Hg-MATE-Db: Hg-cycling **M**icroorganisms in **A**quatic and **T**errestrial **E**cosystems **D**ata**b**ase

This catalogue of HgcAB amino acid sequences was primarily compiled by Caitlin Gionfriddo (Smithsonian Environmental Research Center, USA), Eric Capo (Umeå University, SE), and Benjamin Peterson (University of Wisconsin-Madison, USA), with contributions from Heyu Lin (University of Melbourne, AU), Daniel Jones (New Mexico Institute of Mining and Technology, USA), Andrea García Bravo (Institut de Ciències del Mar, Institute of Marine Sciences, ES), Stefan Bertilsson (SLU, Sveriges lantbruksuniversitet, Swedish University of Agricultural Sciences, SE), John Moreau (University of Glasgow, UK), Katherine McMahon (University of Wisconsin, USA), Dwayne Elias (Oak Ridge National Laboratory, USA), and Cynthia Gilmour (Smithsonian Environmental Research Center, USA).

Citation

Please cite us when using the Hg-MATE-Db in your work.

Gionfriddo, C., Capo, E., Peterson, B., Lin, H., Jones, D., Bravo, AG., Bertilsson, S., Moreau, J., McMahon, K., Elias, D., and Gilmour, C. (2021). Hg-MATE-Db.v1.01142021. doi:10.25573/serc.13105370

Background

Microorganisms play a significant role in regulating the form and fate of mercury (Hg) in aquatic and terrestrial ecosystems. Microbes with the *hgcAB* gene pair can produce a more toxic, and bioaccumulative form of Hg, methylmercury (MeHg). Microbes that possess the *mer* operon can demethylate and/or reduce Hg species as part of a detoxification mechanism. Improved techniques for capturing *hgcAB* and *mer* presence and diversity are necessary for identifying the major microbial players in environmental Hg cycling. The primary goal of the Hg-MATE-Db is to provide an up-to-date collated resource of Hg-cycling genes from pure culture and environmental microbial genomes and meta-omic datasets. The database will be updated regularly.

The Hg-MATE-Db.v1 contains an *hgcAB* dataset with resources for identifying key microbial producers of the toxin MeHg. Future versions of the Hg-MATE-Db will also include *hgcAB* sequences from high-throughput sequencing and clone datasets. Future versions will also contain a *mer* dataset, which will contain resources for identifying genes of the *mer*-operon that encode for demethylation of organomercurials (*merB*), reduction of inorganic Hg(II) (*merA*), as well as operon regulation (*merR*), and Hg transport across the cell (*merTPC*).

Description

The latest release, version 1 (v1), was compiled on 23 October 2020 and finalized on 14 January 2021, and contains an *hgcAB* dataset. The catalogue 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:

- pure culture/environmental microbial isolates (ISO)
- single-cell genome sequences (CEL)

- metagenome-assembled genomes (MAGs)
- environmental meta-omic contig (CON)

Only sequences with genomic identifying information (i.e. 'ISO', 'CEL', 'MAG') were used to compile resources for identifying and classifying HgcAB.

Included in the database are amino acid sequences of HgcA, HgcB, and concatenated HgcA and HgcB. If *hgcB* is not co-localized with *hgcA* in the genome and/or cannot be identified, then 'na' will be listed in the 'HgcB' sequence column. To our knowledge, both genes need to be present and encode functional proteins for a microbe to methylate Hg (see Parks et al. 2013, doi:10.1126/science.123066 and Smith et al. 2015, doi:10.1128/AEM.00217-15). One reason *hgcB* may not be identifiable in a genome is because HgcB is highly homologous to other 4Fe-4S ferredoxins, and therefore *hgcB* can be difficult to differentiate from other ferredoxin-encoding genes if not co-localized with *hgcA.* In addition, *hgcB* may be missing from 'MAGs', 'CEL' and 'CON' sequences due to incomplete coverage of the genome or incomplete contig assembly.

Some *hgcAB* genes are predicted to encode a 'fused HgcAB protein' (as defined in Podar et al. 2015, doi:10.1126/sciadv.1500675). 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 for it is unclear whether they encode for Hg-methylation capability. While they share significant sequence homology to HgcA and HgcB in confirmed Hg-methylators, to date no organism with a 'fused HgcAB' has been shown to methylate Hg in culture (see Podar et al. 2015, doi:10.1126/sciadv.1500675 and Gilmour et al. 2018, doi:10.1128/mBio.02403-17)

Table 1. Summary of HgcAB sequence types

Genome Type	Total HgcA(B) sequences	Encodes both HgcA and HgcB	Encodes 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)

- ISO = pure culture/environmental microbial isolates
- CEL = single-cell genome sequences
- MAG = metagenome-assembled genomes
- CON = environmental meta-omic contig

Resources

The resources within the Hg-MATE-Db.v1 include:

- 'Hg-MATE-Db.v1.01142021.xlsx': Excel spreadsheet catalogue with amino acid sequences and organism metadata
- FASTA files containing amino acid sequences of HgcA ('_HgcA.fas'), HgcB ('_HgcB.fas'), and concatenated HgcA-HgcB sequences ('_Hgc.fas') from pure culture/environmental isolates, single-cell genome sequences, and metagenome-assembled genomes ('ISOCELMAG')
- FASTA formatted alignments of amino acid sequences of HgcA ('_HgcA_msa.fas'), HgcB ('_hgcB_msa.fas'), and concatenated HgcA-HgcB sequences ('_Hgc_msa.fas') from pure culture/environmental isolates, single-cell genome sequences, and metagenome-assembled genomes ('ISOCELMAG')
- Hidden Markov models of aligned HgcA ('_HgcA.hmm'), HgcB ('_HgcB.hmm'), and concatenated HgcA-HgcB ('_Hgc.hmm') built from pure culture/environmental isolates, single-cell genome sequences, and metagenome-assembled genomes ('ISOCELMAG')
- Three reference packages that can be used to identify and classify: 1) the cap-helix encoding region of HgcA (Hg-MATE-Db.v1.01142021_ISOCELMAG_HgcA_CH.refpkg), for example in *Desulfovibrio desulfuricans* ND132, this encompasses the CdhD-like encoding region, sites ~37-156 of HgcA (https://www.uniprot.org/uniprot/F0JBF0); 2) full HgcA (Hg-MATE-Db.v1.01142021_ISOCELMAG_HgcA_Full.refpkg); and 3) concatenated HgcA and HgcB (Hg-MATE-Db.v1.01142021_ISOCELMAG_HgcA-HgcB.refpkg). Each reference package contains sequence alignments, hmm model, phylogenetic tree, and NCBI taxonomy.

Metadata

The following metadata is provided for each sequence in the catalogue:

- **MATE ID**: unique identifier for Hg-MATE database
- Sequence name as **[Organism Name]_Phylum/Class**
- **Type**: either isolate, single cell, contig, or MAG (see Description above for more info)
- **Hg-methylation capability**: whether the organism has been experimentally confirmed to methylate Hg
- **Published % rates**: if a confirmed methylator, then the published Hg-methylation rate as a percentage (e.g., Gilmour et al., 2018, doi.org/10.1128/mBio.02403-17)
- **Contig ID**: unique identifier for environmental contig with *hgcA*(*B*) from source dataset
- **NCBI GenBank ID**: GenBank assembly accession ID (https://www.ncbi.nlm.nih.gov/genbank/)
- **NCBI alt ID:** Alternative ID for sequence/genome in an NCBI data repository
- **NCBI:txid**: classification ID from the NCBI taxonomy database (https://www.ncbi.nlm.nih.gov/Taxonomy)
- **JGI GOLD ID**: ID of genome or metagenome in JGI GOLD database (https://gold.jgi.doe.gov/)
- **DSMZ**: provided if organism is in the DSMZ catalogue

(https://www.dsmz.de/collection/catalogue/microorganisms/catalogue)

- **ATCC**: provided if organism is a type strain from ATCC Collection (https://www.atcc.org/)
- **Strain** taxonomy identifier
- **Microbial group**: Phylum, or in some cases Class, of organism. Labeled as 'Unclassified' if no taxonomic information is available.
- **Reference**: source publication for the sequence (this is primarily for environmental contigs and sequences from meta-omic datasets not in public repositories such as NCBI or JGI). If sequence was pulled from public repository, then labeled 'Hg-MATE_catalogue'
- **Environmental medium**: Environmental material for the *hgcAB* reference sequence. Following metadata guidelines for NCBI BioSample attributes (https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/) identify the environmental material that was sampled for *hgcAB* genes. This metadata field is required for environmental reference sequences (e.g. contigs).
- **Environmental package**: Identifier of environmental setting for the *hgcAB* reference sequence. Following metadata guidelines for NCBI BioSample attributes https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/). This metadata field is required for environmental reference sequences (e.g. contigs).
- **Data repository**: link to data repository for environmental sequencing datasets. This metadata field is required for environmental reference sequences (e.g. contigs).
- **Version**: refers to version of Hg-MATE-Db

Methodology

To compile the catalogue, we started by combining two previously published HgcAB databases: Gionfriddo et al. 2020 (https://doi.org/10.3389/fmicb.2020.541554) and McDaniel et al. 2020 (https://doi.org/10.1128/mSystems.00299-20). We then added HgcAB sequences pulled from three public data repositories: NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/), JGI-IMG GOLD (https://gold.jgi.doe.gov/), and GTDB release 89 (https://gtdb.ecogenomic.org/). Sequences were compiled on 23 October 2020. HgcAB sequences were identified in these databases by hmmsearch with HgcA and HgcB hmm profiles from Gionfriddo et al. 2020. Methods for specific parts of the Hg-MATE-Db are as follows.

- Multiple sequence alignments (MSA) were built with MUSCLE implemented in MEGAX (Kumar et al., 2018) with the cluster method UPGMA
- HMM profiles for HgcA and HgcB encoding genes were built from MSAs using the hmmbuild function from the hmmer software (3.2.1 version, Finn et al., 2011, http://hmmer.org/)
- Reference packages were constructed using the program Taxtastic (https://github.com/fhcrc/taxtastic) for HgcA(B) amino acid sequences from ISO-CEL-MAGs in the Hg-MATE-Db.v1. For example, the methods for the full HgcA reference package are:
    - **RAxML_bipartitions.Hg-MATE-Db.v1.ISOCELMAG-HgcA-Full-ML-100bs-tree-rooted:** a phylogenetic tree estimated by RAxML using the GAMMA model of rate heterogeneity and LG substitution matrix. Phylogeny was inferred from

alignment of HgcA (or concatenated HgcA and HgcB sequences). The tree is rooted by HgcA paralog sequences, carbon monoxide dehydrogenases (PF03599) from non-HgcA organisms *Candidatus* Omnitrophica bacterium CG1_02_41_171 and *Thermosulfurimonas dismutan*s. These organisms were chosen because of their distinct phylogeny to HgcA+ organisms. Confidence values on branches are calculated from 100 bootstraps.

- **RAxML_info.Hg-MATE-Db.v1.ISOCELMAG_HgcA_full-ref-tree-ML100bs:** information file for phylogenetic tree produced by RAxML above
- **Hg-MATE-Db.v1.ISOCELMAG_HgcA_full.hmm:** the HMM model built using hmmbuild from the stockholm-formatted amino acid sequence alignment
- **Hg-MATE-Db.v1.ISOCELMAG_HgcA_full.stockholm:** stockholm-formatted alignment of amino acid sequences used for building HMM profile. Alignment produced from Hg-MATE-Db.v1 HgcA MSA and paralog sequences using MUSCLE in Geneious Prime
- **Hg-MATE-Db.v1.ISOCELMAG_HgcA_full_align-bmge.fasta:** FASTA-formatted alignment of amino acid sequences. Alignment produced from Hg-MATE-Db.v1 HgcA MSA and paralog sequences using MUSCLE in Geneious Prime. Sites with >90% gaps were trimmed from alignment using BMGE (Criscuolo and Gribaldo 2010, https://doi.org/10.1186/1471-2148-10-210). The filtered alignment was used for building the phylogenetic tree.
- **Seq_info.csv:** table containing NCBI taxID and organism (i.e. species name) for each reference sequence in the package.
- **Taxa.csv:** a curated NCBI taxonomy database for all sequences in the reference package. Used for classifying sequences using pplacer (Matsen et al. 2010, https://doi.org/10.1186/1471-2105-11-538). Taxonomy was pulled from the NCBI tax dump version 05-31-2020. Please note, taxonomy for some sequences may have changed since then.

Recommended Usage

For detection of *hgcAB* in metagenomes or MAGs: we recommend using the HgcA and HgcB HMM profiles and MSA files generated from reference sequences of isolates, single-cell genomes, and metagenome-assembled genomes (ISOCELMAG). The Hgc HMM profile and MSA file (i.e. concatenated HgcA and HgcB amino acid sequences) should be used when HgcA and HgcB amino acid sequences are detected in the same contig/scaffold/MAG in your analysis. For an example of how to work with these sorts of datasets, see Capo et al. 2020 (https://doi.org/10.3389/fmicb.2020.574080), Peterson et al. 2020 (https://doi.org/10.1021/acs.est.0c05435) and Lin et al. 2020 (https://doi.org/10.1101/2020.06.03.132969).

The reference packages can be used for phylogenetic analysis of HgcA(B) sequences from amplicon sequencing or meta-omic datasets. An example workflow for how to use the Hg-MATE-Db reference packages for identifying and classifying HgcA(B) sequences using hmmer (http://hmmer.org/) and pplacer (Matsen et al. 2010, https://doi.org/10.1186/1471-2105-11-538) is provided below. Further analysis and visualizations can be produced from the pplacer output files, i.e. the 'jplace' placement file and 'sqlite' classification database, using the R packages BoSSA and phyloseq in R (version 3.5.1 or later). For a more in-depth tutorial for how to use these types of reference packages for identifying and classifying HgcA sequences in high-throughput sequencing data, see Gionfriddo et al. 2020 (https://doi.org/10.3389/fmicb.2020.541554) and the accompanying tutorial (https://caitlingio.com/tutorial-for-hgcab-amplicon-sequencing-data/).

1) Start with a FASTA-formatted file of predicted protein sequences that you would like to classify (for example, translated amino acid sequences from amplicon sequencing data, or predicted open reading frames from a metagenomic dataset), sequences-to-be-classified.fasta**.** Use hmmsearch to filter out non-HgcA sequences (i.e. those that do not align with reference HgcA sequences in HMM model, Hg-MATE-Db.v1.ISOCELMAG_HgcA_full.hmm from reference package) using an inclusion E-value cutoff. We suggest starting with an E-value of 10^{-25} (1E-25), although this will differ by dataset. The E-value may need to be adjusted higher or lower in order to exclude non-HgcA CdhD-encoding sequences without losing HgcA. Look at sequences in the hmm-table and hmm-output that are around the inclusion threshold cut-off to determine the best cut-off for your dataset). Outputs include a table of query sequences and E-values (hmm_table), text file showing alignment of all query sequences with reference HgcA and E-values (hmm_output) and alignment of query sequences that passed inclusion threshold with reference sequences (hmm_alignment).

```
hmmsearch --tblout hmm_table -o hmm_output --incE 1E-25 -A hmm_alignment Hg-MATE-Db.v1.ISOCELMAG_HgcA_full.hmm sequences-to-be-classified.fasta
```

2) Align filtered query sequences from previous step (hmm_alignment) to stockholm formatted alignment of reference sequences in reference package (Hg-MATE-Db.v1.ISOCELMAG_HgcA_full.stockholm) using hmm model (Hg-MATE-Db.v1.ISOCELMAG_HgcA_full.hmm) producing a Stockholm formatted alignment of filtered query sequences and reference sequences (hmm_alignment.sto) that can be used for classification

```
hmmalign -o hmm_alignment.sto -- mapali Hg-MATE-Db.v1.ISOCELMAG_HgcA_full.stockholm Hg-MATE-Db.v1.ISOCELMAG_HgcA_full.hmm hmm_alignment
```

3) Using the program pplacer (Matsen et al. 2010, https://doi.org/10.1186/1471-2105-11-538) place aligned query sequences (hmm_alignment.sto) onto the HgcA reference tree (RAxML_bipartitions.Hg-MATE-Db.v1.ISOCELMAG-HgcA-Full-ML-100bs-tree-rooted) in the reference package (Hg-MATE-Db.v1.01142021_ISOCELMAG_HgcA_Full.refpkg). We suggest specifying the following arguments when using pplacer: calculate posterior probabilities based on alignment of query sequence with reference sequence (-p), specify that the maximum number of placements to keep is one (--keep-at-most 1), and set the maximum branch length to 1 (--max-pend 1) (this ensures that sequences that

are highly dissimilar to HgcA are not placed on the tree. Output will be a jplace file (hmm_alignment.jplace).

```
pplacer --keep-at-most 1 --max-pend 1 -p -c Hg-MATE-Db.v1.01142021_ISOCELMAG_HgcA_Full.refpkg hmm_alignment.sto
```

4) Make a sqlite-formatted database (classify_output) for classifications in the next step. The ‘rppr’ and ‘guppy’ commands are part of the program pplacer.

```
rppr prep_db --sqlite classify_output -c Hg-MATE-Db.v1.01142021_ISOCELMAG_HgcA_Full.refpkg
```

5) Assign taxonomy to query sequences based on placement from step 3 (hmm_alignment.jplace). If you used posterior probability in step 3, also specify it here (--pp). This step will use the lowest common ancestor of the branch in the reference tree to classify the query sequences, and will use a 90% confidence cut-off for identification as default. Output will write classifications to the sqlite database made in the previous step (classify_output).

```
guppy classify -c Hg-MATE-Db.v1.01142021_ISOCELMAG_HgcA_Full.refpkg --pp --sqlite classify_output hmm_alignment.jplace
```

6) Use ‘guppy to_csv’ to write classifications to csv file:

```
guppy to_csv --point-mass --pp -o classifications.csv hmm_alignment.jplace
```

7) Use ‘guppy tog’ to produce a visualization showing placements of query sequences on reference tree. The Newick-formatted tree can be opened in tree-viewing software such as archeopteryx (https://github.com/cmzmasek/archaeopteryx-js) or FigTree (https://github.com/rambaut/figtree/releases).

```
guppy tog --pp -o classifications_on_tree.nwk hmm_alignment.jplace
```