A consensus protocol for the recovery of mercury methylation genes from metagenomes

Eric Capo[1,2], Benjamin D. Peterson[3], Minjae Kim[4], Daniel S. Jones[5,6], Silvia G. Acinas[1], Marc Amyot[7], Stefan Bertilsson[2], Erik Björn[8], Moritz Buck[2], Claudia Cosio[9], Dwayne Elias[10], Cynthia Gilmour[11], Maria Soledad Goñi Urriza[12], Baohua Gu[10], Heyu Lin[13], Yu-Rong Liu[14], Katherine McMahon[3], John W. Moreau[15], Jarone Pinhassi[16], Mircea Podar[10], Fernando Puente-Sánchez[2], Pablo Sánchez[1], Veronika Storck[7], Yuya Tada[17], Adrien Vigneron[12], David Walsh[18], Marine Vandewalle-Capo[2], Andrea G. Bravo[1]*, Caitlin Gionfriddo[11]*

Corresponding author: Eric Capo eric.capo@hotmail.fr

*joint last authors

[1]Department of Marine Biology and Oceanography, Institute of Marine Sciences, CSIC, Barcelona, 08003, Spain
[2]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, 75007, Sweden
[3]Department of Bacteriology, University of Wisconsin at Madison, Madison, WI 53706, United States
[4]Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523, USA
[5]Department of Earth and Environmental Science, New Mexico Institute of Mining and Technology, Socorro, NM 87801, USA.
[6]National Cave and Karst Research Institute, Carlsbad, NM 88220, USA
[7]Department of Biological Sciences, University of Montréal, Montréal, QC, H3C 5J9, Canada
[8]Department of Chemistry, Umeå University, Umeå, 90736, Sweden
[9]University of Reims Champagne-Ardenne, UMR-I 02 SEBIO, Reims, 51100, France.
[10]Oak Ridge National Lab, Oak Ridge, TN 37830, USA
[11]Smithsonian Environmental Research Center, Edgewater, MD 21037, USA
[12]University of Pau et des Pays de l'Adour, E2S UPPA, CNRS, IPREM, Pau, 64000, France
[13]School of Geography, Earth and Atmospheric Sciences, The University of Melbourne, Parkville, VIC 3010, Australia
[14]College of Resources and Environment, Huazhong Agricultural University, Wuhan, 430070, China
[15]School of Geographical and Earth Sciences, University of Glasgow, Glasgow, G12 8RZ, UK
[16]Centre for Ecology and Evolution in Microbial Model Systems - EEMiS, Linnaeus University, Kalmar, 39231, Sweden
[17]National Institute for Minamata Disease, Department of Environment and Public Health, Kumamoto, 867-0008, Japan
[18]Department of Biology, Concordia University, Montreal, Quebec H4BIR6, Canada

Abstract

Mercury methylation genes (*hgcAB)* mediate the formation of the toxic methylmercury and have been identified from diverse environments, including freshwater and marine ecosystems, Arctic permafrost, forest and paddy soils, coal-ash amended sediments, chlor-alkali plants discharges and geothermal springs. Here we present the first attempt at a standardized protocol for the detection, identification and quantification of *hgc* genes from metagenomes. Our Hg-MATE (Hg-cycling Microorganisms in Aquatic and Terrestrial Ecosystems) database, a catalogue of *hgc* genes, provides the most accurate information to date on the taxonomic identity and functional/metabolic attributes of microorganisms responsible for Hg methylation in the environment. Furthermore, we introduce "marky-coco", a ready-to-use bioinformatic pipeline based on *de novo* single-metagenome assembly, for easy and accurate characterization of *hgc* genes from environmental samples. We compared the recovery of *hgc* genes from environmental metagenomes using the marky-coco pipeline with an approach based on co-assembly of multiple metagenomes. Our data show similar efficiency in both approaches for most environments except those with high diversity (i.e., paddy soils) for which a co-assembly approach was preferred. Finally, we discuss the definition of true *hgc* genes and methods to normalize *hgc* gene counts from metagenomes.

Keywords: mercury, *hgcAB* genes, Hg methylation, metagenomics, bioinformatics, Hg-MATE, marky-coco

Running title: Recovering *hgcAB* genes from metagenomes

Introduction

Environmental mercury methylation is primarily a biotic process carried out by microorganisms that transform inorganic mercury (Hg) into the more toxic and bioaccumulative monomethylmercury (MeHg). The capacity to perform Hg methylation was historically associated with certain sulfate-reducing bacteria, iron-reducing bacteria and methanogenic archaea (Compeau and Bartha, 1985; Fleming et al., 2006; Kerin et al., 2006; Hamelin et al., 2011). Field observations revealed links between Hg methylation and sulfate-reduction, iron-reduction and methanogenesis in organic matter-rich anaerobic environments (Bravo and Cosio, 2020 for review), as well as subsequent studies that tested cultured representatives of these clades for Hg-methylation capability (Fleming et al. 2006; Gilmour et al., 2011; 2013; 2018). The discovery of the *hgc* genes (Parks et al., 2013) has facilitated the detection of novel putative Hg methylating bacteria and archaea through cultivation-independent molecular methods (Podar et al., 2015; Gionfriddo et al., 2016). Recent works analyzing publicly available genomes and environmental metagenome-assembled genomes (MAGs) identified *hgc*-containing *(hgc*$^+$) microorganisms from microbial lineages not formerly associated with Hg-methylation, such as members of the PVC superphylum (Jones et al., 2019; Gionfriddo et al., 2019; Peterson et al., 2020; McDaniel et al., 2020; Lin et al., 2021). Identifying *hgc* genes in microbial genomes from meta-omic datasets greatly expanded our view of the phylogenetic diversity of putative Hg methylators (Fig 1), but we still do not fully understand which microorganisms are the main drivers of Hg methylation in diverse environments, particularly outside of anoxic sediments.

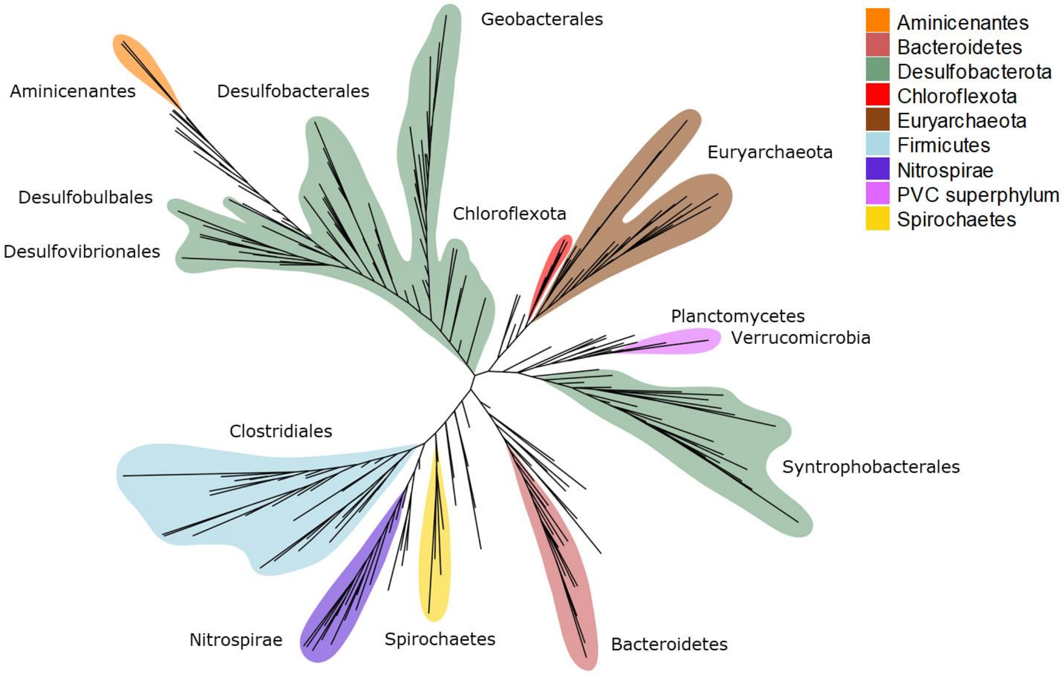


Figure 1. Simplified unrooted phylogenetic tree of *hgcA* sequences from the Hg-MATE database. Taxonomy is based on NCBI classification with the exception of Deltaproteobacteria (Desulfobacterota with GTDB classification) and Chloroflexi (Chloroflexota with GTDB classification). For visualization ease, microbial groups were collapsed by the dominant monophyletic group. Microbial groups with the highest diversity of *hgc*$^+$ microorganisms are denoted by colors.

Significant knowledge gaps in the identification of microorganisms capable of Hg methylation remain, largely because of the absence of *hgc*$^+$ cultured representatives from novel clades (i.e., outside the Desulfobacterota, Firmicutes, Methanomicrobia) with experimentally validated Hg-methylating capability (Gilmour et al., 2018). One reason for this is the difficulty in selecting for *hgc*$^+$ microorganisms during cultivation, and another is the lack of a successful methodology for isolating all relevant microbes in controlled laboratory conditions. Microbes that have yet to be cultivated, and for which successful laboratory growth parameters need to be identified, are often referred to as the “unculturable” (Hug et al., 2016; Steen et al., 2019). High-throughput meta-omic and targeted amplicon sequencing studies have become the main methods for identifying putative Hg methylating microorganisms of this unculturable fraction (Bravo et al., 2018; Gionfriddo et al., 2020; Xu et al., 2021). While directly testing for Hg methylation capacity may not be a viable strategy, pairing these sequencing methods with biogeochemical measurements, Hg methylation assays, and other manipulation studies can connect a Hg-methylating microbiome to MeHg production and metabolic activity and help to elucidate the potential contribution of these novel clades to Hg methylation (Kronberg et al., 2016; Bouchet et al., 2018; Schaefer et al., 2020; Roth et al., 2021).

The detection of hgc$^+$ MAGs provides the most precise information about the taxonomic and metabolic characteristics of putative Hg methylators (Jones et al., 2019; Peterson et al., 2020; Lin et al., 2021; Vigneron et al., 2021). However, the microbial diversity in some environments is too high and/or Hg methylators are too rare to identify them effectively (Podar et al., 2015; Christensen et al., 2019). In these cases, read-based metagenomic analyses and *hgc* metabarcoding are easier and more economical. Accurately identifying Hg-methylating clades (and metabolic guilds) from *hgc* sequences alone therefore requires a universally used and updated *hgcAB* database, coupled to consistent and robust bioinformatic practices, in order to identify precisely the target genes in complex meta-omic datasets.

In this work, we introduce Hg-MATE (Hg-cycling Microorganisms in Aquatic and Terrestrial Ecosystems) database version 1 (https://doi.org/10.25573/serc.13105370.v1), an up-to-date *hgcAB* catalog compiled from isolated, single-cell and metagenome-reconstructed genomes. Additionally, we present marky-coco (https://github.com/ericcapo/marky-coco), a ready-to-use bioinformatic pipeline to detect, identify and count *hgc* genes from metagenomes (**Fig 2**). We apply this pipeline to metagenomes collected from paddy soils, brackish and lake waters, as well as sediments from reservoirs and lakes in which *hgc* genes have been previously detected (Liu et al., 2018; Jones et al., 2019; Capo et al., 2020; Millera Ferriz et al., 2021). Further, we specifically compared the reliability of (i) applying the marky-coco pipeline based on *de novo* single assembly approach from single metagenomes with (ii) co-assembling of multiple metagenomes (co-assembly) prior to mapping and identification. Finally, we discuss appropriate definitions and cutoff criteria for *hgc* genes and also best practices to normalize data for an accurate count of *hgc* genes in metagenomes from environmental samples.

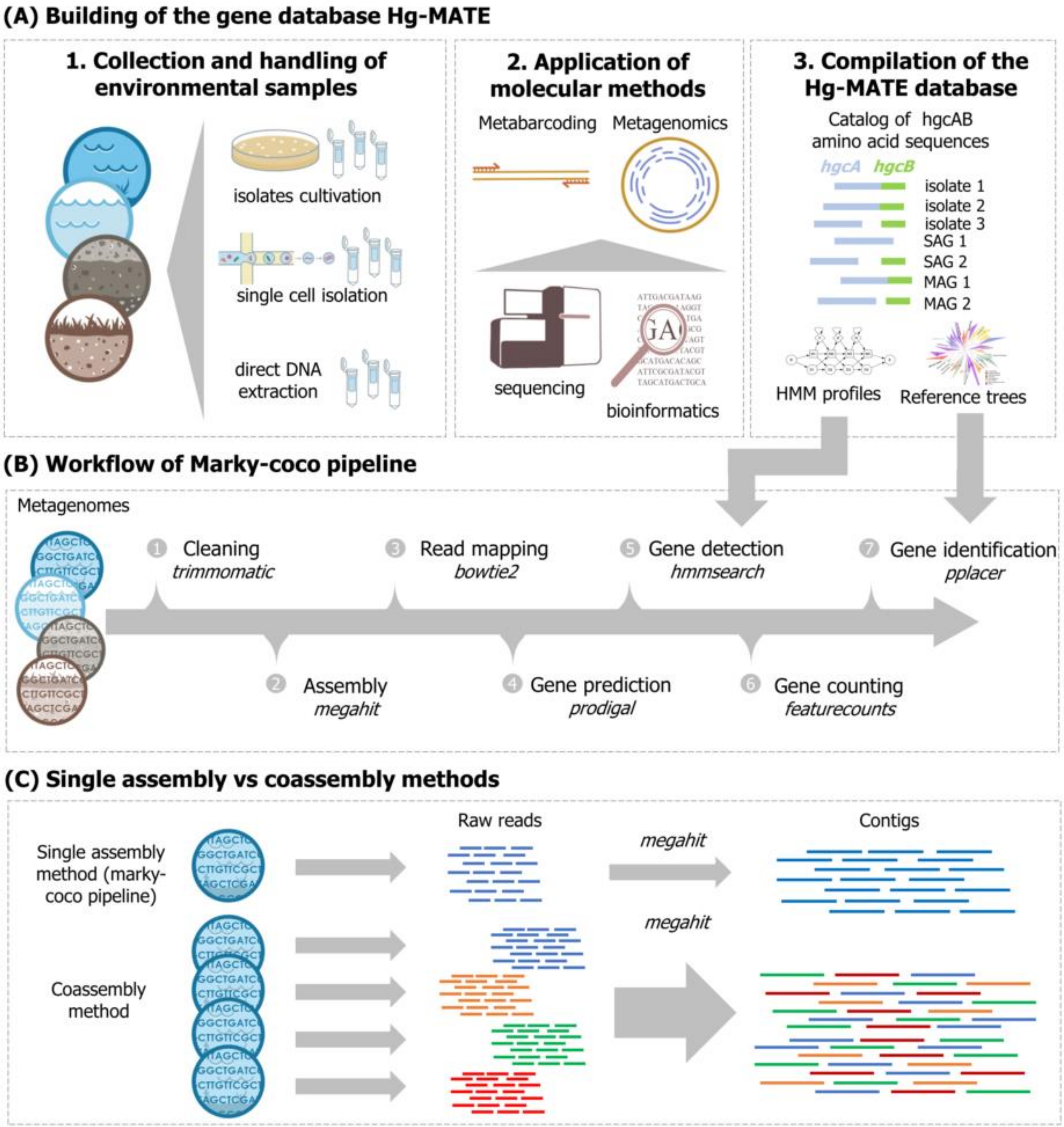


Figure 2. (A) Workflow illustrating how the *hgcAB* gene catalogue Hg-MATE database was built, (B) Simplified workflow of the marky-coco pipeline (C) Illustration of the two assembly approaches compared in this work: single assembly vs co-assembly.

2. Material and Methods

2.1 Description of the Hg-MATE database v1

The Hg-MATE database v1 was released on 14 January 2021 (https://doi.org/10.25573/serc.13105370.v1), and contains an extensive *hgcAB* dataset from a wide range of microorganisms and environments. The catalog contains 1053 unique HgcA/B amino acid sequences (Table 1). We categorized the HgcAB amino acid sequences into four types depending on whether they were encoded in (i) pure culture/environmental microbial isolates (ISO) (ii) single-cell genome sequences (CEL) (iii) metagenome-assembled genomes (MAGs) (iv) or an environmental meta-omic contig (CON). Amino acid sequences of HgcA, HgcB, and concatenated HgcA and HgcB were included in the database. If *hgcB* was not co-localized with *hgcA* in the genome and/or could not be identified, then 'na' was listed in the 'HgcB' sequence column. Both genes need to be present and encode functional proteins for a microbe to methylate Hg (see Parks et al., 2013; Smith et al., 2015). One reason *hgcB* may not be identified in some genomes carrying *hgcA* is because HgcB is highly homologous to other 4Fe-4S ferredoxins. Therefore, *hgcB* can be difficult to differentiate from other ferredoxin-encoding genes if not co-localized with *hgcA* on a contiguous sequence. In addition, *hgcB* may be missing from 'MAGs', 'CEL' and 'CON' sequences due to incomplete coverage of the genome or incomplete contig assembly, or failure to bin the contig carrying *hgcB*. Some *hgc* genes are predicted to encode a 'fused HgcAB protein' which has been previously described (Podar et al., 2015), and is characterized by one gene that encodes for a 4Fe-4S ferredoxin-like protein with shared homology to HgcA and HgcB. This 'fused HgcAB' protein contains the corrinoid iron-sulfur and transmembrane domains characteristic of HgcA as well as the 4Fe-4S ferredoxin motif of HgcB (e.g., Uniprot Q8U2U9, NCBI Refseq: WP_011011854.1, *Pyrococcus furiosus* DSM 3638). These sequences are provided in the 'HgcA' column, and labeled 'fused HgcAB' in the HgcB column. These 'fused HgcAB' sequences should be treated with caution because, while they share significant sequence homology to HgcA and HgcB from confirmed Hg methylators, to date all organisms with a 'fused HgcAB' that have been tested do not seem to produce MeHg in culture (Podar et al., 2015; Gilmour et al., 2018).

Table 1. Summary of HgcAB sequence types in version 1 of the Hg-MATE database.

Genome type	Total HgcA(B) sequences	Encode both HgcA and HgcB	Encode fused HgcAB	Only HgcA (or HgcB) present
ISO	204	173	10	21
CEL	29	4	18	7
MAG	787	696	17	74
CON	33	9	0	21(3)

The resources within the Hg-MATE database v1 include a catalog with the amino acid sequences and metadata of all microorganisms. Only sequences with genomic identifying information (i.e., 'ISO', 'CEL', 'MAG') were used to compile further resources. Resources include: (i) FASTA files containing Hgc amino acid sequences; (ii) Multiple Sequence Alignments (MSA) in FASTA format of Hgc amino acid sequences built with MUSCLE implemented in MEGAX (Kumar et al., 2018) with the cluster method UPGMA; and (iii)

Hidden Markov models (HMM) of aligned Hgc amino acid sequences built from MSAs using the *hmmbuild* function from the hmmer software (3.2.1 version, Finn et al., 2011). Additionally, resources include reference packages that can be used to identify and classify: (1) the corrinoid-binding domain of HgcA which corresponds to residues ~37-156 of the HgcA sequence from *Pseudodesulfovibrio mercurii* ND132 and includes the characteristic cap helix domain (2) full HgcA sequence and (3) concatenated HgcA and HgcB. Each reference package contains sequence alignments, an HMM model, a phylogenetic tree, and NCBI taxonomy. Reference packages were constructed using the program Taxtastic (https://github.com/fhcrc/taxtastic) for HgcA(B) amino acid sequences from ISO, CEL & MAG. Phylogenetic trees were built from MSA files by RAxML using the GAMMA model of rate heterogeneity and LG amino acid substitution matrix (Le and Gascuel, 2008). Trees were rooted by HgcA paralog sequences, carbon monoxide dehydrogenases (PF03599) from non-HgcA coding microorganisms Candidatus Omnitrophica bacterium CG1_02_41_171 and *Thermosulfurimonas dismutan*s. These organisms were chosen because of their distant phylogenetic relationship to hgcA$^+$ microorganisms. Confidence values on branches were calculated from 100 bootstraps. Using the HgcA reference tree, a simplified tree of 'ISO', 'CEL', 'MAG' *hgcA* genes was built using iTOL (Letunic and Bork, 2019) and clades were collapsed by the dominant monophyletic group, when possible, for visualization ease.

2.2 Data collection

A total of 29 metagenomes from recent studies studying *hgc* genes in environments with known active Hg methylation were used for the bioinformatic analyses performed in this work (Table 1, Datasheet 1A). Metagenomes from brackish waters (BARM8s) were collected in 2014 in the Gotland Deep basin of the Central Baltic Sea. Out of 81 available metagenomes (Alneberg et al., 2018; BioProject ID PRJEB22997), 8 metagenomes where *hgc* genes have been detected (Capo et al., 2020) were used in the present analysis. Water depths of these metagenomes ranged from 76 to 200 m with oxygen concentrations either low (hypoxic zone) or undetectable (anoxic zone), salinity ranging between 9.2-12.1 psu and MeHg concentrations measuring up to 1640 fM (Soerensen et al., 2018). Lake sediments and water metagenomes (MANGA6s) were obtained in 2013-2014 from the sulfate-impacted Manganika lake in Northern Minnesota (Jones et al., 2019, BioProject ID PRJNA488162). This hypereutrophic lake is characterized by dissolved oxygen approaching 16 mg/L (nearly 200% saturation) near the surface, pH exceeding 8.7 and MeHg accumulating over 3 ng/L in bottom waters. Dissolved oxygen and pH decreased with depth, and anoxic conditions were encountered below 4 m. Sulfide concentrations up to 2 mM were observed in bottom waters and sediments. Water samples were collected at these anoxic depths. Five metagenomes (RES5S) were obtained from reservoir sediments from the St. Maurice River near Wemotaci, Canada in 2017 and 2018 (Millera-Ferriz et al., 2021, GOLD-JGI Ga0393614 Ga0393582, Ga0393617, Ga0393586, Ga0393589). The studied river section has been affected by the construction of two run-of-river power plant dams and its watershed has been disturbed by a forest fire, logging, and the construction of wetlands. MeHg concentrations in samples varied from <0.02 to 19 ng/g. Metagenomes from paddy and upland soils (PADDY10s) were collected from two historical Hg mining sites, Fenghuang (FH) and Wanshan (WS), in Southwest China in August 2016 (Liu et al., 2018, BioProject ID PRJNA450451). The pH of paddy soils ranged from 6 to 7.5. Historical discharge from Hg mining operations and ongoing atmospheric deposition contribute to high concentrations of MeHg in the soils around these areas with values up to 7.9 ng g^{-1} in the collected samples.

Table 2. Metagenomes collected from previously published papers investigating the presence of Hg methylators in the environment.

Systems	#metagenomes	dataset id	References
Brackish waters	8	BARM8s	Capo et al., 2020
Lake sediments and water	6	MANGA6s	Jones et al., 2019
River/reservoir sediments	5	RES5s	Millera Ferriz et al., 2021
Paddy soils/upland soils	10	PADDY10s	Liu et al., 2018

2.3 Bioinformatics

The detection, taxonomic identification and counting of *hgc* genes was done with the marky-coco snakemake-implemented pipeline (https://github.com/ericcapo/marky-coco). A brief overview of this workflow is as follows: the metagenomes were trimmed and cleaned using fastp (Chen et al., 2018) with the following parameters: quality threshold of 30 (-q 30), length threshold of 25 (-l 25), and with trimming of adapters and polyG tails enabled (--detect_adapter_for_pe --trim_poly_g --trim_poly_x). A *de novo* single assembly approach, in which each metagenome was assembled individually, was applied using the assembler megahit 1.1.2 (Li et al., 2016) with default settings. The annotation of the contigs for prokaryotic protein-coding gene prediction was done with the software Prodigal 2.6.3 (Hyatt et al., 2010). The DNA reads were mapped against the contigs with bowtie2 (Langdmead and Salzberg, 2012), and the resulting .sam files were converted to .bam files using samtools 1.9 (Li et al., 2009). The .bam files and the prodigal output .gff file were used to estimate read counts by using featureCounts (Liao et al., 2014). In order to detect *hgc* homologs, HMM profiles derived from the Hg-MATE database v1 were applied to the amino acid FASTA file generated with Prodigal from each assembly with the function *hmmsearch* from HMMER 3.2.1 (Finn et al., 2011). The reference package 'hgcA' from Hg-MATE.db was used for phylogenetic analysis of the HgcA amino acid sequences. Briefly, the predicted amino acid sequences from gene identified as putative *hgcA* gene were (i) compiled in a FASTA file, (ii) aligned to the Stockholm formatted HgcA alignment from the reference package with the function *hmmalign* from HMMER 3.2.1 (iii) placed onto the HgcA reference tree and classified using the functions *pplacer*, *rppr* and *guppy_classify* from the program pplacer (Matsen et al., 2010). For more details, see the README.txt of the Hg-MATE database v1 (https://doi.org/10.25573/serc.13105370.v1). Additionally, to compare the efficiency of the marky-coco pipeline to detect *hgc* genes from metagenomes with a co-assembly approach (multiple metagenomes used for assembly), we performed co-assemblies on metagenomes within each environmental system (BARM8s, MANGA6s, RES5s, PADDY10s, Table 2). Post-assembly, all other steps of the analysis procedure were performed similarly to the marky-coco pipeline. Detection of *dsrA* genes were detected in metagenomes with the function *hmmsearch* and HMM profile from TIGRFAM (Selengut et al., 2007). The amount of sequencing required to cover the total diversity and the estimated diversity of each metagenome were evaluated using the Nonpareil method (Rodriguez-R and Konstantinidis, 2014).

2.4 Stringency cut-offs for the definition of true *hgc* genes

Based on knowledge from confirmed isolated Hg methylators, we propose several stringency cutoffs that could be used to distinguish between an *hgcA* gene homolog and an *hgcA*-like gene that encodes for a protein of unknown Hg methylation capability. (i) High stringency cutoff: amino acid sequence includes one of the cap-helix motifs with the conserved cysteine (Cys93 in *P. mercurii* ND132), NVWCAAGK, NVWCASGK, NVWCAGGK, NIWCAAGK, NIWCAGGK or NVWCSAGK. This cutoff is based on previous findings that showed isolated microorganisms carrying HgcA proteins with the cap helix domain are capable of Hg methylation (Parks et al., 2013; Smith et al., 2015; Gilmour et al., 2018; Cooper et al., 2020). Within the high stringency cutoff, there is a possible need to distinguish between the amino acid sequences from fused HgcAB-like proteins and those from true HgcA proteins, since isolates that encode fused HgcAB-like genes do not have the capacity to methylate Hg in culture (Podar et al., 2015; Gilmour et al., 2018). The fused HgcAB include the cap-helix and ferredoxin motifs of HgcA and HgcB. (ii) Moderate stringency cutoff: in addition to amino acid sequences that include the motifs described above, any sequence with a bitscore value obtained from the HMM analysis greater than or equal to 100 is included (iii) Low stringency cutoff: in addition to amino acid sequences that include the motifs described above, any sequence with a bitscore value greater than or equal to 60 is included. For *hgcB* gene homologs, we propose two cutoffs that could be used for their description as *hgcB* genes. (i) High stringency cutoff: their amino acid sequences include one of the following motifs featuring the conserved Cys (Cys73 in *P. mercurii* ND132, Cooper et al., 2020), C(M/I)ECGA motifs and that the genes are found on the same contig as an *hgcA* genes. (ii) Moderate stringency cutoff: amino acid sequences include the C(M/I)ECGA motif, but the gene are not co-located on a contig with an *hgcA* gene.

2.5 Estimation of *hgcA* abundance in metagenomes

Coverage values of *hgcA* genes were calculated, for each gene and each sample, as the number of reads mapping to the gene divided by the length of the gene (read/bp). We compared the reliability of four procedures for normalizing read counts of *hgcA* genes. Normalization metrics were (i) the total number of mapped reads (ii) the summed coverage values of *rpoB* genes, (iii) the median coverage values of 257 marker genes (GTDB-Tk r89 release, Chaumeil et al., 2019), or (iv) the genome equivalents values calculated using the software MicrobeCensus (Nayfach and Pollard, 2015) which normalizes the relative abundance by the metagenomic dataset size and the community average genome size of the microbial community. The coverage of each marker gene was calculated as the sum of the coverages of all the ORFs assigned to that gene (Datasheet 1A). The *rpoB* and the 256 other marker genes were detected using the function *hmmsearch* from hmmer software (v3.2.1, Finn et al., 2011) and applying the trusted cut-off provided in HMM files (GTDB-Tk r89 release, Chaumeil et al., 2019).

2.6 Data analysis

A non-metric multidimensional scaling analysis (nMDS) was performed applying the function *metaMDS* from the R package vegan (Oksanen et al., 2015) to the table of *hgcA* gene coverage values, clustered at the lowest level of NCBI taxonomic identification (txid), obtained with single assembly and co-assembly approaches (Datasheet 1B). A PROTEST permutation procedure analysis (1000 permutations) was performed using the function Procrustes to evaluate the level of concordance of the outputs between both approaches. The functions *rcorr* from the R package Hmisc (Harrell and Harrell, 2019), *corrplot* from the R package corrplot (Taiyun et al., 2017) and *plot3D* from the R package rgkl (Adler et al., 2019) were used to investigate correlations between normalization methods.

3. Results

3.1 Dataset outputs

A total of 29 single assemblies (one for each metagenome) and 4 co-assemblies (reads from each of the BARM8s, MANGA6s, RES5S, and PADDY10s metagenome sets assembled together) were used to compare the efficiency of a single assembly using the marky-coco pipeline and a co-assembly approach to detect, identify and count *hgc* genes from metagenomes (**Fig. 2**). The number of mapped reads of the analyzed metagenomes ranged between 10.2-110.9 M reads (average, 29.4 ± 19.6) with single assembly and 16.6-120.7 M reads (average, 36.0 ± 19.9) with co-assembly, with the percentage of mapped reads ranging between 16-76 % and 24-89 %, respectively (Datasheet 1A). Nonpareil diversity index values (N_d) of metagenomes were between 18.7 and 23.7 with the highest found in paddy soil metagenomes (Fig. S1, Table 3). Nonpareil curves showed that paddy soil samples from this study required the highest sequencing effort for nearly complete coverage followed by reservoir sediments, and then lake sediment and lake waters and brackish waters (Fig S1). Estimated coverage of paddy soils metagenomes was relatively low (average, 0.30-0.37) compared to other metagenomes (0.49-0.83) showing that only a portion of the diversity of these environmental samples was recovered despite the relatively high sequencing depth (88.6 ± 5.6 M reads) (Table 3). Seven metagenomes (S02, S03, S19, S22, S26, S28, S29) that were used in coassemblies but with low *hgcA* coverage values (i.e., <0.40 obtained from co-assemblies) were not used for further comparison analysis. The remaining 22 metagenomes, labeled MG01 - MG22, had *hgcA* unnormalized coverage values between 0.44 and 3.06 (1.22 ± 0.79) (Datasheet 1A). Only *hgcA* genes (and not *hgcB*) from these metagenomes were used for comparison of the two assembly approaches as *hgcAB* gene pairs were not 100 % similar between the two approaches (Datasheet 1B). Additionally, *hgcAB*-like homologs that are predicted to encode for fused HgcAB proteins were excluded from further analysis.

3.2 Distribution of *hgcA* genes with different stringency cutoffs

By definition, all *hgcA* genes detected with the high stringency cutoff are predicted to encode proteins that include the conserved amino acid motifs characteristic of functional HgcA proteins, while this is not the case for those additionally detected when lowering the stringency cutoffs (i.e., moderate or low). We therefore considered that gene homologs to *hgcA* found with bitscore values below 100 and without conserved motifs cannot with confidence be defined as true *hgcA* genes. Nevertheless, we wanted here to highlight how "false" *hgcA* genes i.e., detected without the conserved amino acid motifs characteristic of functional HgcA proteins, were taxonomically assigned using the *pplacer* approach applied to the Hg-MATE *hgcA* reference tree. The *hgcA* genes detected with a high stringency cutoff and those additionally detected with moderate stringency cutoffs were predominantly identified as Desulfobacterota, Chloroflexota and Euryarchaeota (Fig S2). In contrast, the *hgcA* genes additionally detected with low stringency cutoff were primarily identified as members of the PVC superphylum but were unclassified at lower taxonomic levels. For further comparison, we used information only from *hgcA* genes detected with the high stringency.

3.3 Comparison between co-assembly vs single assembly approaches

For all metagenomes, 1.50-7.25 times more *hgcA* genes were detected in co-assemblies (19-147 genes) compared to linked single assemblies (4-69 genes) (Table 3). We investigated the differences in *hgcA* gene lengths, discriminating between genes (i) found at the extremity of contigs (potentially truncated) and (ii) between other genes in contigs therefore expected to be complete. A higher number of 'complete' *hgcA* gene sequences were detected with the co-assembly (1-17, average 6.8 ± 4.4 genes) compared to the single assembly (0-6, average 2.0 ± 2.7 genes), e.g., for metagenomes from brackish and lake waters (Datasheet 1A). No complete genes were identified in the single assemblies that were not also identified in the co-assembly. Violin plots illustrated that, overall, a higher number of 'complete' *hgcA* sequences (> 950 bp) were found with the co-assembly versus the single assembly (Fig. S3).

Table 3. For each metagenome, Non-pareil diversity index values, estimated average coverage, number of mapped reads, number of *hgcA* genes and *hgcA* coverage values (reads/bp) for co-assembly ´c´ and single assembly ´s´ approaches. See Datasheet 1A for extended description of the dataset.

Environments	**Metagenomes id**	**Non pareil diversity index (N_d)**	**Estimated average coverage**	**Number of mapped reads (millions reads)**		**Number of *hgcA* genes**		***hgcA* coverage values**	
				c	s	c	s	c	s
brackish water	MG01	19.51	0.83	120.7	110.9	38	14	2.04	1.85
	MG02	21.12	0.55	33.9	25.5	40	16	1.05	0.91
	MG03	19.49	0.70	32.0	25.8	29	7	1.01	0.75
	MG04	20.52	0.63	35.1	26.9	34	10	0.84	0.75
	MG05	18.69	0.76	35.6	30.5	23	5	0.52	0.46
	MG06	20.73	0.48	16.6	10.2	29	4	0.58	0.35
reservoir sediment	MG07	22.46	0.59	28.6	21.8	147	69	3.06	2.36
	MG08	21.99	0.64	33.6	29.4	103	53	2.19	1.98
	MG09	21.82	0.68	47.2	43.5	74	35	1.98	1.78
	MG10	22.10	0.63	36.1	32.7	102	68	2.32	3.00
	MG11	22.15	0.63	36.7	29.8	122	62	2.69	2.43
lake sediment	MG12	20.55	0.62	22.7	26.7	23	10	0.83	0.78
	MG13	20.75	0.57	27.2	22.5	26	9	0.41	0.32
lake water	MG14	21.57	0.49	29.6	24.8	31	13	1.19	1.05

paddy soils	MG15	20.24	0.66	38.5	34.6	31	8	0.62	0.50
	MG16	19.51	0.67	30.5	29.2	19	10	0.47	0.50
	MG17	23.48	0.34	31.1	20.4	77	21	0.69	0.45
	MG18	23.31	0.33	30.8	18.5	60	13	0.59	0.33
	MG19	23.67	0.27	27.1	14.3	85	20	0.76	0.43
	MG20	23.14	0.37	37.5	28.8	61	15	0.58	0.32
	MG21	23.49	0.30	30.5	18.7	57	25	0.60	0.51
	MG22	23.64	0.30	30.6	20.5	84	33	1.12	0.89

In a comparison of HgcA amino acid sequences recovered from the two assembly approaches, no HgcA sequence from the single assembly had 100% sequence identity to sequences in the co-assembly (Datasheet 1B). The highest sequence similarity of HgcA sequences from different assemblies of the same dataset was 99%. To compare, we investigated differences between assemblies for detecting *dsrA* gene, which encodes for dissimilatory sulfite reductase subunit A, an essential enzyme in sulfate reduction and expected to be present in these datasets. Identical amino acid sequences of DsrA-encoding genes were found when comparing single assemblies to the related co-assembly with numbers ranging from 1 to 33 depending on metagenomes (Datasheet 1D). Comparatively, *dsrA* genes were 3-34x more abundant (in coverage) than *hgcA* genes. This higher abundance helps explain why more identical *dsrA* were found between co-assembly and single assembly approaches than for *hgcA* genes.

Distribution plots showed unnormalized coverage values of *hgcA* clustered by environment types (Fig. 3A) or for each metagenome (Fig. S4). Importantly, unnormalized values were used here to compare single assembly vs coassembly results for each metagenome but not to compare difference between environments for which normalization would be required (Fig S4). Overall, higher *hgcA* coverage values were observed with the co-assembly for all types of environments (Fig 3A) and for each metagenome with the exception of reservoir sediment MG10 (Fig. S4, Table 3). The application of normalization methods (as described in the section below) revealed contrasting patterns in *hgcA* relative abundance, with higher values observed for single assembly methods when applying, for instance, a normalization method based on *rpoB* coverage values (Fig. S4). For each metagenome, the nMDS analysis showed a high level of similarity in taxonomy-based *hgcA* inventories obtained from single assembly vs co-assembly (Fig. 3B). This was confirmed by a procrustean analysis that showed significant levels of concordance for the *hgcA* inventories obtained between both approaches ($p \leq 0.001$). Looking at each dataset independently, reservoir sediments and brackish waters showed significant levels of concordances ($p \leq 0.008$, $p \leq 0.002$) while lake waters and paddy soils had non-significant levels of concordances ($p \leq 0.17$, $p \leq 0.30$; no statistics possible with only two metagenomes for lake sediments).

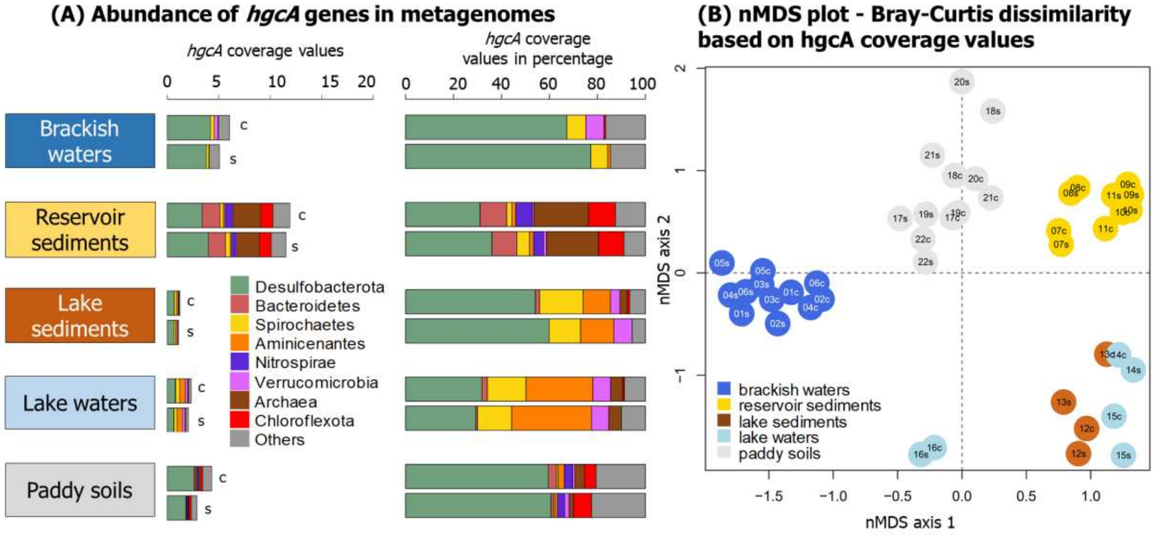


Figure 3 (A) Distribution of *hgcA* genes in the metagenomes obtained from five types of environments with the co-assembly 'c' and the single assembly 's' methods. For these barplots, unnormalized hgcA coverage values were used. (B) Dissimilarities in the structure of *hgcA* inventories obtained with the co-assembly 'c' and the single assembly 's' approaches. nMDS stress values = 0.1909. The id of each metagenome is denoted as follows: numbers corresponding to the metagenome id (e.g., MG01 is 01), ´c´ or ´s´ stands for analysis with the co-assembly or the single assembly.

3.4 Comparison between normalization methods

In order to compare normalization methods to estimate the abundance of *hgcA* genes, we calculated the (i) total number mapped prokaryotic reads, (ii) *rpoB* genes coverage values, (iii) median coverage value of 257 marker genes and (iv) genome equivalents values (Microbe Census) (Fig 4, Datasheet 1E). Overall, significant correlations were observed between the total number of reads, *rpoB* coverage values, and the median coverage values of 257 marker genes (Fig. 4A), while no significant correlations were observed between these metrics and genome equivalent values. The 3D plot shows the relationships between the total number of reads, the median coverage values of 257 marker genes and genome equivalent values (Fig. 4B).

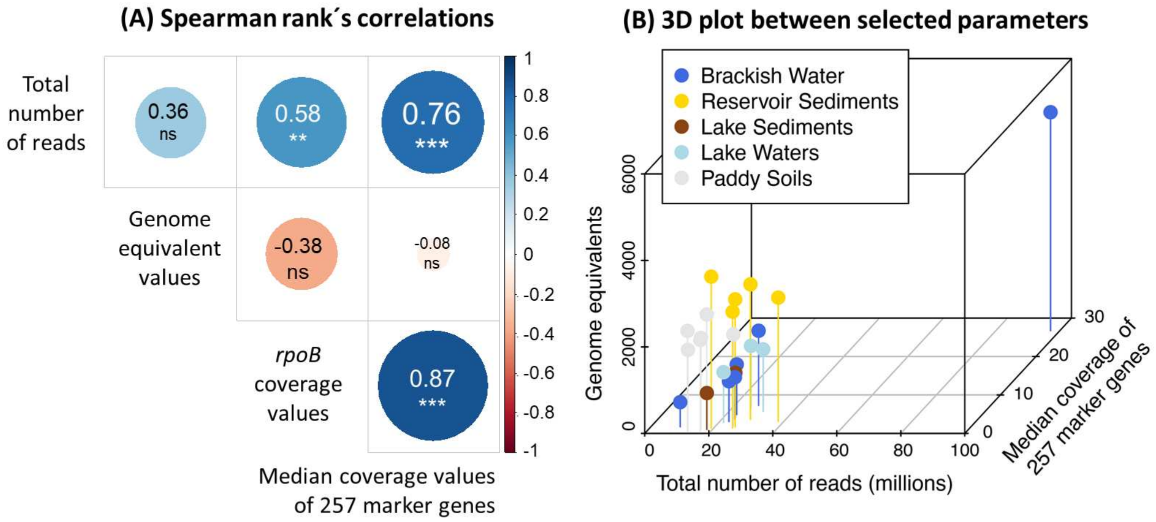


Figure 4. Plots showing correlations between metrics used for normalization. Only outputs presented here were calculated from data obtained with the single assembly approach.

4. Discussion

4.1 Identification of true *hgc* genes from environmental genomic data

The absence of cultured representatives of *hgc*$^+$ microorganisms from novel clades (i.e., outside the Desulfobacterota, Firmicutes, Methanomicrobia) with experimentally validated Hg-methylating capability (Gilmour et al., 2013; 2018) hampers confirmation that newly discovered *hgc* genes from environmental samples truly code for Hg methylating enzymes. Indeed, the recent analysis of publicly available metagenomes revealed the high diversity of microbial lineages with *hgc*$^+$ microorganisms, with the vast majority yet uncultured and therefore unstudied for Hg methylation activity (Gionfriddo et al., 2019; McDaniel et al., 2020). To date, all *hgcA*$^+$ microorganisms that have been experimentally tested have been found shown to produce MeHg (except for those with fused *hgcAB*-like sequences) (Gilmour et al., 2013; 2018), and protein modeling of novel *hgcA* sequences suggest they have comparable active sites to HgcA sequences in experimentally verified Hg methylators. Therefore, although recent findings revealed relationships between microbial expression of *hgc* transcripts and MeHg formation in the environment (Capo, Feng et al. 2022 bioRxiv), and some putative *hgcAB* genes have been computationally modelled to possess functionality for methylation (Gionfriddo et al. 2016, Lin et al. 2021), we remain cautious about defining true *hgc* genes from environmental samples. As such, some studies have qualified *hgc* genes found in the environment as *hgc* genes (e.g., Gionfriddo et al., 2016; Bowman et al., 2020; Villar et al., 2020; Capo et al., 2020).

Here, we defined three stringency cutoffs to describe *hgcA* genes in environmental metagenomes. By definition, the HgcA-encoding genes detected with the high stringency cutoffs include the key amino acid residues (i.e., the cap helix motif N[V/I]WC[A/S][A/G/S]GK), Parks et al., 2013) present in HgcA from known Hg methylators.

In contrast, all other hits to the HMM, from moderate and low stringency cut-offs, lack these amino acid residues. To date none of the isolates lacking these key amino acid residues has been found to methylate Hg, or no cultured isolate exists to test for Hg methylation capability (Gilmour et al., 2018). Substitution of some of these amino acids in the cap helix of HgcA may not result in loss of Hg methylation activity, as demonstrated by site-directed mutagenesis experiments with *P. mercurii* ND132 (Smith et al., 2015). However, in addition to the cap helix domain of HgcA, the transmembrane domain of HgcA may also be required for Hg methylation activity. Unfortunately, the transmembrane region of HgcA has no detectable sequence homology (Cooper et al., 2020).

Thus, we recommend using the high stringency cutoff defined in the present study for routine identification of *hgcA* from environmental metagenomes. Lower stringency could reveal novel HgcA sequences that have lower similarity to HgcA from known Hg methylators, but if the lower stringency cutoff is used, we advise careful manual inspection of the sequences to ensure that they have important motifs and other HgcA features like the cap-helix region. If the amino acid sequence in the cap helix domain is highly divergent from known sequences, we recommend protein modeling efforts to determine if the active site is similar enough to known sequences to validate classification as HgcA. Additional verification of true HgcA sequences include prediction of transmembrane domain regions (e.g., using TMHMM software, Krogh et al., 2001) and identification of other key conserved residues (Parks et al., 2013; Smith et al., 2015; Jones et al., 2019). A combination of several methods will certainly help to improve our description of *hgcA* genes in the coming years.

4.2 Effectiveness of the Hg-MATE database

The Hg-MATE database originates from the combination of two recent works (Gionfriddo et al., 2019; McDaniel et al., 2020). The present work is a collaborative project of the Meta-Hg working group that aimed to provide a living database that will be periodically updated. It provides several useful tools (HMM profiles and references phylogenetic trees) and a documented workflow that allows for the identification of *hgc* genes for easy comparison between studies. One major advantage of Hg-MATE is the assignment of NCBI taxonomy IDs (txid) to *hgcA* genes allowing for easy comparison with datasets from other studies that also use the Hg-MATE database (Datasheet 1B). In contrast, outputs from previous *hgc*-related studies are difficult to compare with each other because *hgc* taxonomic identification is usually done with different in-house databases and/or phylogenetic tools, and is based on the manual inspection of phylogenetic trees increasing the level of uncertainties and subjectivity in taxonomic identification. While the used *pplacer* approach here is not perfect - since phylogenetic relatedness of the gene does not necessarily mean the same organismal taxonomy because of potential horizontal gene transfer (McDaniel et al., 2020) - it is a standardized approach allowing for a robust and automated identification of *hgc* genes from metagenomes.

A side-by-side comparison of previous and present taxonomic identification of putative Hg methylators is presented in this section. For water and sediment metagenomes from Lake Manganika our identification by HgcA phylogeny showed consistent results with previous identification from *hgc*$^+$ MAGs (Jones et al., 2019), with Desulfobacterota, Aminicenantes, Kiritimatiellaeota and Spirochaetes being the predominant putative Hg methylators. In the case of Baltic Sea water metagenomes, the comparison of our Hg-MATE taxonomy identification with the previous identification using a set of *hgc* sequences from Podar et al. (2015) revealed consistency in the predominant *hgc*$^+$ groups detected (Desulfobacterota, Spirochaetes, Kiritimatiellota) but noticeable differences for others i.e., Planctomycetes and Verrucomicrobia (Datasheet 1B). Consistent with previous characterization, reservoir sediments were characterized by predominant *hgc*$^+$ Methanomicrobia, Desulfobacterota, Bacteroidetes, and Chloroflexota. Finally, in paddy soils, Liu et al. (2018) identified mostly *hgc*$^+$ Desulfobacterota, Firmicutes and Methanomicrobia while, in the present study, the two last microbial groups were found less predominant to the benefit of *hgc*$^+$ Nitrospirae and Chloroflexota.

In addition to using phylogenetic placements of *hgc* genes in reference trees from the Hg-MATE database, a more precise approach to identification of putative Hg methylators is probably the identification of *hgc*$^+$ MAGs (i.e., Jones et al., 2019; Peterson et al., 2020; Lin et al., 2021). However, the recovery of MAGs from metagenomes is not always possible due to (i) the difficulty of obtaining MAGs from certain environments such as sediments and (ii) the low predominance of Hg methylators compared to other microorganisms in the environment, and therefore the lower probability of recovering *hgc*$^+$ MAGs. A recent work revealed the good congruence between the identification of hgc$^+$ MAGs and a *hgc* phylogeny based on Hg-MATE phylogeny (Capo, Feng et al., 2022 bioRxiv) highlighting that both approaches could be used to ensure the reliability in the identification of Hg methylators.

4.3 Assembly methods depend of the diversity of the metagenome

The increasing amount of publicly available environmental genomic data (Thompson et al., 2017; Nayfach et al., 2021) opens avenues to answer ecological questions related to the biogeography patterns and dispersal barriers of Hg methylators in interconnected systems (such as the global ocean and coastal systems). Co-assembly of multiple metagenomes has been shown to have many important benefits compared to single assemblies including improved binning and better recovery of low abundance environmental genomes from studies that use multiple low-coverage metagenomes. However, co-assembly requires higher computational costs and potentially masks microdiversity by collapsing the genomes of multiple related strains into a single MAG (Narasingarao et al., 2012; Van der Walt et al., 2017; Ramos-Barbero et al., 2019; Tamames et al., 2020; Paoli et al., 2021). Here, we compared *hgcA* recovery from single assembled metagenomes versus co-assemblies of multiple metagenomes from the same environment. In all cases except one, co-assembly significantly increased the recovery of *hgcA* genes (Fig S4). Additionally, we showed that when the diversity and composition of the *hgcA*$^+$ community was compared across all the samples included in the analysis, single assemblies and co-assemblies performed similarly in this regard, suggesting that also single metagenomes can provide adequate information (similar level of *hgc* coverage and detected diversity) on the *hgc*$^+$ community.

Differences in the diversity of environments can have an effect on the recovery of *hgc* genes from metagenomes. Nonpareil diversity index values of the metagenomes ranged between 18.7 and 23.7 with the highest being found in paddy soils metagenomes (Fig. S1, Datasheet 1A). Here, for the paddy soils that exhibited higher Non-pareil diversity index values (Fig S1), consistently with Rodriguez-R and Konstantinidis (2014), the co-assembly approach outperforms single sample assemblies in the recovery of *hgc* genes (Fig 3). Noticeably, although no identical HgcA amino acid sequences were detected between single assembly and co-assembly approach, identical DsrA amino acid sequences were observed. We hypothesized that the low proportion of *hgcA* genes in metagenomes, compared to *dsrA* genes, explained such discrepancies, although it did not strongly impact the overall *hgcA* coverage values recovery. In these situations, we recommend aiming for either higher depth of coverage or sequencing of multiple adjacent or linked metagenomes or replicates from a single sample. In contrast, we recommend avoiding the co-assembly of metagenomes from different environments that could produce more misassembles and chimerism (Mikheenko et al., 2016; Sczyrba et al., 2017; Tamames et al., 2020). For other environments such as brackish and lake waters, our work highlights that using the marky-coco pipeline based on a single assembly approach provide similar results to a co-assembly approach in detecting *hgc* genes.

4.4 Robust normalization methods are needed for quantitative inferences

The normalization of gene counts from environmental metagenomes and metatranscriptomes is a key aspect of works aiming to study the prevalence of certain microorganisms in specific environments (Pereira et al., 2018; Salazar et al., 2019; Pierella Karlusich et al., 2022). In *hgcAB* omics studies, the number of mapped reads and the coverage values of marker genes or housekeeping genes is usually used to normalize the coverage values of *hgc* genes (Lin et al., 2021; Vigneron et al., 2021; Tada et al., 2021; Capo et al., 2022). Tests here revealed that a wide range of contrasting normalization methods all provided reasonable abundance estimates that were significantly correlated with one another with the exception of genome equivalent values (Fig 4). Non-significant correlations found between genome equivalent values and other metrics can be explained by the weaker relationships observed for the metrics in paddy soils and reservoir sediments metagenomes, while metrics from brackish waters, lake sediment and waters appear to have linear relationships. Therefore, we do not strongly recommend any single method over others. Instead, we suggest that it may be prudent to report data that employ multiple normalization methods to allow for easy comparisons to be carried out between studies. Such normalizations can without too much of an effort be included in the supporting information for later usage. Suggested normalization methods include the total number of prokaryotic reads, coverage values of *rpoB* genes and the median coverage values of 257 marker genes (example in Datasheet 1E).

5. Conclusion

The study of the taxonomic diversity and metabolic capacities of microorganisms involved in Hg methylation will lead to a better understanding of the environmental factors triggering microbial methylation of divalent Hg. Although metagenomic and metatranscriptomic-based studies have provided better insights into the environmental role of those microorganisms, there is still a need to standardize methods to detect *hgc* genes from environmental omic data. Furthermore, since Hg methylators often constitute such a small proportion of the microbiome, methods outlined in this study provide best practices for improving their detection and recovery from metagenomes. We provide here an up-to-date *hgc* gene catalogue, Hg-MATE database v1, and the marky-coco bioinformatic pipeline to detect, identify and count *hgc* genes from metagenomes. We recommend using our high stringency cutoff to detect hgcA genes in metagenomes and applying our protocol in future prospects of Hg methylation genes, especially for cross-comparison between studies. Finally, although a co-assembly approach should be chosen when analyzing metagenomes from highly diverse environments (i.e., paddy soils), we recommend using marky-coco pipeline, based on a de novo assembly for recovering *hgc* genes in metagenomes from aquatic environments.

Acknowledgements

This work was funded by the Severo Ochoa Excellence Program postdoctoral fellowship awarded in 2021 to Eric Capo (CEX2019-000928-S), the Swedish research council Formas (grant 2018-01031), the EMFF-Blue Economy project MER-CLUB (grant 863584). Caitlin Gionfriddo was a Robert and Arlene Kogod Secretarial Scholar with the Smithsonian Environmental Research Center while conducting work described in this manuscript. Benjamin Peterson was funded as a postdoctoral research associate by the National Science Foundation (Award 1935173) during the work in this manuscript. The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) using the compute project SNIC 2021/5-53. Some of the computations for compiling the Hg-MATE database were conducted on the Smithsonian High Performance Cluster (SI/HPC), Smithsonian Institution. https://doi.org/10.25572/SIHPC. The authors are thankful to the members of the Meta-Hg working group (https://ercapo.wixsite.com/meta-hg) and Mersorcium network (https://mersorcium.github.io/).

Conflict of interest

The author declares no conflict of interest

Benefit-sharing statement

Benefits from this research is the creation and curation of Hg-MATE database (https://doi.org/10.25573/serc.13105370.v1) and release of the bioinformatic pipeline marky-coco (https://github.com/ericcapo/marky-coco).

Data availability statements

All metagenomes analyzed in this study are of public access as described in Table 2.

References

Adler, D., Murdoch, M. D., (2019). Package 'rgl'.

Alneberg, J., Sundh, J., Bennke, C., Beier, S., Lundin, D., Hugerth, L. W., … Andersson, A. F. (2018). Data descriptor: BARM and balticmicrobeDB, a reference metagenome and interface to meta-omic data for the baltic sea. *Scientific Data*, *5*, 1–10. doi: 10.1038/sdata.2018.146

Bouchet, S., Goñi-Urriza, M., Monperrus, M., Guyoneaud, R., Fernandez, P., Heredia, C., … Amouroux, D. (2018). Linking Microbial Activities and Low-Molecular-Weight Thiols to Hg Methylation in Biofilms and Periphyton from High-Altitude Tropical Lakes in the Bolivian Altiplano. *Environmental Science and Technology*, *52*(17), 9758–9767. doi: 10.1021/acs.est.8b01885

Bowman, K. L., Collins, R. E., Agather, A. M., Lamborg, C. H., Hammerschmidt, C. R., Kaul, D., … Elias, D. A. (2020). Distribution of mercury-cycling genes in the Arctic and equatorial Pacific Oceans and their relationship to mercury speciation. *Limnology and Oceanography*, *65*(S1), S310–S320. doi: 10.1002/lno.11310

Bravo, A. G., & Cosio, C. (2020). Biotic formation of methylmercury: A bio–physico–chemical conundrum. *Limnology and Oceanography*, *65*(5), 1010–1027. doi: 10.1002/lno.11366

Bravo, A., Peura, S., Buck, M., Ahmed, O., Mateos-Rivera, A., Ortega, S., … Bertilsson, S. (2018). Methanogens and Iron-Reducing Bacteria : the Overlooked Members of Mercury-Methylating Microbial Communities in. *Applied and Environmental Microbiology*, *84*(23), 1–16.

Capo, E., Bravo, A. G., Soerensen, A. L., Bertilsson, S., Pinhassi, J., Feng, C., … Björn, E. (2020). Deltaproteobacteria and Spirochaetes-Like Bacteria Are Abundant Putative Mercury Methylators in Oxygen-Deficient Water and Marine Particles in the Baltic Sea. *Frontiers in Microbiology*, *11*(574080), 1–11. doi: 10.3389/fmicb.2020.574080

Capo, E., Broman, E., Bonaglia, S., Bravo, A. G., Bertilsson, S., Soerensen, A. L., … Björn, E. (2022). Oxygen-deficient water zones in the Baltic Sea promote uncharacterized Hg methylating microorganisms in underlying sediments. *Limnology and Oceanography*, *67*(1), 135–146. doi: 10.1002/lno.11981

Capo, E., Feng, C., Bravo, A. G., Bertilsson, S., Soerensen, A. L., Pinhassi, J., … Björn, E. (2022). Abundance and expression of hgcAB genes and mercury availability jointly explain methylmercury formation in stratified brackish waters. *BioRxiv*.

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, *36*(6), 1925–1927. doi: 10.1093/bioinformatics/btz848

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. doi: 10.1093/bioinformatics/bty560

Christensen, G. A., Gionfriddo, C. M., King, A. J., Moberly, J. G., Miller, C. L., Somenahally, A. C., … Elias, D. A. (2019). Determining the Reliability of Measuring Mercury Cycling Gene Abundance with Correlations with Mercury and Methylmercury Concentrations. *Environmental Science and Technology*, *53*(15), 8649–8663. doi: 10.1021/acs.est.8b06389

Compeau, G. C., & Bartha, R. (1985). Sulfate-reducing bacteria: Principal methylators of Mercury in Anoxic Estuarine Sediment. *Applied and Environmental Microbiology*, *50*(2), 498–502.

Cooper, C. J., Zheng, K., Rush, K. W., Johs, A., Sanders, B. C., Pavlopoulos, G. A., … Parks, J. M. (2020). Structure determination of the HgcAB complex using metagenome sequence data: insights into microbial mercury methylation. *Communications Biology*, *3*(1), 1–9. doi: 10.1038/s42003-020-1047-5

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*(suppl), W29–W37. doi: 10.1093/nar/gkr367

Fleming, E. J., Mack, E. E., Green, P. G., & Nelson, D. C. (2006). Mercury methylation from unexpected sources: molybdate-inhibited freshwater sediments and an iron-reducing bacterium. *Applied and Environmental Microbiology*, *72*(1), 457–464. doi: 10.1128/AEM.72.1.457-464.2006

Gilmour, C. C., Elias, D. A., Kucken, A. M., Brown, S. D., Palumbo, A. V., Schadt, C. W., & Wall, J. D. (2011). Sulfate-reducing bacterium Desulfovibrio desulfuricans ND132 as a model for understanding bacterial mercury methylation. *Applied and Environmental Microbiology*, *77*(12), 3938–3951. doi: 10.1128/AEM.02993-10

Gilmour, C. C., Podar, M., Bullock, A. L., Graham, A. M., Brown, S. D., Somenahally, A. C., … Elias, D. A. (2013). Mercury methylation by novel microorganisms from new environments. *Environmental Science and Technology*, *47*(20), 11810–11820. doi: 10.1021/es403075t

Gilmour, C. C., Bullock, A. L., McBurney, A., Podar, M., & Elias, D. A. (2018). Robust mercury methylation across diverse methanogenic Archaea. *MBio*, *9*(2), 1–13. doi: 10.1128/mBio.02403-17

Gionfriddo, C. M., Tate, M. T., Wick, R. R., Schultz, M. B., Zemla, A., Thelen, M. P., … Moreau, J. W. (2016). Microbial mercury methylation in Antarctic sea ice. *Nature Microbiology*, *1*(August), 1–12. doi: 10.1038/nmicrobiol.2016.127

Gionfriddo C, Podar M, Gilmour C, Pierce E, Elias D. (2019) ORNL Compiled Mercury Methylator Database https://www.osti.gov/dataexplorer/biblio/dataset/1569274

Gionfriddo, C. M., Wymore, A. M., Jones, D. S., Wilpiszeski, R. L., Lynes, M. M., Christensen, G. A., … Elias, D. A. (2020). An Improved hgcAB Primer Set and Direct High-Throughput Sequencing Expand Hg-Methylator Diversity in Nature. *Frontiers in Microbiology*, *11*, 2275. doi: 10.3389/fmicb.2020.541554

Hamelin, S., Amyot, M., Barkay, T., Wang, Y., & Planas, D. (2011). Methanogens: Principal methylators of mercury in lake periphyton. *Environmental Science and Technology*, *45*(18), 7693–7700. doi: 10.1021/es2010072

Harrell, F., & Harrell, M. (2013). Package 'Hmisc .' *CRAN*, *235*(6).

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., … Ban, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, *1*(16048), 1–6. doi: 10.1038/nmicrobiol.2016.48

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. doi: 10.1186/1471-2105-11-119

Jones, D. S., Walker, G. M., Johnson, N. W., Mitchell, C. P. J., Coleman Wasik, J. K., & Bailey, J. V. (2019). Molecular evidence for novel mercury methylating microorganisms in sulfate-impacted lakes. *ISME Journal*, *13*(7), 1659–1675. doi: 10.1038/s41396-019-0376-1

Kerin, E. J., Gilmour, C. C., Roden, E., Suzuki, M. T., Coates, J. D., & Mason, R. P. (2006). Mercury methylation by dissimilatory iron-reducing bacteria. *Applied and Environmental Microbiology*, *72*(12), 7919–7921. doi: 10.1128/AEM.01602-06

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. . (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology*, *305*(3), 567–580. doi: 10.1006/JMBI.2000.4315

Kronberg, R.-M., Jiskra, M., Wiederhold, J. G., Björn, E., & Skyllberg, U. (2016). Methyl Mercury Formation in Hillslope Soils of Boreal Forests: The Role of Forest Harvest and Anaerobic Microbes. *Environmental Science & Technology*, *50*(17), 9177–9186. doi: 10.1021/acs.est.6b00762

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. doi: 10.1093/molbev/msy096

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. doi: 10.1038/nmeth.1923

Le, S. Q., & Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, *25*(7), 1307–1320. doi: 10.1093/molbev/msn067

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*, *47*(W1), W256–W259. doi: 10.1093/nar/gkz239

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2016). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674–1676. doi: 10.1093/bioinformatics/btv033

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. doi: 10.1093/bioinformatics/btt656

Lin, H., Ascher, D. B., Myung, Y., Lamborg, C. H., Hallam, S. J., Gionfriddo, C. M., … Moreau, J. W. (2021). Mercury methylation by metabolically versatile and cosmopolitan marine bacteria. *The ISME Journal*, *15*, 1810–1825. doi: 10.1038/s41396-020-00889-4

Liu, Y. R., Johs, A., Bi, L., Lu, X., Hu, H. W., Sun, D., … Gu, B. (2018). Unraveling Microbial Communities Associated with Methylmercury Production in Paddy Soils. *Environmental Science and Technology*, *52*(22), 13110–13118. doi: 10.1021/acs.est.8b03052

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*(1), 538. doi: 10.1186/1471-2105-11-538

McDaniel, E., Peterson, B., Stevens, S., Tran, P., Anantharaman, K., & McMahon, K. (2020). Expanded Phylogenetic Diversity and Metabolic Flexibility of Mercury-Methylating Microorganism. *MSystems*, *5*(4), 1–21.

Mikheenko, A., Saveliev, V., & Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, *32*(7), 1088–1090. doi: 10.1093/bioinformatics/btv697

Millera Ferriz, L., Ponton, D. E., Storck, V., Leclerc, M., Bilodeau, F., Walsh, D. A., & Amyot, M. (2021). Role of organic matter and microbial communities in mercury retention and methylation in sediments near run-of-river hydroelectric dams. *Science of the Total Environment*, *774*(February), 145686. doi: 10.1016/j.scitotenv.2021.145686

Narasingarao, P., Podell, S., Ugalde, J. A., Brochier-Armanet, C., Emerson, J. B., Brocks, J. J., … Allen, E. E. (2012). De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *The ISME Journal*, *6*(1), 81–93. doi: 10.1038/ismej.2011.78

Nayfach, S., & Pollard, K. S. (2015). Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biology*, *16*(1), 1–18. doi: 10.1186/s13059-015-0611-7

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., … Eloe-Fadrosh, E. A. (2021). A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, *39*(4), 499–509. doi: 10.1038/s41587-020-0718-6

Oksanen, A. J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., Hara, R. B. O., … Wagner, H. (2015). The vegan package. Community ecology package. *Http://CRAN.R-Project.Org/Package=vegan*.

Paoli, L., Ruscheweyh, H.-J., Forneris, C. C., Kautsar, S., Clayssen, Q., Salazar, G., … Sunagawa, S. (2021). Uncharted biosynthetic potential of the ocean microbiome. *BioRxiv*, 2021.03.24.436479. doi: 10.1101/2021.03.24.436479

Parks, J. M., Johs, A., Podar, M., Bridou, R., Hurt, R. A., Smith, S. D., … Liang, L. (2013). The genetic basis for bacterial mercury methylation. *Science*, *339*(6125), 1332–1335. doi: 10.1126/science.1230667

Pereira, M. B., Wallroth, M., Jonsson, V., & Kristiansson, E. (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics*, *19*(1), 274. doi: 10.1186/s12864-018-4637-6

Peterson, B. D., McDaniel, E. A., Schmidt, A. G., Lepak, R. F., Janssen, S. E., Tran, P. Q., … McMahon, K. D. (2020). Mercury Methylation Genes Identified across Diverse Anaerobic Microbial Guilds in a Eutrophic Sulfate-Enriched Lake. *Environmental Science & Technology*, *54*(24), 15840–15851. doi: 10.1021/acs.est.0c05435

Pierella Karlusich, J. J., Pelletier, E., Zinger, L., Lombard, F., Zingone, A., Colin, S., … Bowler, C. (2022). A robust approach to estimate relative phytoplankton cell abundances from metagenomes. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.13592

Podar, M., Gilmour, C. C., Brandt, C. C., Soren, A., Brown, S. D., Crable, B. R., … Elias, D. A. (2015). Global prevalence and distribution of genes and microorganisms involved in mercury methylation. *Science Advances*, *1*(9), 1–13. doi: 10.1126/sciadv.1500675

Ramos-Barbero, M. D., Martin-Cuadrado, A.-B., Viver, T., Santos, F., Martinez-Garcia, M., & Antón, J. (2019). Recovering microbial genomes from metagenomes in hypersaline environments: The Good, the Bad and the Ugly. *Systematic and Applied Microbiology*, *42*(1), 30–40. doi: 10.1016/J.SYAPM.2018.11.001

Rodriguez-R, L. M., & Konstantinidis, K. T. (2014). Nonpareil: A redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics*, *30*(5), 629–635. doi: 10.1093/bioinformatics/btt584

Roth, S., Poulin, B. A., Baumann, Z., Liu, X., Zhang, L., Krabbenhoft, D. P., … Barkay, T. (2021). Nutrient Inputs Stimulate Mercury Methylation by Syntrophs in a Subarctic Peatland. *Frontiers in Microbiology*, *12*, 741523. doi: 10.3389/fmicb.2021.741523

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H. J., Cuenca, M., … Wincker, P. (2019). Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell*, *179*(5), 1068-1083.e21. doi: 10.1016/j.cell.2019.10.014

Selengut, J. D., Haft, D. H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W. C., … White, O. (2007). TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Research*, *35*(Database), D260–D264. doi: 10.1093/nar/gkl1043

Schaefer, J. K., Kronberg, R. M., Björn, E., & Skyllberg, U. (2020). Anaerobic guilds responsible for mercury methylation in boreal wetlands of varied trophic status serving as either a methylmercury source or sink. *Environmental Microbiology*, *22*(9), 3685–3699. doi: 10.1111/1462-2920.15134

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., … McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods*, *14*(11), 1063–1071. doi: 10.1038/nmeth.4458

Smith, S. D., Bridou, R., Johs, A., Parks, J. M., Elias, D. A., Hurt, R. A., … Wall, J. D. (2015). Site-directed mutagenesis of HgcA and HgcB reveals amino acid residues important for mercury methylation. *Applied and Environmental Microbiology*, *81*(9), 3205–3217. doi: 10.1128/AEM.00217-15

Soerensen, A. L., Schartup, A. T., Skrobonja, A., Bouchet, S., Amouroux, D., Liem-Nguyen, V., & Björn, E. (2018). Deciphering the Role of Water Column Redoxclines on Methylmercury Cycling Using Speciation Modeling and Observations From the Baltic Sea. *Global Biogeochemical Cycles*, *32*(10), 1498–1513. doi: 10.1029/2018GB005942

Tamames, J., Cobo-Simón, M., & Puente-Sánchez, F. (2019). Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes. *BMC Genomics*, *20*(1), 960. doi: 10.1186/s12864-019-6289-6

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., … Zhao, H. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, *551*(7681), 457–463. doi: 10.1038/nature24621

van der Walt, A. J., van Goethem, M. W., Ramond, J.-B., Makhalanyane, T. P., Reva, O., & Cowan, D. A. (2017). Assembling metagenomes, one community at a time. *BMC Genomics*, *18*(1), 521. doi: 10.1186/s12864-017-3918-9

Vigneron, A., Cruaud, P., Aubé, J., Guyoneaud, R., & Goñi-Urriza, M. (2021). Transcriptomic evidence for versatile metabolic activities of mercury cycling microorganisms in brackish microbial mats. *Npj Biofilms and Microbiomes*, *7*(1), 1–11. doi: 10.1038/s41522-021-00255-y

Villar, E., Cabrol, L., & Heimbürger-Boavida, L. E. (2020). Widespread microbial mercury methylation genes in the global ocean. *Environmental Microbiology Reports*, *12*(3), 277–287. doi: 10.1111/1758-2229.12829

Xu, J., Liem-Nguyen, V., Buck, M., … S. B.-F. in, & 2020, U. (2021). Mercury methylating microbial community structure in boreal wetlands explained by local physicochemical conditions. *Frontiers*. Retrieved from https://scholar.google.com/scholar?cluster=13616141145651252883&hl=en&oi=scholarr

Datasheet 1. This file includes information related to different parameters collected or measured in this work from the 29 metagenomes used in this work (A) For each metagenome, metagenome id, type of environment, non-pareil metrics, genome equivalents (Microbe Census) values, number of cleaned and mapped reads, number of *hgcA* genes, *hgcA* coverage values and normalization metrics values, *dsrA* coverage values (B) List of all *hgcA* genes detected in the 29 metagenomes with both a single assembly and co-assembly approaches, with the three stringency cutoffs. Gene length, number of mapped reads, coverage values, NBCI taxonomy txid and amino acid sequences are presented. (C) List of all *hgcA* genes detected in the 29 metagenomes with both a single assembly and co-assembly approaches, with the high stringency cutoff. Gene length, number of mapped reads, coverage values, NBCI taxonomy txid and amino acid sequences are presented. (D) List of all *dsrA* genes detected in the 29 metagenomes with both a single assembly and co-assembly approaches, with the high stringency cutoff. Gene length, number of mapped reads, coverage values and amino acid sequences are presented. (F) Coverage values of the 257 marker genes (including *rpoB*) obtained using the single assembly vs co-assembly approaches.

Supporting Information

Figure S1: Nonpareil curves for the 22 metagenomes. The plot displays the fitted models of the Nonpareil curves. The horizontal dashed lines indicate 100 (gray) and 95% (red) coverage. The empty circles indicate the size and estimated average coverage of the datasets, and the lines after that point are projections of the fitted model.

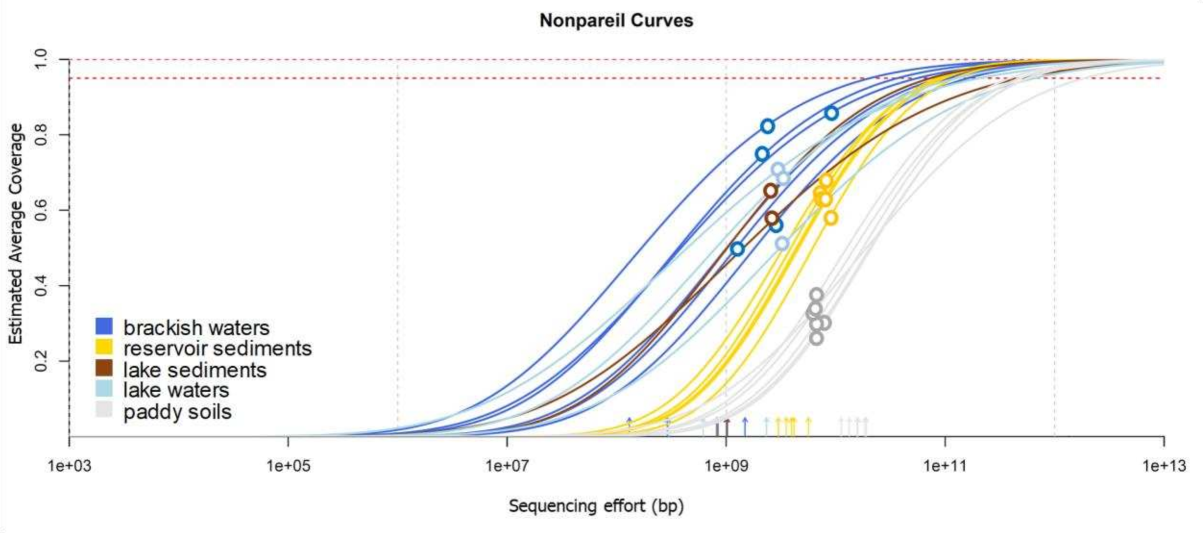


Figure S2. Distribution of *hgcA* genes in the 22 metagenomes recovered using the co-assembly ´c´ and the single assembly ´s´ methods and applying the three stringency cutoffs defined in this manuscript for the definition of *hgcA* genes. Abundance values were calculated as *hgcA* coverage values normalized by *rpoB* normalized values. Colors denote taxonomic affiliations of *hgcA* genes.

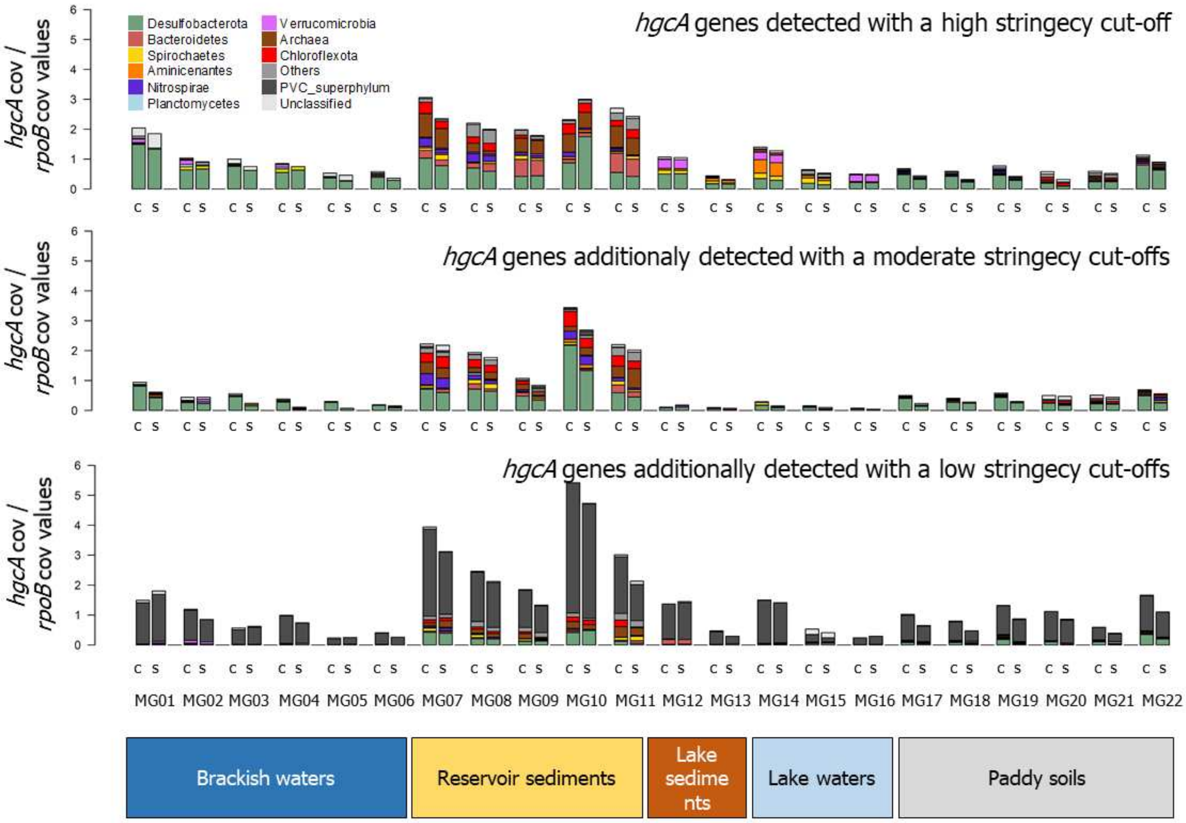
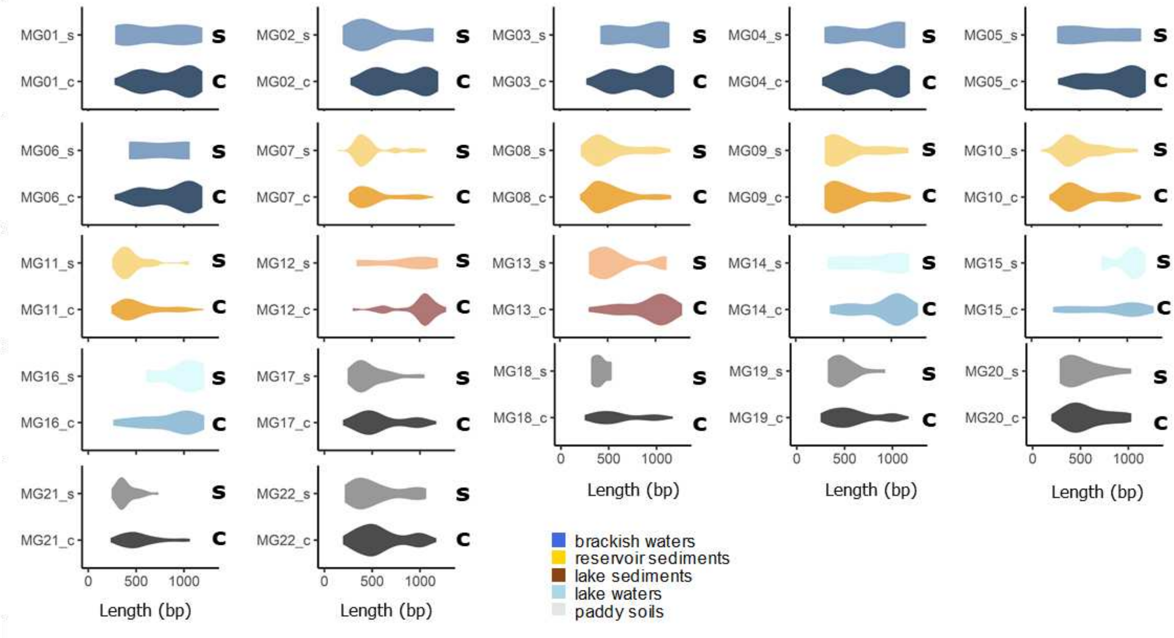
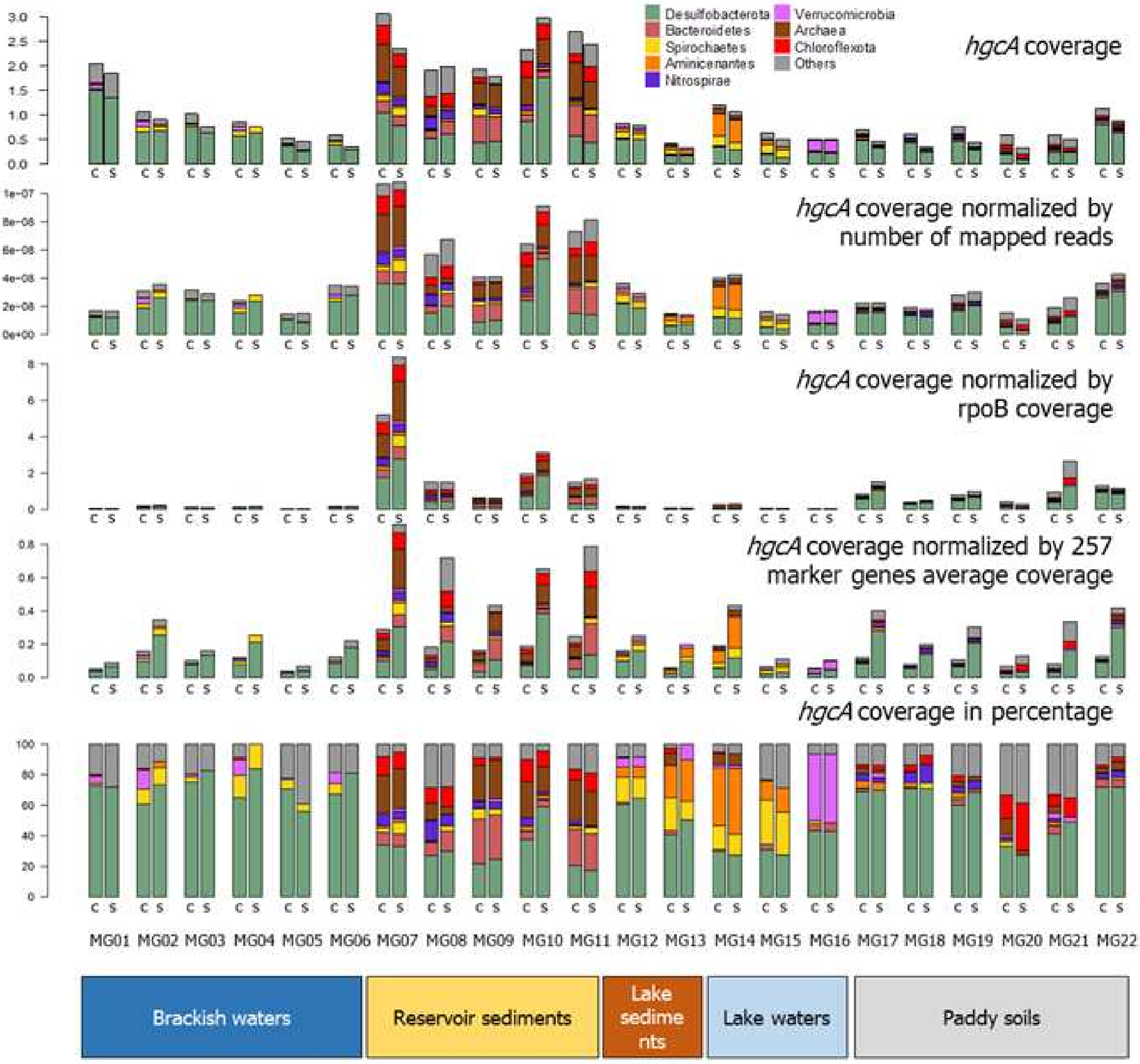


Figure S3: Violin boxplots showing, for each metagenome, the difference in *hgcA* sequence length distribution comparing the outputs of the co-assembly and the single assembly approaches.



844 **Figure S4**: Distribution of *hgcA* genes in the 22 metagenomes with the co-assembly (c) and
845 the single assembly (s) methods with different normalization methods



846