsen M, et al. Biological transformation of Arctic dissolved organic matter in a NE Greenland fjord. Limnol Oceanogr. 2019;64:1014–33.
55. Davis J, Benner R. Seasonal trends in the abundance, composition and bioavailability of particulate and dissolved organic matter in the Chukchi/Beaufort Seas and western Canada Basin. Deep Res Part II Top Stud Oceanogr. 2005;52:3396–410.
56. Tremblay JÉ, et al. Vertical stability and the annual dynamics of nutrients and chlorophyll fluorescence in the coastal, southeast Beaufort Sea. J Geophys Res Ocean. 2008;113:1–14.
57. Lu P, et al. Isolation and characterization marine bacteria capable of degrading lignin- derived compounds. PLoS One. 2020;15:1–19.
58. Ramachandran A, McLatchie S, Walsh DA. A novel freshwater to marine evolutionary transition revealed within Methylophilaceae bacteria from the Arctic Ocean. MBio. 2021;12:e0130621.
59. Kraemer S, Ramachandran A, Colatriano D, Lovejoy C, Walsh DA. Diversity and biogeography of SAR11 bacteria from the Arctic Ocean. ISME J. 2020;14:79–90.
60. Colatriano D, Walsh DA. An aquatic microbial metaproteomics workflow: from cells to tryptic peptides suitable for tandem mass spectrometry-based analysis. J Vis Exp. 2015;103:e52827.
61. Murphy KR, Stedmon CA, Graeber D, Bro R. Fluorescence spectroscopy and multi-way techniques. PARAFAC Anal Methods. 2013;5:6557–66.
62. Bankevich A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.
63. Huntemann M, et al. The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v . 4). Stand Genomic Sci. 2016;1:1–5.
64. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25.
65. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26:139–40.
66. J, O., et al. The vegan package. Community Ecol Packag. 2007;10:631–7.
67. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics. 2010;11:367.
68. Jiang X, et al. Functional biogeography of ocean microbes revealed through non- negative matrix factorization. PLoS One. 2012;7:1–9.
69. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non- negativity-constrained least squares for microarray data analysis. Bioinformatics. 2007;23:1495–502.
70. Bugg TDH, Ahmad M, Hardiman EM, Rahmanpour R. Pathways for degradation of lignin in bacteria and fungi. Nat Prod Rep. 2011;28:1883–96.
71. Kamimura N, et al. Bacterial catabolism of lignin-derived aromatics: new findings in a recent decade: update on bacterial lignin catabolism. Environ Microbiol Rep. 2017;9:679–705.
72. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.
73. Bağcı C, Patz S, Huson DH. DIAMOND+MEGAN: fast and easy taxonomic and functional analysis of short and long microbiome sequences. Curr Protoc. 2021;1:1–29.
74. Kang DD, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7:e7359.
75. Imelfort M, Skennerton CT, Parks DH, Tyson GW, Hugenholtz P. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
76. Parks DH, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36:996.
77. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:1–8.

78. Kraemer S, Ramachandran A, Colatriano D, Lovejoy C, Walsh DA. Diversity and biogeography of SAR11 bacteria from the Arctic Ocean. ISME J. 2019. https://doi.org/10.1038/s41396-019-0499-4.
79. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–5.
80. Kalvari I, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res. 2021;49:D192–200.
81. Aramaki T, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics. 2020;36:2251–2.

Publisher's Note

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

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

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

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions