DiTing: A Pipeline to Infer and Compare Biogeochemical Pathways from Metagenomic and Metatranscriptomic Data

Running title: DiTing for Biogeochemical Pathways

Chun-Xu Xue[1†], Heyu Lin[2†], Xiao-Yu Zhu[1], Jiwen Liu[1,3,4], Gary Rowley[5], Jonathan D Todd[5], Meng Li[6], Xiao-Hua Zhang[1,3,4*]

[1]College of Marine Life Sciences, and Institute of Evolution & Marine Biodiversity, Ocean University of China, Qingdao 266003, China

[2]School of Earth Sciences, University of Melbourne, Parkville, Victoria 3010, Australia

[3]Laboratory for Marine Ecology and Environmental Science, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266071, China

[4]Frontiers Science Center for Deep Ocean Multispheres and Earth System, Ocean University of China, Qingdao 266100, China

[5]School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, United Kingdom.

[6]Shenzhen Key Laboratory of Marine Microbiome Engineering, Institute for Advanced Study, Shenzhen University, Shenzhen 518060, China.

[†]Chun-Xu Xue and Heyu Lin contributed equally to this work.

*To whom correspondence should be addressed.

Abstract

Metagenomics and metatranscriptomics are powerful tools to uncover key microbes and processes driving biogeochemical cycling in natural ecosystems. Currently available databases depicting metabolic functions from metagenomic/metatranscriptomic data are not dedicated to biogeochemical cycles.

There are no databases encompass genes involved in the cycling of dimethylsulfoniopropionate (DMSP), an abundant organosulfur compound. Additionally, a recognized normalization mode to estimate and compare the relative abundance and environmental importance of pathways from metagenomic and metatranscriptomic data has not been available. These limitations impact the ability to accurately relate key microbial driven biogeochemical processes to differences in environmental conditions. Thus, an easy to use specialized tool that infers and visually compares the potential for biogeochemical processes, including DMSP cycling, is urgently required. To solve these issues, we developed DiTing, a tool wrapper to infer and compare biogeochemical pathways among a set of given metagenomic or metatranscriptomic reads in one step, based on the KEGG (Kyoto Encyclopedia of Genes and Genomes) and a manually created DMSP cycling gene database. Accurate and specific formulas for over 100 pathways were developed to calculate their relative abundance. Output reports detail the relative abundance of biogeochemically-relevant pathways in both text and graphical format. We applied DiTing to metagenomes from simulated data, hydrothermal vents and the *Tara* Ocean project. The DiTing outputs were consistent with genetic feature of genomes used in simulated benchmark data, and also demonstrated that the predicted functional profiles correlated strongly with changes in environmental conditions. DiTing can now be confidently applied to wider metagenomic and metatranscriptomic datasets.

Availability and implementation: https://github.com/xuechunxu/DiTing

Contact: xhzhang@ouc.edu.cn

Supplementary information: Supplementary data are available at Molecular Ecology Resources online.

Introduction

Microbial communities play integral and unique roles in mediating global biogeochemical cycles. Sequencing techniques, such as amplicon sequencing (Bokulich et al., 2013), whole-genome sequencing (Jones et al., 2016), genome-resolved metagenomics (Parks et al., 2017; Xue et al., 2020b) and shotgun metagenomic sequencing (Sharpton et al., 2014; Xue et al., 2020a), are widely used to characterize the genetic potential of microbial communities. Metagenomics is an important tool to unravel the diversity, function and ecology of complex microbial ecosystems via quantification of the genetic potential for various biogeochemical pathways within microbial communities (Riesenfeld et al., 2004). Moreover, metatranscriptomic data present a more accurate scenario of processes occurring within ecosystems because these methodologies move past genetic potential and report on the transcription of biogeochemical pathways (Aguiar-Pulido et al., 2016; Shakya et al., 2019). Previous studies have predicted community functions according to gene annotation against several established databases, e.g., KEGG (Ogata et al., 2000), COG (Tatusov et al., 2000), MetaCyc (Caspi et al., 2006), Pfam (Finn et al., 2014), TIGRfam (Selengut et al., 2007), SEED (Ross et al., 2014), and eggNOG (Huertacepas et al., 2016). However, these functional annotations are not dedicated to biogeochemical cycling and lack comprehensive lists of annotated genes for important cycles. FOAM (Functional Ontology Assignments for Metagenomes; Prestat et al., 2014) is a functional gene database for environmental datasets that includes biogeochemical cycles, however, this database lacks visualization, and annotates all protein sequences with a universal threshold value, which may lead to prediction biases. Furthermore, some metabolic pathways, e.g. the cycling of dimethylsulfoniopropionate (DMSP), a key marine osmolyte, nutrient, and signaling molecule, with important roles in sulfur cycling (Curson *et al*., 2011; Zhang *et al*., 2019), lack an accurate and reviewed database for annotating the key metabolic genes. These limitations force researchers to undertake often tricky and time-consuming gathering of gene sequences from primary research and collate them into robust local

databases (Acinas et al., 2019; Dombrowski et al., 2018; Llorens-Marès et al., 2015; Zhang et al., 2018). This can also lead to challenges for downstream interpretation, organization and visualization.

Additionally, there is no recognized and prepared normalization method to estimate and compare the relative abundance of a pathway in metagenomic and metatranscriptomic data. In some studies, the relative abundance of every gene in a biogeochemical pathway was added together to estimate the relative abundance of the pathway (Ganesh et al., 2014; Petter et al., 2013; Smedile et al., 2013), which is unsuitable to infer and compare pathways. For example, thiosulfate disproportionation (thiosulfate -> sulfide & sulfite) is catalyzed by thiosulfate reductase, which is encoded by three genes (*phsABC*). Thus, the relative abundance of thiosulfate disproportionation pathway should be equal to the mean relative abundance of *phsABC* instead of the sum of *phsABC* relative abundance together. This normalization mode was applied in somerecent studies (Llorens-Marès et al., 2015, Graham et al., 2018), but no simple tool to achieve this is currently available. In addition, there are few easy methods for high throughput comparison and visualization of samples. Therefore, new automated tools to identify, quantify, and compare the abundance and/or transcription of genes and pathways for biogeochemical cycles, including the DMSP cycle, are needed.

Here we developed the software DiTing, a pipeline to infer and compare biogeochemical pathways in metagenomic and metatranscriptomic data. DiTing is named after a Chinese mythical creature who knows everything when he puts his ears on the Earth's surface. Similarly, scientists can gain robust knowledge on microbial driven biogeochemical cycles from environmental ‘omic data after analysis with DiTing. DiTing annotates protein sequences based on the KEGG database (Ogata et al., 2000) for most microbial-mediated biogeochemical cycles, and an in-house database specifically for cycling of DMSP, and then estimates the relative abundance of corresponding functional genes. More accurate specific formula for each pathway were developed to calculate the relative abundance of multiple pathways. The output

results consist of user-friendly tables containing a summary of over 100 biogeochemically-relevant pathways and corresponding genes, and their relative abundances in individual metagenomic/metatranscriptomic samples, alongside graphical outputs consisting of heatmaps and multiple sketch plots for easier visualization.

2 Methods

The main procedure of DiTing

DiTing was written in Python 3 and runs on Linux/Unix platforms. The pre-requisites required for running the software are described on the DiTing GitHub page (https://github.com/xuechunxu/DiTing). The input source was a set of metagenomic and/or metatranscriptomic clean reads where low-quality reads, primer and adaptor sequences had been trimmed beforehand (Fig. 1), which were then assembled by Megahit v1.1.2 (Li et al., 2016) (with default parameters) or metaSPAdes v3.12.0 (Nurk et al., 2017) (with default parameters). Users can set distinct parameter to choose which software for reads assembly. Compared to Megahit, MetaSPAdes performs better in recovering long contigs, it has a higher assembly quality index and is the recommended assembler for high-complex metagenomes (Forouzan et al., 2018, Pasolli et al., 2019). However, Megahit has a low error rate, is highly memory-efficient and is ideal for large datasets (Forouzan eta al., 2018). Genes were predicted and translated from the assembled contigs by Prodigal v2.6.3 with the '-p meta' option (Hyatt et al., 2010). To determine the relative abundance of each gene, the input metagenomic reads were mapped against predicted genes (nucleotide) by BWA-MEM (Li, 2013) (bwa v0.7.15, default settings) to generate sequence alignment map (SAM) files. We used the unsorted SAM files as input for pileup.sh (bbmap v38.22) (Bushnell, 2014) (with default parameters) to calculate the average coverage of each gene or transcript. The TPM methodology was used to indicate the relative abundance of a gene by the following formula.

$$TPM_i = \frac{b_i}{\sum_j b_j} \cdot 10^6 = \frac{\frac{X_i}{L_i}}{\sum_j \frac{X_j}{L_j}} \cdot 10^6$$

Where TPM_i is the relative abundance of gene i, b_i is the copy number of gene i, L_i is the length of gene i, X_i is the number of times that gene i is detected in a sample (that is, the number of reads in alignment), and j is the number of genes in a sample. The translated protein sequences were queried against KOfam database (HMM database of KEGG Orthologs; KOs) (Aramaki et al., 2019) using hmmsearch implemented within HMMER (Finn et al., 2011) (parameter: hmmsearch -T <threshold> --tblout <output> <hmm database> <input protein sequence> when score type is full; hmmsearch --domT <threshold> --domtblout <output> <hmm database> <input protein sequence> when score type is domain), which employs methods detecting remote homologs sensitively and efficiently. Kofam suggested values (ftp://ftp.genome.jp/pub/db/kofam/) were used as the cutoff threshold values for hmmsearch, in which each KEGG Orthology (KO) entry had its unique cutoff threshold values (Aramaki et al., 2019). To test the accuracy of the gene annotation from DiTing, we also submitted translated protein sequences to the KofamKOALA web server (https://www.genome.jp/tools/kofamkoala/). KofamKOALA assigns KOs numbers to protein sequences with its accuracy being comparable to the best existing KO assignment tools (Aramaki et al., 2019). For genes assigned into multiple KOs numbers, all the corresponding functions were associated to the genes. To specifically probe DMSP catabolism, 20 verified gene sequences (DMSP lyase genes *dddD*, *dddK*, *dddL*, *dddP*, *dddQ*, *dddY*, *dddW*, *Alma1*; DMSP synthesis genes *dsyB*, *DSYB*, *mmtN*; DMSP demethylation pathway genes *dmdA*, *dmdB*, *dmdC*, *dmdD*; acryloyl-CoA hydratase *acuH*, methanethiol *S*-methylase *mddA*, DMS monooxygenase *dmoA*, methanethiol oxidase *MTO*, and DMSO reductase *dorA*) were collected manually to create the profile HMM (Song et al., 2020). A table with the relative abundance and annotation of genes is used to estimate the relative abundance of approximately one

hundred biogeochemical pathways in each sample.

The formula for each pathway is specifically designed to estimate the relative abundance of the pathway according to the definitions (https://github.com/xuechunxu/DiTing/blob/master/Pathway_formulas.txt). For example, assimilatory sulfite reduction (ASR) that converts sulfite to sulfide has two known possible pathways: (1) Sir protein (K00392) mediated pathway (Gisselmann et al., 1993; Bork et al., 1998), and (2) CysJI protein (K00380 + K00381) mediated pathway (Ostrowski et al., 1989a, b; Zeghouf et al., 2000). Thus, the relative abundance of assimilatory sulfite reduction pathway is estimated by the following formula:

$$A_{ASR} = a_{K00392} + \frac{a_{K00380} + a_{K00381}}{2}$$

Where A_{ASR} is the relative abundance of the ASR pathway, a_{KO} is the relative abundance of KO in each sample. Dissimilatory nitrite reduction (DNRA), which converts nitrite to ammonia, can occur via two different enzymatic reactions: (1) *NirBD* proteins (K00362 + K00363) to convert nitrite to ammonia, or (2) *NrfAH* protein (K03385 + K15876) to convert nitrite to ammonia. Thus, the relative abundance of dissimilatory nitrite reduction to ammonia is estimated by the following formula:

$$A_{DNRA} = \frac{a_{K00362} + a_{K00363}}{2} + \frac{a_{K03385} + a_{K15876}}{2}$$

Where A_{DNRA} is the relative abundance of DNRA pathway, a_{KO} is the relative abundance of KO in each sample. For other pathways, a customized formula for each pathway was utilized (see Supplemental Table S1).

DiTing produces a table in the specified output directory. This table contains approximately 100 biogeochemical pathways and their relative abundance in each input sample. Another table of the relative abundances of corresponding KO/genes within these pathways in each sample is also generated (like Supplemental Table S2). Researchers can evaluate the completeness of pathways from this table. For improved

visualization, heatmaps and sketch plots for comparing the relative abundances of biogeochemical pathways in different samples are drawn by a Python script. DiTing can be installed via Conda (https://docs.conda.io).

Construction of the DMSP database and other selected genes

DMSP is a marine organosulfur compound with important roles in global sulfur cycle and may affect climate (Zhang *et al*., 2019), yet genes involved in the cycling of this compound are missing in currently available databases. Profile HMM were manually generated for eight pathways related to the cycling of DMSP (Song et al., 2020), including DMSP biosynthesis (methionine -> DMSP), DMSP demethylation (DMSP -> MMPA), DMSP demethylation (MMPA -> MeSH), DMSP cleavage (DMSP -> DMS), DMS oxidation (DMS -> MeSH), DMS oxidation (DMS -> DMSO), DMSO reduction (DMSO -> DMS), MddA pathway (MeSH -> DMS), MeSH oxidation (MeSH -> Formaldehyde). 20 verified gene sequences encoding key enzymes of these pathways were used to create the profile HMM (Song et al., 2020).

(i) DMSP biosynthesis (methionine -> DMSP). Three gene families participating in DMSP biosynthesis from methionine (Met), including DSYB, DsyB and MmtN are included in DiTing. DSYB and DsyB are methylthiohydroxbutryrate *S*-methyltransferase enzymes found in marine eukaryotes and prokaryotes, respectively (Curson et al., 2018; Curson et al., 2017). The MmtN Met *S*-methyltransferase is found in some Gram-positive bacteria, alpha- and gamma-proteobacteria (Liao et al., 2019; Williams et al., 2019). The cut-off E-values of DSYB, DsyB and MmtN are 1×10^{-30}, 1×10^{-67} and 1×10^{-98}, respectively.

(ii) DMSP demethylation (DMSP -> MMPA). The first step of DMSP demethylation pathway that results in the production of methylmercaptopropionate (MMPA) is initiated by the DmdA enzyme (Reisch et al., 2011). The cut-off E-values of the DmdA is 1×10^{-130}.

(iii) DMSP demethylation (MMPA -> MeSH). Further degradation of MMPA generating gaseous methanethiol (MeSH) catalyzed by the DmdBCD/AcuH enzymes

(Reisch et al., 2011; Shao et al., 2019). The cut-off E-values of DmdB, DmdC, DmdD and AcuH are 1×10^{-75}, 1×10^{-100}, 1×10^{-30} and 1×10^{-56}, respectively.

(iv) DMSP cleavage (DMSP -> DMS). Eight distinct DMSP lyase enzymes (DddD, DddK, DddL, DddP, DddQ, DddW, DddY and Alma1) can cleave DMSP to generate dimethylsulfide (DMS) (Curson et al., 2011; Alcolombri et al., 2015; Johnston et al., 2016; Sun et al., 2016). The cut-off E-values of DddD, DddK, DddL, DddP, DddQ, DddW, DddY and Alma1 are 1×10^{-97}, 1×10^{-35}, 1×10^{-33}, 1×10^{-83}, 1×10^{-20}, 1×10^{-49}, 1×10^{-64} and 1×10^{-26}, respectively.

(v) DMS oxidation (DMS -> MeSH). DMS can be oxidized to generate MeSH via the DMS monooxygenase enzyme DmoA (Boden et al., 2011). The cut-off E-values of the DmoA is 1×10^{-34}.

(vi) DMS oxidation (DMS -> DMSO). DMS can be oxidized to generate dimethylsulfoxide (DMSO) by the DMS dehydrogenase complex (DdhABC) (McDevitt et al., 2002) or trimethylamine monooxygenase (Tmm); (Lidbury et al., 2016). The cut-off E-values of both DdhABC, DdhB and Tmm are 1×10^{-30}.

(vii) MddA pathway (MeSH -> DMS). MeSH can be *S*-methylated to generate DMS by the MddA enzyme (Carrión et al., 2017). The cut-off E-values of MddA is 1×10^{-30}.

(viii) MeSH oxidation (MeSH -> Formaldehyde). MeSH can also be modified through another pathway catalyzed by the MeSH oxidase MTO (Eyice et al., 2018). The cut-off E-values of MTO is 1×10^{-20}.

The sugar 6-deoxy-6-sulfoglucose (sulfoquinovose, SQ) produced by plants, algae, and cyanobacteria, is an important component of carbon and sulfur cycles (Frommeyer et al., 2020). Microbial community can completely degrade SQ into inorganic sulfate or hydrogen sulfide through three pathways, i.e., sulfo-Embden-Meyerhof-Parnas (sulfo-EMP) (Denger et al., 2014), sulfo-Entner-Doudoroff (sulfo-ED) (Felux et al., 2015), and 6-deoxy-6-sulfofructose-transaldolase (SFT) pathways

(Frommeyer et al., 2020).

(i) sulfo-EMP pathway. SQ is converted to 6-deoxy-6-sulfofructose (SF) through an aldose/ketose isomerase YihS. The SF is phosphorylated to 6-deoxy6-sulfofructosephosphate (SFP) by an ATP-dependent SF kinase YihV. The SFP is then cleaved into 3-sulfolactaldehyde (SLA) and dihydroxyacetone phosphate (DHAP) by an SFP aldolase YihT. Finally, the SLA is reduced via an NADH-dependent SLA reductase (YihU) to DHPS, which is excreted from microorganisms. These four genes *YihSVTU* were annotated through K18479, K18478, K01671 and K08318 Orthology in KEGG, respectively.

(ii) sulfo-ED pathway. This pathway starts with an NAD^+-dependent SQ dehydrogenase (EC:1.1.1.390) oxidizing SQ to 6-sulfogluconolactone (SGL). The SGL is hydrolyzed to 6-deoxy-6-sulfogluconate (SG) by an SGL lactonase (EC:3.1.1.99). The SG is then converted by an SG dehydratase (EC:4.2.1.162) to 2-keto-3,6-deoxy-6-sulfo-gluconate (KDSG). The KDSG is cleaved by a KDSG aldolase (EC:4.1.2.58) into pyruvate and 3-sulfolactaldehyde (SLA). The SLA can be oxidized by a NAD^+-dependent SLA dehydrogenase (EC:1.2.1.97) to SL. The reference sequences of these enzymes were collected manually from Uniprot database (https://www.uniprot.org/).

(iii) SFT pathway. Three key enzymes take part in this pathway. The SQ is converted to SF by an aldose/ketose isomerase, which is the same enzyme as the first step of sulfo-EMP pathway. SF is cleaved to 3-sulfolactaldehyde (SLA) by SF transaldolase enzyme. Finally, The SLA is oxidized by a NAD^+-dependent SLA dehydrogenase to SL. The SLA dehydrogenase is same enzyme as the last step of sulfo-ED pathway. The reference sequence of SF transaldolase enzyme was collected from IMG (https://img.jgi.doe.gov/) according to Frommeyer et al., 2020.

Isoprene (2-methyl-1, 3-butadiene) is an important volatile organic compound emitted to the atmosphere, and has significant effect on the climate (Carrión et al., 2018). Isoprene can be degraded by microbial communities with the isoprene

monooxygenase (IsoMO). The gene *isoA* encoding the α-subunit of IsoMO was selected as marker gene for distribution, diversity and abundance of isoprene-degrading pathway in environment (Carrión et al., 2018; Carrión et al., 2020). The reference sequences of IsoA enzyme were collected manually from NCBI according to Carrión et al., 2018.

3 Results and discussion

General information of DiTing

We developed a new metagenomics/metatranscriptomic analysis pipeline, DiTing, to infer and compare the prevalence of genes and pathways of key biogeochemical cycles. DiTing consists of four main features: (i) automated assembly, CDS prediction, mapping and annotation from reads; (ii) a manually created dimethylsulfoniopropionate (DMSP) cycling related gene database; (iii) accurate and specific formula for DMSP and other biogeochemical pathway to calculate the relative abundance of biogeochemically-relevant pathways and genes; (iv) visualization of results comparing biogeochemical cycling potential between different input samples. These features make DiTing a flexible and versatile tool wrapper for studying biogeochemical cycles, or just as a platform to tackle metagenomic shotgun sequencing data. The speed of DiTing is relatively fast. Five samples (from the hydrothermal vent case study below) that are about 500 Gb in total were used to evaluate the speed. The total run time for all analyses from reads to visualization was ~ 33 hours using 60 CPU threads on a Linux version 4.15.0-20-generic server (Ubuntu 18.04; CPU, Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz; RAM, 256 GB).

Accuracy testing of DiTing using simulated benchmark datasets

To verify the accuracy of DiTing in evaluating the abundance of biogeochemical pathways, CAMISIM (Fritz et al., 2019) was used to simulate three group of metagenomic shotgun sequenced samples (photoautotrophs, chemoautotrophs and heterotrophs group). Metagenomes from the photoautotrophic group were simulated by ten *Cyanobacteria* genomes. Metagenomes from the chemoautotrophic group were

simulated by 10 ammonia-oxidasing archaea (AOA) genomes. Metagenomes from the heterotrophic group were simulated by 10 SAR11 genomes. Each group comprised five metagenomic samples sequenced by Illumina 2 × 150 bp paired-end reads, and each generated sample had a size of 5 Gb. These 15 simulated samples were fed into DiTing. The overall relative abundance of biogeochemical pathways in simulated samples was consistent with features of genomes used in each group (Fig. 2). For example, metagenomes in the photoautotroph group possessed a high relative abundance of photosynthesis related pathway genes (photosystem I, II and cytochrome b_6/f complex), which were absent in other two groups (Fig. 2). AOA are the typical known bacterial ammonia oxidisers, which possesses *amoABC* genes encoding the ammonia monooxygenase complex. Correspondingly, in the chemoautotroph group simulated by AOA, the ammonia oxidation pathway was found but was absent in other two groups analysed by DiTing (Fig. 2). In other nitrogen cycle pathways, *nirKS* encoding nitrite reductase and *hzs* encoding hydrazine synthase were only seen the chemoautotroph group of the DiTing results. Consistently, these genes were annotated in ammonia-oxidasing archaea genomes through RAST annotation manually. Additionally, bacteria and archaea use F-type ATPase and V/A-type ATPases, respectively, to hydrolyze ATP to ADP, respectively (Pisa et al., 2007; Fillingame et al., 1997). Thus, F-type ATPase was detected in groups simulated by *Cyanobacteria* and SAR11 genomes, and V/A-type ATPase was only detected in the chemoautotroph group simulated by ammonia-oxidasing archaea genomes. The translated gene sequences (amino acid) from simulated metagenomes were submitted to the KofamKOALA web server for annotation. The gene annotation results derived from DiTing were the same as those from KofamKOALA web server, verifying the accuracy of gene annotation.

Application of DiTing on five real hydrothermal vent datasets and 15 *Tara* Ocean project datasets

DiTing was used to analyze the biogeochemical potential of five marine metagenomic samples (Table 1; NCBI accession number: ERR1679394-1679398) generated from

hydrothermal vent samples taken at PACManus and North Su fields in the Manus Basin (Meier et al., 2017; Table 1). The metagenomic clean reads ranged in size from 81 to 112 Gbp from each sample. The reads were assembled into 799,269 to 1,182,847 contigs with the total assembly sizes ranging from 0.58 to 1.00 Gbp. A total of 5,639,558 Open Reading Frames (ORFs) within these contigs were then predicted. ~18.9% (1,065,097) ORFs were annotated against KEGG databases and affiliated to 8128 KO entries. The relative abundances of ~100 biogeochemically-relevant pathways were calculated according to our new formulas (Supplementary Table S1). The relative abundance of genes within these pathways was also prepared for further analyses at the gene level (Supplementary Table S2). The summary sketch and heatmap plots for visualization of these pathways were generated, and these reflect the different patterns of community function within metagenomic samples (Fig. 3, 4).

Of the five metagenomes collected in diffuse hydrothermal vent fluids, NSu-F2b and NSu-F5 originated from acidic samples with sulfide (1.6 mmol l^{-1} and 0.7 mmol l^{-1} H_2S, respectively) and methane (0.2 mmol l^{-1} and 0.01 mmol l^{-1} CH_4, respectively) levels detected. The Fw-F1b, Fw-F3 and RR-F1b metagenomes originated from sites with no detectable H_2S and CH_4. Reassuringly, the NSu-F2b and NSu-F5 samples, with similar environmental parameters, showed the most similar distribution patterns for genes and pathways involved in the cycling of nitrogen, carbon and sulfur (Fig. 3, 4). Indeed, hierarchical clustering of samples according to their microbial function composition showed NSu-F2b and NSu-F5 fall into one cluster and the other three samples into another cluster (Supplementary Fig. S1).

At hydrothermal vents, chemolithoautotrophic microorganisms carry out carbon fixation coupled with oxidation of reduced sulfur compounds (Meier et al., 2017). In accordance, we found the relative abundance of thiosulfate oxidation, sulfite oxidation, and first step of dissimilatory sulfate reduction pathways (reversible conversion of sulfate to sulfite) to be more highly represented compared to other sulfur cycle pathways in all five samples (Fig. 3, 4), indicating sulfate reduction and sulfur oxidation as major processes in microbial sulfur cycling. This finding is

supported by the presence of sulfate-reducing *Nitrospirae* and sulfur-oxidizing *Gammaproteobacteria* dominating microbial communities at these hydrothermal vents (Meier *et al*., 2017, 2019). In addition, assimilatory sulfate reduction and thiosulfate disproportionation pathways were almost only found in NSu-F2b and NSu-F5 (Fig. 3), the only samples with detectable sulfide levels, indicating microbes in these samples may incorporate sulfide into the amino acids cysteine (Cys) or homo-Cys. Here, the relative abundance of thiosulfate disproportionation was estimated by dividing the sum of relative abundance of *phsABC* by the number (n = 3) of essential subunits. The relative abundances of each subunit of thiosulfate reductase were often not equal to each other in the metagenomes (Supplementary Table S2). For example, *phsA* (encoding thiosulfate reductase subunit A) was always far more abundant than *phsC* (thiosulfate reductase cytochrome B subunit) and *phsB* (thiosulfate reductase electron transport protein) was not detected in any sample. This may be due to insufficient sequencing depth and/or protein redundancy. Whatever the reason for these discrepancies it cannot be easily solved by bioinformatics alone and culture-dependent work is necessary. This phenomenon highlighted for the thiosulfate disproportionation genes may also occur in other pathways, thus further analyses at the gene level, not only at the pathway level, are essential in predicting the biogeochemical potential of microbial communities after DiTing analysis.

In previously tested seawater and sediment samples, known DMSP synthesis genes are always much less abundant than those for its catabolism (Curson *et al* 2017, Curson *et al* 2018, Williams *et al*., 2019). This was not the case in previously studied hydrothermal samples (Song et al., 2020), with the DMSP lyase gene *dddP* being the only detected DMSP catabolic gene. In three out of five hydrothermal samples interrogated here, the genetic potential to synthesize DMSP, through prokaryotic *dsyB* and *mmtN* genes, is far less than that for DMSP catabolism (DMSP synthesis:DMSP catabolism = 1:16.9) and not so dissimilar to ratios seen in seawater samples (Curson *et al* 2017, Curson *et al.,* 2018, Williams *et al*., 2019). Reasons for this discrepancy between the distinct samples are unknown. The DsyB sequences retrieved from this

data were clustered with ratified DsyB proteins, not with DSYB and non-functional DsyB-like proteins from *Streptomyces varsoviensi*s, which support their function in DMSP synthesis (Supplementary Fig. S2). Interestingly, sample NSu-F2b has higher DMSP synthesis potential than any other samples due to relatively high levels of bacteria with *mmtN*. As in Song et al 2020, the potential for DMSP cleavage was more prominent than for DMSP demethylation (*dmdA*) in all hydrothermal samples, although catabolism of MMPA, the initial product of DMSP demethylation by DmdA (Howard *et al* 2006), was very abundant. This data supports DMSP cleavage being the dominant DMSP catabolic pathway in hydrothermal sediments, as proposed in Song et al., 2020. Alternatively, there could be novel DMSP demethylase enzymes. This would explain why there were such low *dmdA* levels in hydrothermal sediment, yet very high MMPA degradation potential. The potential for oxidation and reduction of DMSP catabolites, DMS and methanethiol, was similar to that described in Song et al., 2020, with sites NSU-F2b and F5 showing the greatest potential. Thus, some interesting predictions of DMSP cycling were enabled by DiTing analysis on the metagenomes analyzed here. Note any predictions made from genetic potential alone require further investigation regarding function and expression and, importantly, substantiation for synthesis and turnover rate analysis.

The samples NSu-F2b and NSu-F5 had lower oxygen concentration than Fw-F1b, Fw-F3 and RR-F1b samples, especially NSu-F2b (0.07 and 0.14 mmol l^{-1} for NSu-F2b and NSu-F5, respectively; 0.17 - 0.2 mmol l^{-1} for other three). Indeed, compared to the other three samples, NSu-F2b and NSu-F5 had significantly more genes encoding *bd* ubiquinol cytochrome oxidases ($p < 0.01$) that are associated with low oxygen concentrations (Fig. 4). It is worth noting that the *bd* oxidase was enriched most in NSu-F2b under the highest sulfide concentration (1.6 mmol l^{-1}) and lowest oxygen concentration. A previous study found that *bd* oxidase could promote sulfide-resistant O_2 consumption and growth in *E. coli* (Forte et al., 2016), implying the important role of *bd* oxidases in the low oxygen NSu-F2b environment.

The NSu-F2b and NSu-F5 samples showed enrichment for denitrification,

nitrification and nitrogen fixation potential, which may be due to the lower oxygen levels of these samples or is possibly reflecting the nitrogen availability at higher temperatures. Notably, in NSu-F5, genes encoding for the denitrification enzymes responsible for reduction of the cytotoxic gaseous intermediates, nitric oxide (NO), *norBC*, and nitrous oxide (N_2O), *nosZ*, are significantly enriched, alongside the nitrifying genes responsible for aerobic conversion of nitrite to nitrate (*nxrAB*). The importance of nitrification and denitrification to nitrogen cycling of hydrothermal vents has previously been reported (Bourbonnais et al., 2012), but not at the resolution allowed by DiTing. The transcriptional and enzymatic activity of these systems at these pH levels would certainly need experimental validation. These metagenomes highlight metabolic importance of nitrogen cycling with the potential for all other pathways being at similarly high levels (Supplementary Table S2) in all samples, with the exception of nitrite assimilation (nitrite to ammonia) and hydroxylamine oxidation to nitrite (*hao*) was not detected. Again, this may reflect nitrogen availability but is also indicative of nitrogen source preference of the microbiomes under the highly reactive physicochemical constraints of the vent environment. This study illustrates the need for comprehensive measurements of nitrogen flux, metatranscriptional analyses to ascertain the most active pathways and identification of the dominant organisms responsible for nitrogen cycling in these ecosystems. Overall, these results highlight potential microbial metabolic differences in communities from different hydrothermal samples that likely reflect changes in environmental conditions.

DiTing was also applied to analyze 15 metagenomic samples from chlorophyll *a* (Chl*a*) maximum layer in Mediterranean Sea from *Tara* Ocean project. The metagenomic clean reads ranged in size from 1.24 to 52.53 Gbp from each sample. The reads were assembled into 71,183 to 1,601,956 contigs with the total assembly sizes ranging from 0.045 to 1.38 Gbp. A total of 18,431,131 ORFs within these contigs were then predicted. ~24% (1,065,097) ORFs were annotated against KEGG databases and affiliated to 8759 KO entries. The 74 pathways related biogeochemical cycles were found (Supplementary Table S3). Compared to the sample derived

hydrothermal vents, the Chl*a* maximum layer contains remarkable high relative abundance of photosystem pathway as expected (Supplementary Table S3 and S4). Additionally, eukaryotic DMSP synthesis gene, *DSYB* was detected in 10 out of 15 Chl*a* maximum samples, which were absent in the hydrothermal vent samples. The relative abundance of *DSYB* was comparable to that of prokaryotic DMSP synthesis gene *dsyB* in Chl*a* maximum layers (Supplementary S4), indicating that the DMSP was produced by both prokaryotes and eukaryotes in these environments. For DMSP degradation, in six out of 15 samples, the genetic potential to DMSP demethylation, through the *dmdA* gene, was higher than that for DMSP cleavage (*ddds* and *alma1*) (DMSP demethylation:DMSP cleavage = 1.69:1). This is contrasted with the hydrothermal vent samples. In other nine samples, the potential for DMSP demethylation was comparable to that for DMSP cleavage (DMSP demethylation:DMSP cleavage = 0.82:1). These data support both DMSP demethylation and cleavage being the dominant DMSP catabolic pathways in the Chl*a* maximum layer.

Conclusion

In summary, this study developed a pipeline (DiTing) to infer and compare biogeochemical pathways from metagenomic and metatranscriptomic data. DiTing is a portable tool for metagenomic and metatranscriptomic datasets, providing automatic, multi-threaded bioinformatic workflows for data handling, including read assembly, ORF prediction, annotation, and more accurate specific formulas for calculating the relative abundance of biogeochemical pathways. The visualization module is designed to more easily compare functions between samples via graphical outputs. In addition, a verified database was built manually for the annotation of genes involved in the production and cycling of DMSP. As validation of the outputs produced by DiTing, comparisons of the relative abundance of biogeochemical pathways in published metagenomes and metatranscriptomes to those calculated by DiTing were consistent. By applying DiTing to analyze five hydrothermal shotgun metagenomes, we showed that the functional profile could accurately reflect changes

in environmental conditions (H_2S and O_2 concentrations). DiTing can be readily applied to metagenomic and/or metatranscriptomic studies, with relatively straightforward user intervention. This bioinformatics framework will facilitate our understanding of spatial and temporal changes in microbiome-mediated biogeochemical cycles.

Acknowledgements

The authors thank Saiyi Zhu from Shenzhen Nanshan Foreign Language School for logo designing.

Funding

This work was supported by Marine S & T Fund of Shandong Province for Pilot National Laboratory for Marine Science and Technology (Qingdao) (No. 2018SDKJ0406-4), the National Key Research and Development Program of China (No. 2018YFE0124100), and the National Natural Science Foundation of China (Nos. 41730530 and 91751202) in X-HZ's laboratory, and Natural Environmental Research Council standard, UK, standard grants (NE/N002385, NE/P012671, and NE/S001352) in JDT's laboratory.

Conflicts of Interest: none declared.

References

Acinas, S.G., Sánchez, P., Salazar, G., Cornejo-Castillo, F.M., Sebastián, M., Logares, R., … Gasol, J.M. (2019) Metabolic architecture of the deep ocean microbiome. *bioRxiv*, 635-680.

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., & Narasimhan, G. (2016) Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis. *Evolutionary Bioinformatics*, 12, 5-16.

Alcolombri, U., Ben-Dor, S., Feldmesser, E., Levin, Y., Tawfik, D.S., & Vardi, A. (2015) Identification of the algal dimethyl sulfide–releasing enzyme: a missing link in the marine sulfur cycle. *Science*, 348, 1466–1469.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2019) KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, doi: 10.1093/bioinformatics/btz859.

510 Boden, R., Borodina, E., Wood, A. P., Kelly, D. P., Murrell, J. C., & Schäfer, H. (2011) Purification
511 and characterization of dimethylsulfide monooxygenase from *Hyphomicrobium sulfonivorans*.
512 *Journal of Bacteriology*, 193, 1250–1258.

513 Bokulich, N.A., Subramanian S., Faith J.J., Gevers, D., Gordon J.I., Knight, R., Mills, D.A., &
514 Caporaso, J.G. (2013) Quality-filtering vastly improves diversity estimates from Illumina
515 amplicon sequencing. *Nature Method*, 10, 57-59.

516 Bork, C., Schwenn, J.D., & Hell, R. (1998) Isolation and characterization of a gene for assimilatory
517 sulfite reductase from *Arabidopsis thaliana*. *Gene*, 212, 147-153.

518 Bourbonnais, A., Lehmann, M.F., Butterfield, D.A., & Juniper, S.K. (2012) Subseafloor nitrogen
519 transformations in diffuse hydrothermal vent fluids of the Juan de Fuca Ridge evidenced by the
520 isotopic composition of nitrate and ammonium. *Geochemistry Geophysics Geosystems*, 13, 1-23.

521 Bushnell, B. (2014) BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab.
522 (LBNL), Berkeley, CA (United States). https://www.osti.gov/biblio/1241166

523 Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., … Karp, P.D. (2006)
524 MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*,
525 34, D511-D516.

526 Carrión, O., Larke-Mejía, N.L., Gibson, L., Haque, M.F.U., Ramiro-García, J., McGenity, T.J., &
527 Murrell, J.C. (2018) Gene probing reveals the widespread distribution, diversity and abundance of
528 isoprene-degrading bacteria in the environment. *Microbiome*, 6, 219.

529 Carrión, O., McGenity, T. J., & Murrell, J. C. (2020) Molecular ecology of isoprene-degrading
530 bacteria. *Microorganisms*, 8, 967.

531 Carrión, O., Pratscher, J., Curson, A. R., Williams, B. T., Rostant, W. G., Murrell, J. C., & Todd, J.D.
532 (2017) Methanethiol-dependent dimethylsulfide production in soil environments. *The ISME*
533 *Journal*, 11, 2379–2390.

534 Curson, A.R., Todd, J.D., Sullivan, M.J., & Johnston, A.W.B. (2011) Catabolism of
535 dimethylsulphoniopropionate: microorganisms, enzymes and genes. *Nature Reviews*
536 *Microbiology*, 9, 849–859.

537 Curson, A.R., Liu, J., Martínez, A.B., Green, R.T., Chan, Y., Carrión, O., … Todd, J.D. (2017).
538 Dimethylsulfoniopropionate biosynthesis in marine bacteria and identification of the key gene in
539 this process. *Nature Microbiology*, 2, 17009. doi:10.1038/nmicrobiol.2017.9

540 Curson, A.R., Williams, B.T., Pinchbeck, B.J., Sims, L.P., Martínez, A.B., Rivera, P.P.L., … Todd,
541 J.D. (2018). DSYB catalyses the key step of dimethylsulfoniopropionate biosynthesis in many
542 phytoplankton. *Nature Microbiology*, 3, 430–439. doi:10.1038/s41564-018-0119-5.

543 Denger, K., Weiss, M., Felux, A.K., Schneider, A., Mayer, C., Spiteller, D., … Schleheck, D. (2014)
544 Sulphoglycolysis in *Escherichia coli* K-12 closes a gap in the biogeochemical sulphur cycle.

545 *Nature* 507, 114–117.

546 Dombrowski, N., Teske, A.P., & Baker, B.J. (2018) Expansive microbial metabolic versatility and
547 biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nature Communications*, 9, 1-
548 13.

549 Eyice, O., Myronova, N., Pol, A., Carrión, O., Todd, J.D., Smith, T.J., … Schäfer, H. (2018) Bacterial
550 SBP56 identified as a Cu-dependent methanethiol oxidase widely distributed in the biosphere. *The*
551 *ISME Journal*, 12, 145–160.

552 Felux, A-K., Spiteller, D., Klebensberger, J., & Schleheck, D. (2015) Entner–Doudoroff pathway for
553 sulfoquinovose degradation in *Pseudomonas putida* SQ1. *Proceedings of the National Academy of*
554 *Sciences*, 112, 4298-4305.

555 Fillingame, R.H. (1997) Coupling H+ transport and ATP synthesis in F1F0-ATP synthases: glimpses
556 of interacting parts in a dynamic molecular machine. *Journal of Experimental Biology*, 200, 217-
557 224.

558 Finn, R.D., Clements, J., & Eddy, S.R. (2011) HMMER web server: interactive sequence similarity
559 searching. *Nucleic Acids Research*, 39, D29-D37.

560 Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., … Bateman, A. (2014) Pfam:
561 the protein families database. *Nucleic Acids Research*, 42, D222-D230.

562 Firtz, A., Hofmann, P., Majda, S., Dahms, E., Dröge, J., Fiedler, J., … McHardy, A.C. (2019).
563 CAMISIM: simulating metagenomes and microbial communities. *Microbiome*, **7**, 17.

564 Forouzan, E., Shariati, P., Maleki, M.S.M., Karkhane, A.A., & Yakhchali, B. (2018) Practical
565 evaluation of 11 de novo assemblers in metagenome assembly. *Journal of Microbiological*
566 *Methods*, 151, 99-105.

567 Fortunato, C.S., Larson, B., Butterfield, D.A., & Huber, J.A. (2018) Spatially distinct, temporally
568 stable microbial populations mediate biogeochemical cycling at and below the seafloor in
569 hydrothermal vent fluids. *Environmental Microbiology*, 20, 769-784.

570 Friedrich, C.G. (1998) Physiology and genetics of sulfur-oxidizing bacteria. *Advances in Microbial*
571 *Physiology*, 39, 235-289.

572 Frommeyer, B., Fiedler, A.W., Oehler, S.R., Hanson, B.T., Loy, A., Franchini, P., … Schleheck, D.
573 (2020) Environmental and intestinal phylum Firmicutes bacteria metabolize the plant sugar
574 sulfoquinovose via a 6-deoxy-6-sulfofructose transaldolase pathway. *iScience*, 23, 101510.

575 Ganesh, S., Parris, D.J., DeLong, E.F., & Stewart, F.J. (2014) Metagenomic analysis of size-
576 fractionated picoplankton in a marine oxygen minimum zone. *The ISME Journal*, 8, 187-211.

577 Gisselmann, G., Klausmeier, P., & Schwenn, J.D. (1993) The ferredoxin: sulphite reductase gene from
578 *Synechococcus* PCC7942. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1144, 102-106.

579 Graham, E.D., Heidelberg, J.F., & Tully, B.J. (2018) Potential for primary productivity in a globally-
580 distributed bacterial phototroph. *The ISME Journal*, 350, 1–6.

581 Howard, E.C., Henriksen, J.R., Buchan, A., Reisch, C.R., Bürgmann, H., Welsh, R., … Moran, M.A.
582 (2006). Bacterial taxa that limit sulfur flux from the ocean. *Science*, 314, 649–652.
583 doi:10.1126/science.1130657

584 Huertacepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., … Bork, P. (2016)
585 eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for
586 eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, 44, D286-D293.

587 Hyatt, D., Chen, G., Locascio, P.F., Land, M.L., Larimer, F.W., & Hauser, L.J. (2010) Prodigal:
588 prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11,
589 119.

590 Johnston, A.W., Green, R.T., & Todd, J.D. (2016) Enzymatic breakage of dimethylsulfoniopropionate
591 —a signature molecule for life at sea. *Current Opinion Chemical Biology*, 31, 58–65.

592 Jones, M.R., & Good, J.M. (2016) Targeted capture in evolutionary and ecological genomics.
593 *Molecular. Ecology*, 25, 185-202.

594 Li, D., Luo R., Liu, C-M., Leung, C-M, Ting, H-F., Sadakane, K., … Lam, T-W. (2016) MEGAHIT
595 v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and
596 community practices. *Methods*, 102, 3-11.

597 Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*
598 *preprint arXiv:*1303.3997.

599 Liao, C., & Seebeck, F.P. (2019) In vitro reconstitution of bacterial DMSP biosynthesis. *Angewandte*
600 *Chemie International Edition*, 58, 3553-3556.

601 Lidbury, I., Kröber, E., Zhang, Z., Zhu, Y., Murrell, J. C., Chen, Y., & Schäfer, H. (2016) A
602 mechanism for bacterial transformations of DMS to DMSO: a missing link in the marine organic
603 sulfur cycle. *Environmental Microbiology*. 18, 2754–2765.

604 Llorens-Marès, T., Yooseph, S., Goll, J., Hoffman, J., Vila-Costa, M., Borrego, C.M., … Casamayor,
605 E. O. (2015) Connecting biodiversity and potential functional role in modern euxinic
606 environments by microbial metagenomics. *The ISME Jounal*, 9, 1648-1661.

607 McDevitt, C.A., Hanson, G.R., Noble, C.J., Cheesman, M.R., & McEwan, A.G. (2002)
608 Characterization of the redox centers in dimethyl sulfide dehydrogenase from *Rhodovulum*
609 *sulfidophilum*. *Biochemistry*, 41, 15234–15244.

610 Meier, D.V., Pjevac, P., Bach, W., Hourdez, S., Girguis, P.R., Vidoudez, C., … Meyerdierks, A. (2017)
611 Niche partitioning of diverse sulfur-oxidizing bacteria at hydrothermal vents. *The ISME Jounal*,
612 11, 1545-1558.

Meier, D.V., Pjevac, P., Bach, W., Markert, S., Schweder, T., Jamieson, J., … Meyerdierks, A. (2019) Microbial metal-sulfide oxidation in inactive hydrothermal vent chimneys suggested by metagenomic and metaproteomic analyses. *Environmental Microbiology*, 21, 682-701.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, 27, 824-834.

Ogata, H., & Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27, 29-34.

Ostrowski, J., Barber, M.J., Rueger, D.C., Miller, B.E., Siegel, L.M., & Kredich, N.M. (1989a) Characterization of the flavoprotein moieties of NADPH-sulfite reductase from *Salmonella typhimurium* and *Escherichia coli*. Physicochemical and catalytic properties, amino acid sequence deduced from DNA sequence of *cysJ*, and comparison with NADPH-cytochrome P-450 reductase. *Journal of Biological Chemistry*, 264, 15796-15808.

Ostrowski, J., Wu, J-Y., Rueger, D.C., Miller, B.E., Siegel, L.M., & Kredich, N.M. (1989b) Characterization of the *cysJIH* regions of *Salmonella typhimurium* and *Escherichia coli* B. DNA sequences of *cysI* and *cysH* and a model for the siroheme-Fe4S4 active center of sulfite reductase hemoprotein based on amino acid homology with spinach nitrite reductase. *Journal of Biological Chemistry*., 264, 15726-15737.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P-A., Woodcroft, B.J., Evans, P.N., … Tyson, G.W. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2, 1533-1542.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., … Segata, N. (2019) Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 176, 649-662.

Petter, T., Lundin, D., Plathan, J., Poole, A.M., Sjöberg, B-M., & Sjöling, S. (2013) A metagenomics transect into the deepest point of the Baltic Sea reveals clear stratification of microbial functional capacities. *Plos One*., 8, e74983.

Pisa, K.Y., Huber, H., Thomm, M., & Muller, V. (2009) A sodium ion-dependent A1AO ATP synthase from the hyperthermophilic archaeon *Pyrococcus furiosus*. *The FEBS Journal*, 274, 3928-3938.

Prestat, E., David, M.M., Hultman, J., Taş N., Lamendella, R., Dvornik, J., … Jansson, J.K. (2014) FOAM (Functional Ontology Assignments for Metagenomes): a Hidden Markov Model (HMM) database with environmental focus. *Nucleic Acids Research*, 42, e145.

Reisch, C.R., Moran, M.A., & Whitman, W.B. (2011) Bacterial catabolism of dimethylsulfoniopropionate (DMSP). *Frontier in Microbiology*, 2, 172.

Reisch, C.R., Stoudemayer, M.J., Varaljay, V.A., Amster, I.J., Moran, M.A., & Whitman, W.B. (2011) Novel pathway for assimilation of dimethylsulphoniopropionate widespread in marine bacteria.

Nature, 473, 208–211.

Riesenfeld, C.S., Schloss, P.D., & Handelsman, J. (2004) Metagenomics: genomic analysis of microbial communities. *Annual Reviews of Genetics*, 38, 525-552.

Ross, O., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., … Stevens, R. (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, 42, D206-D214.

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, D260-D264.

Shakya, M., Lo, C-C., & Chain, P.S.G. (2019) Advances and challenges in metatranscriptomic analysis. *Frontier in Microbiol*ogy, 10, 904.

Shao, X., Cao, H-Y., Zhao, F., Ming, P., Wang, P., and Li, C-Y., … Zhang, Y-Z. (2019) Mechanistic insight into 3-methylmercaptopropionate metabolism and kinetical regulation of demethylation pathway in marine dimethylsulfoniopropionate-catabolizing bacteria. *Molecular Microbiology*, 111, 1057–1073.

Sharpton, T.J. (2014) An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5, 00209.

Smedile, F., Messina, E., Cono, V.L., Tsoy, O., Monticelli, L.S., … Yakimov, M.M. (2013) Metagenomic analysis of hadopelagic microbial assemblages thriving at the deepest part of Mediterranean Sea, Matapan-Vavilov Deep. *Environmental Microbiology*, 15, 167-182.

Song, D., Zhang, Y., Liu, J., Zhong, H., Zheng, Y., Zhou, S., … Zhang, X-H. (2020) Metagenomic insights into the cycling of dimethylsulfoniopropionate and related molecules in the Eastern China Marginal seas. *Frontiers in Microbiology*, 11, 157.

Sun, J., Todd, J.D., Thrash, J.C., Qian, Y., Qian, M.C., Temperton, B., … Giovannoni, S.J. (2016) The abundant marine bacterium *Pelagibacter* simultaneously catabolizes dimethylsulfoniopropionate to the gases dimethyl sulfide and methanethiol. *Nature Microbiology*, 1, e16065.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., & Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28, 33-36.

Williams, B.T., Cowles, K., Martínez, A.B., Curson, A.R.J., Zheng Y., Liu, J., … Todd, J.D. (2019) Bacteria are important dimethylsulfoniopropionate producers in coastal sediments. *Nature Microbiology*, 4, 1815-1825.

Xue, C-X., Liu, J., Lea-Smith, D.J., Rowley, G., Lin, H., Zheng, Y., … Zhang, X-H. (2020a) Insights into the vertical stratification of microbial ecological roles across the deepest seawater column on Earth. *Microorganisms*, 8, 1309.

Xue, C-X. Zhang, H., Lin, H., Sun, Y., Luo, D., Huang, Y., … Luo, H. (2020b) Ancestral niche separation and evolutionary rate differentiation between sister marine Flavobacteria lineages. *Environmental Microbiology*, 22, 3234-3247.

Zeghouf, M., Fontecave, M., & Coves, J. (2000) A simplifed functional version of the *Escherichia coli* sulfite reductase. *Journal of Biological Chemistry*, 275, 37651-37656.

Zhang, X., Xu, W., Liu, Y., Cai, M., Luo, Z., & Li, M. (2018) Metagenomics reveals microbial diversity and metabolic potentials of seawater and surface sediment from a hadal biosphere at the Yap Trench. *Frontiers in Microbiology*, 9, 2402.

Zhang, X.H., Liu, J., Liu, J.L., Yang, G., Xue, C-X., Curson, A.R.J., & Todd, J.D. (2019). Biogenic production of DMSP and its degradation to DMS—their roles in the global sulfur cycle. *Science China Life Sci*ence, 62, 1296–1319.

Fig. 1. A flowchart of the major steps involved in running DiTing. First (**A**), clean reads of metagenomes or/metatranscriptomes are assembled, annotated and mapped. Second (**B**), a table for relative abundances of KO number in KEGG among samples is constructed and relative abundances of biogeochemical pathways are estimated according to unbiased specific formulas. Third (**C**), heatmap and sketch plots are drawn to aid visualization.

Fig. 2. Bubble plots depicting the DiTing result of the relative abundance of pathways in simulated metagenomes. Photoautotroph group contains sample1-5 that simulated by *Cyanobacteria* genomes. Chemoautotroph group contains sample6-10 that simulated by ammonia-oxidasing archaea genomes. Heterotroph group contains sample11-15 that simulated by SAR11 genomes.

Fig. 3. Pie charts representing the relative abundance of carbon (**A**), nitrogen (**B**), sulfur (**C**) and DMSP (**D**) cycle related pathways for five metagenomic samples from the Manus Basin. Normalized relative abundance was calculated through dividing the relative abundance of a pathway in an individual sample by the sum of this pathway's relative abundance in all samples. Pie chart area reflects the relative abundance of the

process according to the scale shown in pink. The dashed line in panel D means the data was not shown. (**A**) CBB, Calvin-Benson-Bassham cycle; rTCA, reductive citric acid cycle; WL, Wood-Ljungdahl pathway; 3HB, 3-hydroxypropionate bicycle. (**B**) ANRA, assimilatory nitrate reduction to ammonia; DNRA, Dissimilatory nitrate reduction to ammonia; Anammox, anaerobic ammonia oxidation. (**C**) ASR, assimilatory sulfate reduction; DSR, dissimilatory sulfate reduction. (**D**) DMSP, dimethylsulfoniopropionate; MMPA, methylmecaptopropionate; MeSH, methanethiol; DMSO, dimethylsulfoxide; *L*-Met, *L*-methionine.

Fig. 4. Bubble plots depicting the relative abundance of pathways for carbon (**A**), sulfur (**B**), nitrogen (**C**) and other selected (**D**) processes. The key marker genes used to report on the genetic potential for pathways (as the relative abundances) are indicated in brackets. ASR, assimilatory sulfate reduction; DSR, dissimilatory sulfate reduction. The full name of these key marker genes can be found in Supplementary Table S1. For better visualization, we multiply the relative abundance by 10^{-3} and transformed by log(10).

Table 1: A summary of sampling sites and environmental parameters for collected samples

Sample name	Sample type	Latitude	Longitude	Depth [m]	T [°C]	pH	H_2S [mM]	CH_4 [mM]	DIC [mM]	O_2 [mM]
NSu-F2b	water/fluid	S 03°47.995'	E 152°06.052'	1155	51.7	4.3	1.61	0.2	3.07	0.07
NSu-F5	water/fluid	S 03°47.955'	E 152°06.080'	1199	31.4	5.1	0.7	0.01	0.18	0.14
Fw-F1b	water/fluid	S 03°43.700'	E 151°40.344'	1709	3.7	6.5	0	0	0.24	0.17
Fw-F3	water/fluid	S 03°43.698'	E 151°40.350'	1705	3.2	7.2	ND	ND	ND	ND
RR-F1b	water/fluid	S 03°43.238'	E 151°40.519'	1685	6.6	7.5	0	0	2.34	0.2

ND – 'not determined'. 0 – below detection limit

733

734 **Table 2** The relative abundance of biogeochemical pathways in metagenomes from the Manus Basin

Pathway	NSu-F2b	NSu-F5	Fw-F1b	Fw-F3	RR-F1b
Photosystem II (psbABCDEF)	0.0687	2.83	0.0315	0	0.264
Photosystem I (psaABCDEF)	0	0.105	0	0	0
Cytochrome b6/f complex (petABCDGLMN)	0.95	0.728	0.448	0.456	1.15
Anoxygenic photosystem II (pufML)	0	0	0	0	0
Anoxygenic photosystem I (pscABCD)	0	0	0	0	0
RuBisCo	13	34.1	31.3	40.9	57.9
CBB cycle (prkB)	12.6	74.8	45.7	55	55.9
rTCA cycle (aclAB, ccsAB, ccl)	74.4	53.7	4.2	2.09	0.871
Wood-Ljungdahl pathway (acsABCDE)	15.5	2.32	0	0	0
3-Hydroxypropionate Bicycle	2.02	2	0.39	0.335	0.661
Glycolysis (glk, pfk, pyk)	123	158	49.8	64.2	94.7
Entner-Doudoroff pathway, glucose-6P -> glyceraldehyde-3P + pyruvate	19.6	31.4	3.39	3.92	8.68
Gluconeogenesis (fbp, pck)	383	281	66.4	75	103
TCA cycle	178	184	38.5	45	76.5
Methanogenesis (mcrABG)	0	0	0	0	0
Methanogenesis, methanol -> methane (mtaABC)	0	0	0	0	0
Methanogenesis, amines -> methane (mtbA, mtmC, mtbC, mttC)	0	0	0	0	0
Methanogenesis, acetate -> methane (cdhCDE)	2.63	0.565	0	0	0
Methanogenesis, CO2 -> methane	4.49	1.74	0.438	0.646	2.65
Methane oxidation, methane -> methanol (mmoBCDXYZ, amoABC)	22	6.15	7.86	6.53	5.7
Methane oxidation, methanol -> formaldehyde (mxaFI, xoxF)	0.101	0	0	0	0
Fermentation to lactate, pyruvate -> lactate (LDH)	4.96	0.19	0	0	0
Fermentation to formate, pyruvate -> formate (pf1D)	0.563	3.88	0	0	0
Fermentation to formate -> CO2 & H2 (fdh)	14.7	14.4	2.88	2.87	4.64
Fermentation to acetate, pyruvate -> acetate (poxB, poxL, acyP)	76.8	40.7	10.2	13.5	21.1
Fermentation to acetate, acetyl-CoA -> acetate (ach1, eutD, pta, acyP)	83.5	80.5	11.3	13.9	21.4
Fermentation to acetate, lactate -> acetate (EC:1.13.12.4)	0	0	0	0	0
Fermentation to ethanol, acetate to acetylaldehyde (ald)	16	39	7.52	9.45	6.19

Fermentation to ethanol, acetyl-CoA to acetylaldehyde (reversible)	2.01	10.1	0.409	0.453	0.403
Fermentation to ethanol, acetylaldehyde to ethanol (adh, mdh)	40.1	69.4	11.8	17.2	16.5
Fermentation to succinate	245	216	32.8	38.8	62.9
Anaplerotic genes (pyruvate -> oxaloacetate)	699	627	88.6	107	137
Dissimilatory nitrate reduction, nitrate -> nitrite (narGHI or napAB)	187	160	28.1	33.4	33.3
Dissimilatory nitrate reduction, nitrite -> ammonia (nirBD or nrfAH)	19.7	86.9	31.6	36.5	65.1
Assimilatory nitrate reduction, nitrate -> nitrite (narB or NR or nasAB)	4.56	7.21	0	0	0.132
Assimilatory nitrate reduction, nitrite -> ammonia (NIT-6 or nirA)	0	0	0	0	0.911
Denitrification, nitrite -> nitric oxide (nirK or nirS)	9.05	70.9	9.87	2.98	4.02
Denitrification, nitric oxide -> nitrous oxide (norBC)	68	338	31	34.7	8.62
Denitrification, nitrous oxide -> nitrogen (nosZ)	31.4	111	11.8	9.11	1.9
Nitrogen fixation, nitrogen -> ammonia (nifKDH)	0.981	2.12	0	0	0
Nitrification, ammonia -> hydroxylamine (amoABC)	22	6.15	7.86	6.53	5.7
Nitrification, hydroxylamine -> nitrite (hao)	0	0	0	0	0
Nitrification, nitrite -> nitrate (nxrAB)	62.9	23.3	4.29	4.79	4.88
Anammox, nitric oxide + ammonia -> hydrazine (hzs)	0.976	5.89	0.286	0.131	1.35
Anammox, hydrazine -> nitrogen (hdh)	0.219	0.254	0	0	0
Assimilatory sulfate reduction, sulfate -> sulfite	89.3	91.9	12.9	15.7	23.6
Assimilatory sulfate reduction, sulfite -> sulfide (cysJI or sir)	10.2	17.5	0.353	0.187	1.81
Dissimilatory sulfate reduction, sulfate -> sulfite (reversible) (sat and aprAB)	103	134	53.7	64.8	92.8
Dissimilatory sulfate reduction, sulfite -> sulfide (reversible) (dsrAB)	6.34	82.6	73.7	83	71.7
Thiosulfate oxidation by SOX complex, thiosulfate -> sulfate	20.3	183	77.5	90.6	147
Alternative thiosulfate oxidation (doxAD)	2.5	3.8	0.871	0.784	1.49
Alternative thiosulfate oxidation (tsdA)	18.8	39.8	1.79	1.05	0.667
Sulfur reduction, sulfur -> sulfide (sreABC)	0	0	0	0	0
Thiosulfate disproportionation, thiosulfate -> sulfide & sulfite (phsABC)	31.9	16.2	1.48	0.818	0.0703
Sulfhydrogenase, (sulfide)n -> (sulfide)n-1	6.64	0.367	0	0	0
Sulfur disproportionation, sulfur -> sulfide & sulfite	0	0	0	0	0
Sulfur dioxygenase	15.9	65.7	47.1	44.3	84.6
Sulfite oxidation, sulfite -> sulfate (sorB, SUOX, soeABC)	64.9	287	56	64.9	80.9
Sulfide oxidation, sulfide -> sulfur (fccAB)	3.05	22.5	15.5	18.6	14.6

DMSP biosynthesis, Met -> DMSP (DSYB or dsyB or mmtN)	5.16	0.286	0	0	0.168
DMSP demethylation, DMSP -> MMPA (dmdA)	9.31	5.7	2.44	0.933	5.57
DMSP demethylation, MMPA -> MeSH (dmdBCD or acuH)	85.2	102	13.1	14	26.8
DMSP cleavage, DMSP -> DMS (ddds or alma1)	25.3	13.9	7.74	8.96	15
DMS oxidation, DMS -> MeSH (dmoA)	8.51	16.4	2.96	1.34	6.59
DMS oxidation, DMS -> DMSO (ddhABC or tmm)	75	53.1	8.73	10.6	13.7
DMSO reduction, DMSO -> DMS (dms or dorA)	32.8	87.4	26.1	43.6	39.2
MddA pathway, MeSH -> DMS (mddA)	1.33	31.1	0.488	0.505	0.501
MeSH oxidation, MeSH -> Formaldehyde (MTO)	0.5	2.55	0	0	0.49
F-type ATPase	232	248	59.3	72.9	113
V/A-type ATPase	23.2	11.1	5.85	5.8	7.34
NADH-quinone oxidoreductase	89.7	173	52.5	57.4	96.3
NAD(P)H-quinone oxidoreductase	0.0815	0.232	0.103	0.17	0.151
Succinate dehydrogenase (ubiquinone)	0	0	0	0	0
Cytochrome c oxidase, cbb3-type	34.8	147	34.3	40.3	74
Cytochrome *bd* ubiquinol oxidase	239	128	5.67	3.47	5.09
Cytochrome o ubiquinol oxidase	11.2	13.5	0.188	0.184	0.944
Cytochrome c oxidase, prokaryotes, aa3-type	39.5	106	42	54.9	104
Cytochrome aa3-600 menaquinol oxidase	0	0	0	0	0
Cytochrome bc1 complex	13.8	35.3	11.7	14.3	21.8
Type I Secretion	7.82	18.8	1.29	1.09	2.93
Type III Secretion	0.0069	0.00643	0.19	0.118	0.0265
Type II Secretion	40.3	57.8	9.38	10.3	9.41
Type IV Secretion	9.82	10.9	0.252	0.218	0.172
Type VI Secretion	3.81	23.7	1.71	1.29	1.3
Sec-SRP	196	200	50.6	56.9	90.5
Twin arginine targeting	183	199	49.7	59.7	83
Type Vabc secretion	0	0	0	0	0
Bacterial chemotaxis	119	70.5	5.09	3.93	8.31
Flagellum assembly	112	51.9	5.06	4.64	10.4
Dissimilatory arsenic reduction	105	181	24.5	26.4	37