







# SCycDB: A curated functional gene database for metagenomic profiling of sulphur cycling pathways

Xiaoli Yu<sup>1</sup>  | Jiayin Zhou<sup>2</sup> | Wen Song<sup>2</sup> | Mengzhao Xu<sup>2</sup> | Qiang He<sup>3</sup> |  
Yisheng Peng<sup>1</sup> | Yun Tian<sup>4</sup>  | Cheng Wang<sup>1</sup> | Longfei Shu<sup>1</sup>  | Shanquan Wang<sup>1</sup> |  
Qingyun Yan<sup>1</sup>  | Jihua Liu<sup>2</sup> | Qichao Tu<sup>2</sup>  | Zhili He<sup>1,5</sup> 

<sup>1</sup>Environmental Microbiomics Research Center, School of Environmental Science and Engineering, Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Sun Yat-sen University, Guangzhou, China

<sup>2</sup>Institute of Marine Science and Technology, Shandong University, Qingdao, China

<sup>3</sup>Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, TN, USA

<sup>4</sup>Key Laboratory of the Ministry of Education for Coastal and Wetland Ecosystems, School of Life Sciences, Xiamen University, Xiamen, China

<sup>5</sup>College of Agronomy, Hunan Agricultural University, Changsha, China

## Correspondence

Zhili He, Environmental Microbiomics Research Center, School of Environmental Science and Engineering, Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Sun Yat-sen University, Guangzhou, China.  
Email: hezhili@mail.sysu.edu.cn

Qichao Tu, Institute of Marine Science and Technology, Shandong University, Qingdao, China  
Email: tuqichao@outlook.com

## Funding information

National Natural Science Foundation of China, Grant/Award Number: 91951207, 31770539, 31971446, 31700427 and 92051110; National Key Research and Development Program of China, Grant/Award Number: 2019YFA0606700, 2020YFA0607600 and 2017YFA0604300; Natural Science Foundation of Shandong Province, Grant/Award Number: ZR201911110287

## Abstract

Microorganisms play important roles in the biogeochemical cycling of sulphur (S), an essential element in the Earth's biosphere. Shotgun metagenome sequencing has opened a new avenue to advance our understanding of S cycling microbial communities. However, accurate metagenomic profiling of S cycling microbial communities remains technically challenging, mainly due to low coverage and inaccurate definition of S cycling gene families in public orthology databases. Here we developed a manually curated S cycling database (SCycDB) to profile S cycling functional genes and taxonomic groups for shotgun metagenomes. The developed SCycDB contains 207 gene families and 585,055 representative sequences affiliated with 52 phyla and 2684 genera of bacteria/archaea, and 20,761 homologous orthology groups were also included to reduce false positive sequence assignments. SCycDB was applied for functional and taxonomic analysis of S cycling microbial communities from four habitats (freshwater, hot spring, marine sediment and soil). Gene families and microorganisms involved in S reduction were abundant in the marine sediment, while those of S oxidation and dimethylsulphoniopropionate transformation were abundant in the soil. SCycDB is expected to be a useful tool for fast and accurate metagenomic analysis of S cycling microbial communities in the environment.

## KEYWORDS

functional gene database, metagenome sequencing, microbial community, sulphur cycling

## 1 | INTRODUCTION

Sulphur (S) is an essential component of important biomolecules such as amino acids, vitamins and enzymes. S cycling is an important biogeochemical process in the Earth's biosphere (Fike et al., 2015; Moran & Durham, 2019; Muyzer & Stams, 2008; Wasmund et al., 2017), and is usually coupled with carbon (C), nitrogen (N) and metal cycling in natural ecosystems (Buongiorno et al., 2019; Landa et al., 2019; Zhu et al., 2018). Microorganisms play important roles in the biogeochemical cycling of S compounds, which are present in a large variety of chemical forms and redox states (Wasmund et al., 2017). S is abundant with active metabolism in diverse environments, such as marine sediments, hot springs, peatlands and coastal sediments (Baker et al., 2015; Hausmann et al., 2018; Lin et al., 2015; Wasmund et al., 2017). Characterizing the function and taxonomy of S cycling microbial communities is therefore of critical importance to understand microbially mediated S cycling processes and their regulatory mechanisms in the environment.

The S cycle consists of inorganic and organic S transformations. In inorganic S transformation, assimilatory sulphate reduction and dissimilatory sulphate reduction processes as well as their key functional genes such as *sat*, *aprAB* and *dsrAB* have been well studied (Müller et al., 2015), while other inorganic S forms, such as thiosulphate, tetrathionate, polysulphide and elemental S, need further clarification in terms of functional genes, pathways and associated microorganisms involved in biotransformation. Organic S transformation and the linkages between inorganic and organic S transformations are also important in the S cycle. As one of the most abundant organosulphur compounds in marine ecosystems, dimethylsulphoniopropionate (DMSP) is mainly produced by phytoplankton and degraded by cleavage and demethylation pathways, subsequently resulting in the generation of dimethyl sulphide (DMS), a climate-active gas, which may influence global warming (Curson et al., 2011, 2018; Li et al., 2014; Moran et al., 2012). In addition, a previous study found that inorganic S oxidation was linked to the biodegradation of volatile organosulphur compounds via *hdr*-like genes (Koch & Dahl, 2018), indicating the importance of linkages between inorganic and organic S transformations in S cycling. However, microbially mediated S cycling is complex in the environment, and much remains to be learned regarding the genes and pathways, especially for organic S transformation and linkages between inorganic and organic S transformation. Thus, it is critical to develop capabilities for the rapid and accurate analysis of S cycling microbial communities via advanced technologies.

Recently, high-throughput amplicon sequencing of functional genes, such as *dsrA* and *dsrB*, has expanded our knowledge on the diversity and composition of sulphite-/sulphate-reducing microorganisms (Pelikan et al., 2016; Vigneron et al., 2018). For example, amplicon sequencing analysis of *dsrB* genes revealed that the diversity of sulphate reducers could be underestimated, with approximately one-third of detected genes as uncharacterized lineages (Vigneron et al., 2018). However, as universal primers are not available for many S cycling genes, characterization of S cycling gene

families and pathways as well as associated microorganisms cannot be resolved by amplicon sequencing approaches. The development of shotgun metagenome sequencing approaches has provided new insights into our understanding of biogeochemical cycling in natural ecosystems (Knight et al., 2018; Nayfach & Pollard, 2016; Quince et al., 2017; Sharpton, 2014). For metagenome sequencing data analysis, comprehensive and reliable orthology databases are of critical importance for accurate metagenomic profiling of functional gene families. An undesired observation is that results of metagenomic analysis are substantially affected by the selection of orthology databases (Nayfach & Pollard, 2016).

To date, several orthology databases such as arCOG (Archaeal Clusters of Orthologous Genes) (Makarova et al., 2015), COG (Clusters of Orthologous Groups) (Galperin et al., 2015), eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) (Huerta-Cepas et al., 2019) and KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa et al., 2016) have been developed and widely used for functional annotation in both genomic and metagenomic studies. These databases have their own distinct features due to differences in the design concept, with arCOG for archaeal annotation (Makarova et al., 2015), COG and eggNOG for annotation of orthologous groups (Galperin et al., 2015; Huerta-Cepas et al., 2019), and KEGG for linking genes with pathways (Kanehisa et al., 2016). These databases present several limitations for analysis of S cycling microbial communities, such as low coverage of S cycling genes (Baker et al., 2015; Vavourakis et al., 2019), difficulties in distinguishing homologous genes (e.g., *sat* vs. *cysC*, *sbp* vs. *cysP*, or *psrA* vs. *phsA*) (Marietou et al., 2018; Rückert, 2016; Wasmund et al., 2017), and long database searching time (Kim et al., 2013; Scholz et al., 2012). Recently, a specific "small database" NCycDB was developed to facilitate shotgun metagenome sequencing data analysis of nitrogen cycling gene families (Tu et al., 2019). NCycDB has been applied to profile N cycling microbial communities from various environments (Anwar et al., 2019; Zhang et al., 2020), demonstrating its high coverage, accuracy and efficiency. Therefore, it is essential to develop a comprehensive and accurate database for fast functional and taxonomic analysis of S cycling microbial communities in metagenomic studies.

In the present study, to understand the microbial ecology of the S cycle in the environment, we present a curated sulphur cycling database (SCycDB) containing 207 S cycling gene families and associated homologous groups involved in eight pathways, including assimilatory sulphate reduction, dissimilatory S reduction and oxidation, S reduction, SOX systems, S oxidation, S disproportionation, organic S transformation, and the linkages between inorganic and organic S transformation. By integrating multiple orthology databases, SCycDB is characterized by high specificity, comprehensiveness, representativeness and accuracy for rapid profiling of S cycling microbial communities. SCycDB was applied to functionally and taxonomically analyse metagenome sequencing data from four different habitats (freshwater, hot spring, marine sediment and soil). The results demonstrate that SCycDB is a powerful tool for rapid

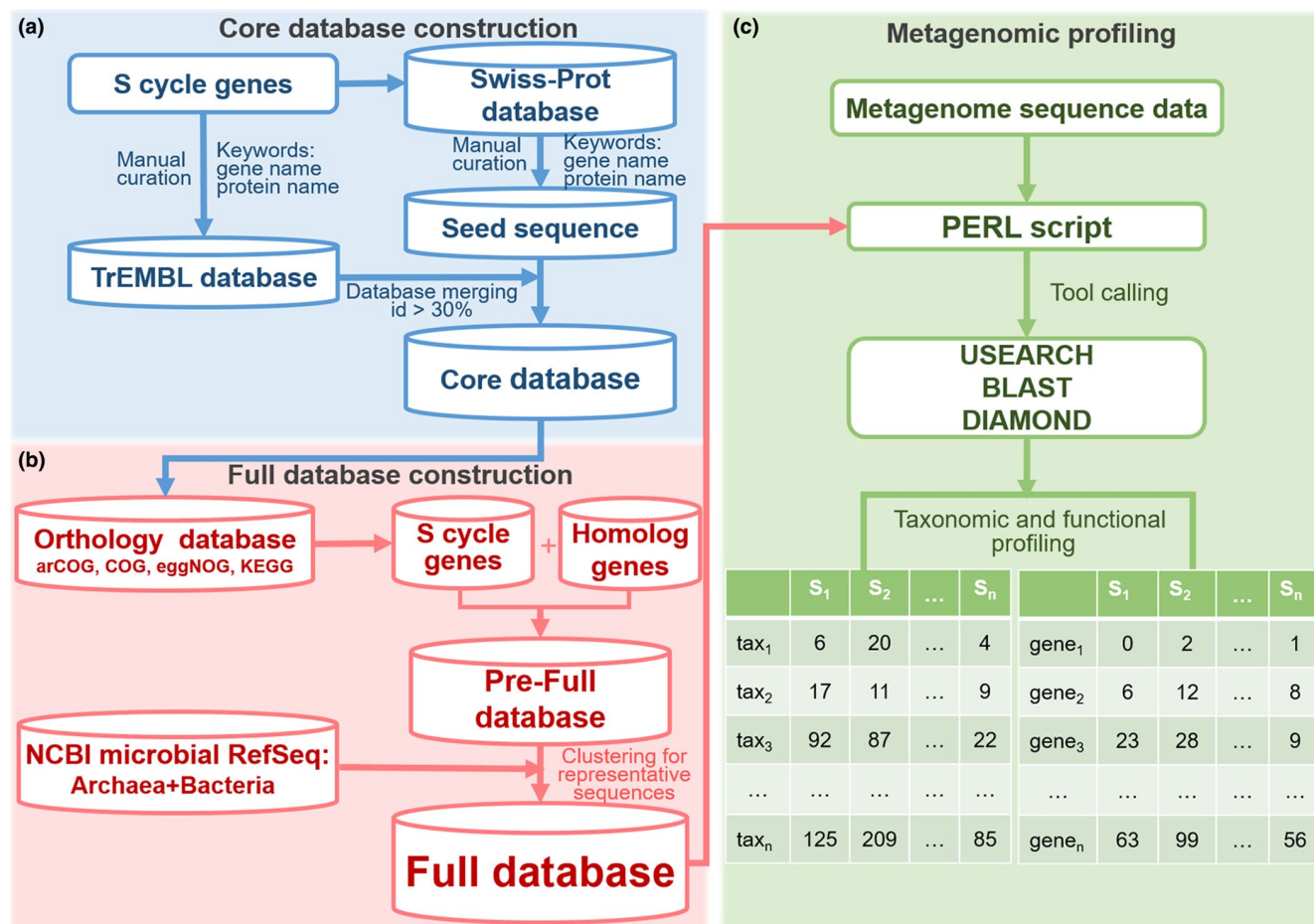
and accurate profiling of S cycling microbial communities, and can be widely used to analyse microbially mediated S cycling processes and underlying mechanisms in the environment.

## 2 | METHODS

### 2.1 | Database construction

An improved pipeline built upon a previous study was used to construct SCycDB (Tu et al., 2019) (Figure 1). First, a core database was manually constructed based on current knowledge of and literature on S cycling processes (Hausmann et al., 2018; Moran & Durham, 2019; Rückert, 2016; Vavourakis et al., 2019; Wasmund et al., 2017). S metabolism pathways in KEGG were also referenced (Kanehisa et al., 2016). By creating and refining keywords for each gene family involved in S cycling, seed sequences for each gene family were downloaded from the Swiss-Prot database (The Uniprot

Consortium, 2017). For gene families without reference sequences in Swiss-Prot, manually checked high-quality sequences were downloaded from TrEMBL (The Uniprot Consortium, 2017). To ensure the accuracy of SCycDB, seed sequences for each gene family were manually checked based on their annotation and similarities with other sequences, especially for those without reference sequences in Swiss-Prot. Sequences downloaded from TrEMBL with the same keywords sharing  $\geq 30\%$  identity with seed sequences were merged with seed sequences, forming the core database (Figure 1a). Second, sequences belonging to S cycling gene families and their orthologues in public databases were identified and merged with the core database, forming the full database (Figure 1b). Publicly available orthology databases including arCOG, COG, eggNOG and KEGG were recruited and searched against the core database. Gene families involved in S cycling and their homologues were identified. Corresponding sequences were extracted and included in SCycDB. By doing so, the comprehensiveness of SCycDB was expected to improve, while the “small database” issue that may lead to increased



**FIGURE 1** A framework of SCycDB construction. (a) Core database construction: seed sequences were retrieved from the Swiss-Prot database using manually refined keywords, and sequences retrieved from the TrEMBL database were merged with the seed sequences at a 30% identity cutoff, generating the core database. (b) Full database construction: S cycling gene families and homologous gene families were retrieved from the public orthology databases and NCBI RefSeq database, and representative sequences were extracted and included in the full database. (c) Metagenomic profiling: PERL scripts were provided to generate both functional and taxonomic profiles for shotgun metagenomes with selected searching tools [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

false positive assignments was expected to diminish or be eliminated (Tu et al., 2019). In addition, corresponding sequences (S cycling gene families and homologues) in the NCBI archaea and bacteria RefSeq databases were also identified, extracted and merged. Taxonomic coverage of S cycling genes and pathways in SCycDB was summarized from corresponding sequences in the NCBI RefSeq. Sequences of both S cycling gene families and homologous gene families were clustered by CD-HIT (Fu et al., 2012) at 100% identity. All representative sequences and related information were checked and used to construct SCycDB. Finally, we included PERL scripts with three candidate database searching tools (USEARCH, BLAST and DIAMOND) for both functional and taxonomic profiling of shotgun metagenomes (Figure 1c). Both functional and taxonomic profiles can be generated by searching raw reads, predicted genes or protein sequences against SCycDB. A random subsampling function is also provided in the PERL scripts to eliminate sequencing depth differences among different samples. Functional profiles of S cycling microbial communities are provided at the gene family level. Taxonomic profiles of S cycling microbial communities are provided at various taxonomic levels.

## 2.2 | Database sources

We used the UniProt database to retrieve seed sequences and construct the core database (The Uniprot Consortium, 2017). The orthology databases used for database merging and homologous gene identification in this study included arCOG (Makarova et al., 2015), COG (Galperin et al., 2015), eggNOG (Huerta-Cepas et al., 2019) and KEGG (Kanehisa et al., 2016). The NCBI RefSeq database (O'Leary et al., 2016) of archaea and bacteria was used for enriching SCycDB and for taxonomically classifying S cycling microbial communities.

## 2.3 | Case study

We applied SCycDB to analyse S cycling microbial communities from four distinct habitats: freshwater, hot spring, marine sediment and soil. The metagenome sequencing data files were downloaded from the NCBI SRA database (<https://www.ncbi.nlm.nih.gov/sra>) (Table S1) (Bahram et al., 2018; Dodsworth et al., 2013; Mitchell et al., 2018; Seitz et al., 2016; Tran et al., 2018). The forward and reverse reads were merged by the program PEAR (options: -q 30) (Zhang et al., 2014). Merged sequences were searched against the arCOG, COG, eggNOG, KEGG and SCycDB databases with the DIAMOND program (options: -k 1 -e 0.0001, -p 20) (Buchfink et al., 2015). Sequences matched to SCycDB were extracted to generate functional profiles of S microbial communities. These sequences were subsequently used to generate taxonomic profiles of S cycling microbial communities at different taxonomic levels using KRAKEN2 (Wood et al., 2019). One-way analysis of variance (ANOVA) was performed with the IBM SPSS 22 (SPSS Inc.), and

used to compare the abundances of gene families and taxonomic groups among different habitats.

## 3 | RESULTS

### 3.1 | Summary of gene families and pathways in SCycDB

The constructed SCycDB contains 585,055 sequences covering 207 gene families involved in eight key S cycling pathways, including assimilatory sulphate reduction, dissimilatory S reduction and oxidation, S reduction, SOX systems, S oxidation, S disproportionation, organic S transformation, and linkages between inorganic and organic S transformation; S compound transporters are also included as "others" (Table 1; Table S2, Figures S1–S3).

#### 3.1.1 | Assimilatory sulphate reduction

A total of 11 gene families with 117,455 representative sequences and 4580 homologous orthology groups are included for this pathway (Table 1; Figure S1A). Among these, gene families including *cysD*, *cysN* and *sat* participate in sulphate activation to adenosine 5'-phosphosulphate (APS), and *cysC* converts APS to phosphoadenosine 5'-phosphosulphate (PAPS). The gene family *cysNC* encodes the biofunctional enzyme CysN/CysC responsible for sulphate conversion to PAPS, *cysH* for PAPS reduction to sulphite, and *cysI*, *cysJ* and *sir* for sulphite reduction to sulfide.

#### 3.1.2 | Dissimilatory sulphur reduction and oxidation

Twenty-two gene families with 20,354 sequences and 775 homologous orthology groups are covered for dissimilatory S reduction and oxidation (Table 1; Figure S1B). The gene family *sat* participates in the conversion between sulphate and APS, and *aprAB* and *qmoABC* for the transformation between APS and sulphite. The *dsr* gene families are involved in both dissimilatory S reduction and oxidation, with some members of the gene families (e.g., *dsrAB*, *dsrC*, *dsrD*, *dsrEFH*, *dsrL*, *dsrMKJOP*) responsible for the transformation between sulphite and sulphide.

#### 3.1.3 | SOX systems

Seven gene families, including *soxA*, *soxB*, *soxC*, *soxD*, *soxX*, *soxY* and *soxZ*, are involved in SOX systems for thiosulphate oxidation to sulphate (Table 1; Figure S1C). The SOX system genes encode SoxAX, SoxYZ, SoxB and SoxCD proteins. A total of 14,998 sequences and 851 homologous orthology groups are included in SCycDB.

TABLE 1 Summary of sulphur (S) cycling gene families with representative sequences and orthology groups

Pathways	Gene	Annotation	Core database sequences	Orthology groups			
				Full database sequences	arCOG	COG	KEGG
Assimilatory sulphate reduction	<i>cysC</i>	Adenylylsulphate kinase	13,037	20,875	39	51	135
	<i>cysND</i>	Sulphate adenylyltransferase	18,468	29,706	19	38	91
	<i>cysH</i>	Phosphoadenosine phosphosulphate reductase	7962	14,572	10	10	32
	<i>cysIJ</i>	Sulphite reductase	8030	19,497	11	21	81
	<i>cysNC</i>	Bifunctional enzyme CysN/CysC	4029	6383	12	12	43
	<i>cysQ</i>	3'(2'), 5'-bisphosphate nucleotidase	9548	14,916	15	28	72
	<i>nmA</i>	Bifunctional oligoribonuclease and PAP phosphatase	308	2,972	2	3	10
	<i>sat</i>	Sulphate adenylyltransferase	4307	6017	8	11	47
	<i>sir</i>	Sulphite reductase (ferredoxin)	418	2517	1	3	8
	<i>aprAB</i>	Adenylylsulphate reductase	92	558	9	8	18
Dissimilatory sulphur reduction and oxidation	<i>dsrAB</i>	Dissimilatory sulphite reductase	11,869	11,741	11	13	31
	<i>dsrC</i>	Dissimilatory sulphite reductase related protein	35	247	0	0	0
	<i>dsrDNT</i>	Protein DsrD DsrN DsrT	19	167	2	NA	3
	<i>dsrEFH</i>	Sulfurtransferase	63	423	NA	1	5
	<i>dsrL</i>	NADPH: acceptor oxidoreductase DsrL	10	43	3	6	19
	<i>dsrMKJOP</i>	Membrane-Bound DsrMKJOP complex	65	626	2	4	24
	<i>qmoABC</i>	Quinone-modifying oxidoreductase	33	417	6	NA	23
	<i>rdsr</i>	Reverse dissimilatory sulphite reductase	81	115	NA	NA	2
	<i>sat</i>	Sulphate adenylyltransferase	4307	6017	8	11	270
	<i>dsrABC</i>	Anaerobic sulphite reductase	716	1833	11	16	35
Sulphur reduction	<i>fsr</i>	Sulphite reductase (coenzyme F420)	3	26	4	4	9
	<i>hydABDG</i>	Sulphhydrogenase	7	161	2	2	2
	<i>mccA</i>	Dissimilatory sulphite reductase	6	15	NA	1	3
	<i>otr</i>	Octaheme tetrathionate reductase Otr	24	248	2	1	12
	<i>psrABC</i>	Polysulphide reductase	6	172	NA	0	6
	<i>rdIA</i>	Putative rhodanese-like protein	8	54	NA	NA	2
	<i>shyABCD</i>	Sulphhydrogenase 2	12	313	0	2	2
	<i>sreABC</i>	Sulphur reductase	12	66	0	1	3
	<i>sudAB</i>	Sulphide dehydrogenase	119	2574	9	20	54
							16

(Continues)

TABLE 1 (Continued)

Pathways	Gene	Annotation	Core database sequences	Orthology groups			
				Full database sequences	arCOG	COG	eggNOG
				6084	18	17	175
				39			
SOX systems	<i>ttrABC</i>	Tetrathionate reductase	1792	6084	18	17	175
	<i>soxAX</i>	L-cysteine S-thiosulphotransferase	1734	4500	5	8	248
	<i>soxB</i>	S-sulfosulfanyl-L-cysteine sulphohydrolase	1357	2068	2	3	64
	<i>soxCD</i>	Sulphite dehydrogenase	1069	3491	4	14	240
	<i>soxYZ</i>	Sulphur-oxidizing protein SoxYZ	2065	4939	3	10	142
Sulphur oxidation	<i>doxAD</i>	Thiosulphate dehydrogenase [quinone]	3	31	0	NA	1
	<i>fccAB</i>	Sulphide dehydrogenase	75	962	1	1	36
	<i>glpE</i>	Thiosulphate sulfurtransferase	3727	8073	1	11	74
	<i>soeABC</i>	Sulphite dehydrogenase	6	617	NA	2	6
	<i>sorAB</i>	Sulphite cytochrome c oxidoreductase	82	918	1	4	20
	<i>sqr</i>	Sulphide:quinone oxidoreductase	89	559	0	2	1
	<i>sseA</i>	Thiosulphate sulfurtransferase	30	3165	0	1	17
	<i>tsdAB</i>	Thiosulphate dehydrogenase	21	1047	NA	0	6
Sulphur disproportionation	<i>phsABC</i>	Thiosulphate reductase	506	1343	4	7	39
	<i>tetH</i>	Tetrathionate hydrolase TetH	3	22	NA	NA	0
	<i>sor</i>	Sulphur oxygenase/reductase	9	29	0	NA	0
	<i>acul</i>	Acrylyl-CoA reductase Acul	458	4301	1	7	121
Organic sulphur transformation	<i>acuNK</i>	Acrylyl-CoA transferase and hydratase	4	63	3	3	9
	<i>betAB</i>	Betaine biosynthesis protein	9452	32,160	9	28	398
	<i>betC</i>	Choline-sulphatase	1689	4490	NA	2	75
	<i>comABCDE</i>	Coenzyme M biosynthesis protein	1933	4377	10	10	221
	<i>dddAC</i>	3-Hydroxypropionate dehydrogenase	8	67	NA	2	6
	<i>dddDKLPQWY</i>	Dimethylsulphoniopropionate lyase	48	441	1	1	8
	<i>dddT</i>	Betaine/carnitine/choline transporter	2	103	NA	NA	7
	<i>ddhABC</i>	Dimethylsulphide dehydrogenase	11	51	1	2	4
	<i>dmdABCD</i>	Dimethylsulphoniopropionate demethylation protein	156	6234	2	6	34
	<i>dmoA</i>	Dimethyl-sulphide monooxygenase	213	1900	2	2	36
	<i>dmsABC</i>	Anaerobic dimethyl sulphoxide reductase	2670	16,676	27	29	245

(Continues)



TABLE 1 (Continued)

Pathways	Gene	Annotation	Core database sequences	Orthology groups			
				Full database sequences	arCOG	COG	eggNOG
Linkages between inorganic and organic sulphur transformation	<i>dsyB</i>	DsyB	94	64	NA	NA	20
	<i>gdl</i>	Glutamate dehydrogenase (NADP+)	32	3931	0	1	10
	<i>hpsN</i>	Sulphopropanediol 3-dehydrogenase	35	976	1	NA	6
	<i>hpsOP</i>	R or S-dihydroxypropanesulphonate-2-dehydrogenase	4	122	1	2	4
	<i>iseJ</i>	Isethionate dehydrogenase	3	6	NA	1	1
	<i>isfD</i>	Sulphoacetaldehyde reductase	143	2726	2	6	22
	<i>mdaA</i>	Methanethiol S-methyltransferase	6	315	NA	0	2
	<i>mdh</i>	Malate dehydrogenase	28,829	30,400	38	67	858
	<i>mtsAB</i>	Methylthiol:coenzyme M methyltransferase	5	37	0	NA	2
	<i>prpE</i>	Propionate-CoA ligase	1295	7177	5	4	93
	<i>pta</i>	Phosphate acetyltransferase	6697	20,805	27	33	736
	<i>sfnG</i>	Dimethylsulphone monooxygenase	10	721	1	0	4
	<i>slcD</i>	Sulpholactate dehydrogenase	57	3216	5	4	43
	<i>sqaBDX</i>	Sulpholipid sulphoquinovosyl diacylglycerol biosynthesis protein	204	2045	1	6	28
	<i>tauXY</i>	Taurine dehydrogenase	24	140	NA	2	10
	<i>tmm</i>	Trimethylamine monooxygenase	24	382	NA	NA	2
	<i>toa</i>	Taurine:2-oxoglutarate transaminase	3	226	2	2	5
	<i>tpa</i>	Taurine-pyruvate aminotransferase	128	2263	5	4	17
	<i>yihQ</i>	Sulphoquinovosidase	35	816	NA	0	2
	<i>cuyA</i>	L-cysteate sulpho-lyase	62	1116	1	0	8
	<i>cysEKM0</i>	Cysteine biosynthesis protein	25,547	62,541	57	90	1389
	<i>hdrABCDE</i>	Heterodisulphide reductase	122	1348	39	35	94
	<i>mccB</i>	Cystathionine gamma-lyase/homocysteine desulphhydrase	176	1238	NA	7	49
	<i>metABCXYZ</i>	L-Cystathionine biosynthesis protein	16,131	52,563	29	39	1103
	<i>msmAB</i>	Methanesulphonate monooxygenase	9	36	NA	1	3
	<i>mtoX</i>	Methanethiol oxidase	11	27	1	1	3

(Continues)

TABLE 1 (Continued)

Pathways	Gene	Annotation	Core database sequences	Orthology groups				
				Full database sequences	arCOG	COG	eggNOG	KEGG
Others	<i>ssuDE</i>	Alkanesulphonate monooxygenase	7663	15,664	4	8	290	44
	<i>suuAB</i>	(2R)-sulpholactate sulpho-lyase	138	1182	3	1	16	11
	<i>tauD</i>	Taurine dioxygenase	1129	6531	1	1	34	14
	<i>tbuBC</i>	Toluene-3-monooxygenase	7	90	NA	2	2	1
	<i>tmoCF</i>	Toluene-4-monooxygenase	10	140	NA	1	4	4
	<i>touCF</i>	Toluene o-xylene monooxygenase	5	14	NA	2	4	1
	<i>xsc</i>	Sulphoacetaldehyde acetyltransferase	1026	2130	1	6	99	16
	<i>cuyZ</i>	Sulphite exporter	1	6	NA	1	5	1
	<i>cysAPUWZ</i>	Sulphate/thiosulphate transporter	18,458	38,216	93	138	1628	415
	<i>hpsKLM</i>	Dihydroxypropanesulphonate transporter	5	172	NA	NA	4	NA
	<i>iseKLM</i>	Isethionate TRAP transporter	3	153	NA	NA	NA	1
	<i>sbp</i>	Sulphate-binding protein	855	5602	2	0	71	7
	<i>sgpABC</i>	Sulphur globule protein	10	288	2	NA	189	25
	<i>soxL</i>	Sulphur transferase, periplasm	2	38	NA	1	1	1
	<i>ssuABC</i>	Sulphonate transport system	9787	34,648	43	104	970	347
	<i>sulp</i>	Sulphate permease	542	4727	31	16	729	82
	<i>tauABC</i>	Taurine transport system	4332	13,826	20	57	270	143
	<i>tauE</i>	Sulphite/organosulphonate exporter	1	32	NA	NA	3	NA
	<i>tauZ</i>	Membrane protein TauZ	7	236	NA	1	14	2
	<i>tusA</i>	Sulphur carrier protein TusA	3431	3907	2	6	39	14
<i>tusBCDE</i>	tRNA 2-thiouridine synthesizing protein	5641	10,234	7	9	132	24	

**Note:** The gene families responsible for identical reactions were combined together. More detailed information is provided in Table S2. NA: not detected in the database.



### 3.1.4 | Sulphur reduction

The S reduction pathway contains 26 gene families encoding sulphite reductase, tetrathionate reductase, S reductase and polysulphide reductase with a total of 11,546 representative sequences and 496 homologous orthology groups (Table 1; Figure S1D). Among these, *asrABC*, *fsr* and *mccA* are responsible for sulphite reduction to sulphide, *otr* and *ttrABC* for tetrathionate reduction to thiosulphate, *sreABC* and *psrABC* for elemental S reduction and polysulphide reduction, respectively, and *hydABDG*, *shyABCD* and *sudAB* for the reduction of both elemental S and polysulphide to sulphide.

### 3.1.5 | Sulphur oxidation

A total of 14 gene families are involved in S oxidation pathways with a total of 15,372 representative sequences and 231 homologous orthology groups (Table 1; Figure S1D). The *fccAB* and *sqr* gene families participate in sulphide oxidation, *doxAD*, *glpE*, *sseA* and *tsdAB* in thiosulphate oxidation, and *soeABC* and *sorAB* in sulphite oxidation.

### 3.1.6 | Sulphur disproportionation

Gene families such as *phsABC*, *tetH* and *sor* are included for this pathway with 1394 sequences and 64 homologous orthology groups (Table 1; Figure S1D). Among these, *phsABC* gene families encode thiosulphate reductase responsible for the transformation of thiosulphate to sulphite and sulphide, *tetH* for the disproportionation of tetrathionate to elemental S, thiosulphate and sulphate, and *sor* for the conversion of elemental S to sulphite and sulphide.

### 3.1.7 | Organic sulphur transformation

There are 57 gene families involved in organic S transformation with a total of 147,231 sequences and 4103 homologous orthology groups (Table 1; Figure S2). Among these, the gene family *dsyB* encodes methyltransferase, a key enzyme for DMSP biosynthesis. For DMSP degradation, two pathways are involved, including the cleavage pathway with *dddDKLPQWY* encoding DMSP lyase for the conversion of DMSP to DMS and acrylate, and the demethylation pathway with *dmdABCD* for the conversion of DMSP to methylmercaptopropionate (MMPA), further generating methanethiol (MeSH) and acetaldehyde (Curson et al., 2011; Moran & Durham, 2019; Moran et al., 2012). The *acul*, *acuKN* and *prpE* gene families participate in acrylate utilization and detoxification, while the *dmsABC*, *ddhABC* and *tmm* families are involved in the transformation between DMS and sulfoxide (DMSO) (Bilous et al., 1988; Lidbury et al., 2016; Wang et al., 2017). Other diverse organic S compounds, such as sulfolipid, sulphonate and sulphate ester, are also involved in organic S metabolism. Two enzymes encoded by *sqdB* and *sqdDX* are related to the biosynthesis of sulfolipid sulphoquinovosyl diacylglycerides

(SQDG), while sulphoquinovosidase encoded by *yihQ* subsequently converts SQDG to sulphoquinovose (SQ) (Moran & Durham, 2019; Speciale et al., 2016). The *tauXY*, *toa*, *tpa* and *iseJ* gene families are responsible for C2 sulphonate (taurine, isethionate) conversion to sulphoacetaldehyde, and *xsc* and *pta* for the transformation of sulphoacetaldehyde to acetyl-CoA (Durham et al., 2019; Landa et al., 2019). The *hpsOPN* and *slcCD* gene families are related to the transformation of C3 sulphonate DHPs, and *betABC* associated with the utilization of sulphate ester choline-*o*-sulphate (Landa et al., 2019).

### 3.1.8 | Linkages between inorganic and organic sulphur transformation

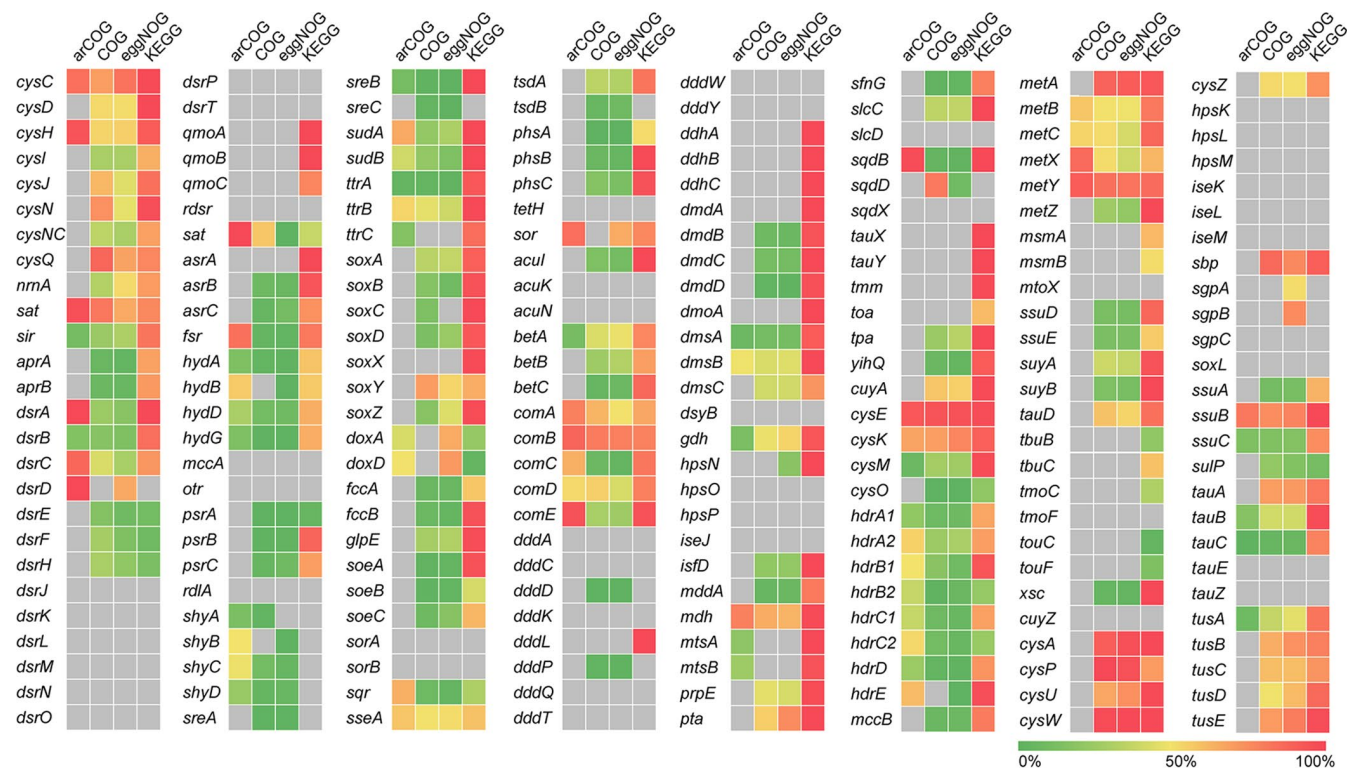
There are 35 gene families responsible for linking inorganic and organic S transformation with a total of 144,620 sequences and 4011 homologous orthology groups (Table 1; Figure S3). The *hdrABCDE* gene families encode a heterodisulphide reductase-like system, which links DMS oxidation with thiosulphate reduction (Koch & Dahl, 2018). Gene families including *cuyA*, *msmAB*, *ssuDE*, *suyAB*, *tbuBC*, *tmoCF*, *touCF* and *xsc* link the transformation between organic S compounds (such as alkanesulphonate, L-cysteate, methanesulphonate, sulfolactate, sulphoacetaldehyde and taurine) and sulphite, with other gene families, namely *cysEKM*, *mccB*, *metABCXYZ* and *mtoX*, linking the transformation between organic S compounds (such as L-cysteine, L-homocysteine, L-serine and MeSH) and sulphide (Byrne et al., 1995; Landa et al., 2019; Moran & Durham, 2019; Wasmund et al., 2017).

### 3.1.9 | Others

Thirty-one gene families encoding various transporters for sulphate, sulphite, thiosulphate and organic S compounds are also included in SCycDB with a total of 112,085 sequences and 5650 homologous orthology groups (Table 1).

## 3.2 | A comparison of gene families detected by SCycDB and other orthology databases

To evaluate the coverage of S cycling gene families in SCycDB, the developed SCycDB was compared with other publicly available orthology databases including arCOG, COG, eggNOG and KEGG. Several critical issues affecting accurate functional assignments in metagenomics were noted. First, there are 207 gene families in SCycDB, while only 62, 130, 138 and 152 gene families are included in the arCOG, COG, eggNOG and KEGG orthology databases, respectively (Figure S4). Second, several key S cycling gene families are included in SCycDB but missing in these four public orthology databases, such as gene families for dissimilatory S reduction and oxidation (*dsrMKJOP*), sulphur reduction (*mccA*, *otr*, *rdIA*), sulphur oxidation (*sorAB*), sulphur disproportionation (*tetH*), organic sulphur



**FIGURE 2** A comparison of S cycling gene families in SCycDB with other public orthology databases. Different colours in the heatmap represent coverage of the selected S cycling gene families in corresponding orthology databases. SCycDB was used as a reference for the comparison. Grey colour indicates the absence of this gene family in the public orthology databases [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

transformation (*dddAC*, *dddKQWY*, *dsyB*) and others (Figure 2). Third, in the four public orthology databases, many different gene families defined by SCycDB were merged into one orthologous group; conversely, single gene families with distinct classification in SCycDB could be correctly found in multiple orthologous groups (Table S3). For instance, *dsrAB*, *asrC* and *fsr* for different S reduction pathways are merged into one orthology group in COG and eggNOG (Table S3). Similarly, *phsA* and *psrA* are not clearly distinguished in COG, eggNOG and KEGG (Table S3), and they were phylogenetically separated in SCycDB (Figure S5). Therefore, SCycDB, specifically designed to target gene families involved in S metabolism, has advantages over existing orthology databases in terms of coverage, representativeness and accuracy.

### 3.3 | Taxonomic composition of S cycling genes and pathways in SCycDB

To understand the taxonomic composition of S cycling genes and pathways in SCycDB, we mapped sequences targeting S cycling genes and pathways to their affiliated reference genomes from the NCBI RefSeq. In total, the developed SCycDB covers 47 phyla, 82 classes, 197 orders, 461 families and 2562 genera of bacteria, and five phyla, 12 classes, 22 orders, 37 families and 122 genera of archaea (Table 2). For bacteria, Proteobacteria (this phylum covers 91.3% of the genes), Firmicutes (67.6%), Actinobacteria (62.8%) and

Bacteroidetes (44.0%) are the dominant phyla, with *Pseudomonas* (this genus covers 51.7% genes), *Escherichia* (45.9%), *Bacillus* (45.4%) and *Vibrio* (36.7%) representing the dominant genera in SCycDB (Table S4). Further analysis shows that organic S transformation has the highest coverage of microorganisms, containing 42 phyla and 2289 genera, and especially assimilatory sulphate reduction accounts for one of the largest coverage groups with 40 phyla and 2059 genera, while 40 phyla and 2204 genera are involved in the linkages between inorganic and organic S transformation (Table 2). For archaea, Euryarchaeota, Crenarchaeota, Thaumarchaeota, Candidatus Bathyarchaeota and Candidatus Korarchaeota are the dominant phyla in SCycDB (Table S4). At the genus level, organic S transformation has the highest diversity with the involvement of 84 genera, followed by assimilatory sulphate reduction (81 genera), and linkages between inorganic and organic S transformation (76 genera) (Table 2). These results indicate that SCycDB covers a high diversity of microorganisms participating in the S cycle, providing a useful platform for the search and annotation of S cycling genes, pathways and associated key microorganisms in the environment.

### 3.4 | Application of SCycDB for functional and taxonomic profiling of environmental samples

We applied SCycDB and four other orthology databases (arCOG, COG, eggNOG and KEGG) to profile S cycling microbial communities

TABLE 2 Summary of S cycling pathways and the number of taxa covered at different taxonomic levels in SCycDB

Pathway	Phylum		Class		Order		Family		Genus	
	Archaea	Bacteria	Archaea	Bacteria	Archaea	Bacteria	Archaea	Bacteria	Archaea	Bacteria
Assimilatory sulphate reduction	5	40	9	74	17	179	25	417	81	2059
Dissimilatory sulphur reduction and oxidation	4	29	8	52	16	117	21	235	49	657
Sulphur reduction	3	23	7	44	11	88	19	173	27	392
SOX systems	1	19	4	35	6	91	8	192	12	683
Sulphur oxidation	2	22	3	34	4	79	7	167	8	506
Sulphur disproportionation	2	4	2	11	2	24	2	37	3	53
Organic sulphur transformation	5	42	12	73	21	182	32	437	84	2289
Linkages between inorganic and organic sulphur transformation	5	40	11	74	19	177	27	405	76	2204
Others	5	33	9	66	13	163	19	378	52	1746
Total	5	47	12	82	22	197	37	461	122	2562
NCBI RefSeq	25	156	18	98	29	224	49	530	194	3815

from four habitats: freshwater, hot spring, marine sediment and soil (Figures 3 and 4; Figure S6). The number of S cycling gene families detected by searching against SCycDB ranged from 174 to 188 in the four habitats, which was significantly ( $p < .05$ ) greater than the other four databases (55–58 in arCOG, 125–128 in COG, 129–134 in eggNOG, 120–135 in KEGG). Notably, the run-time with SCycDB (418–2264 s) was much shorter than with eggNOG (3625–17,749 s) and KEGG (2243–11,161 s) (Table S5).

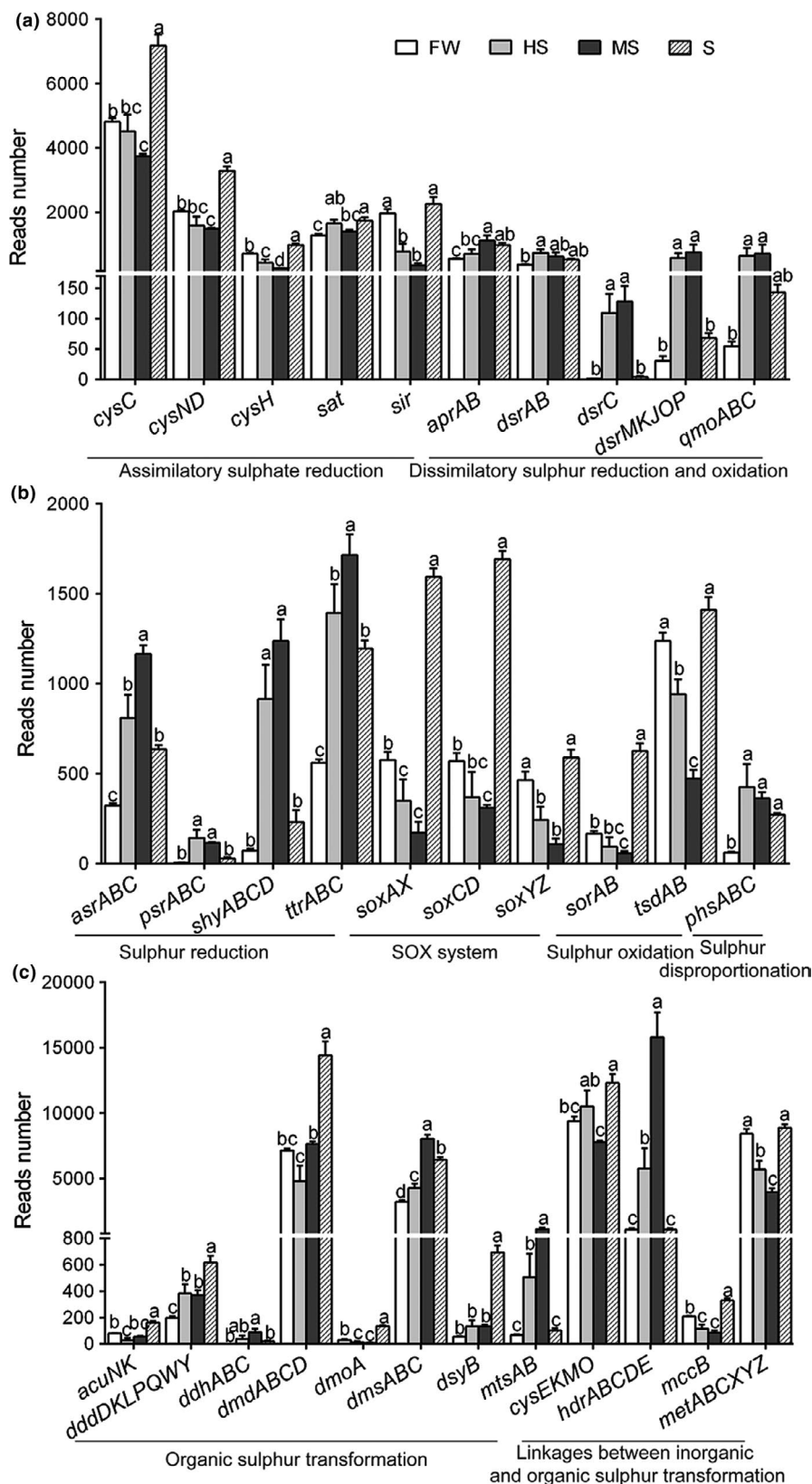
SCycDB profiles of S cycling microbial communities showed that the overall functional or taxonomic composition was significantly different ( $p < .05$ ) among the four habitats profiled in this study (Table S6, Figure S7). Functional profiling of the microbial communities showed S cycling functional genes and pathways were differentially enriched in different habitats (Figure 3). For example, soil habitat exhibited the highest abundance of gene families involved in SOX systems (*soxAX*, *soxCD*, *soxYZ*), and S oxidation (*sorAB*, *tsdAB*) as well as DMSP biosynthesis and degradation (*dsyB*, *dddDKLPQWY*, *dmdABCD*, *acuNK*), with marine sediment having particularly high abundances of S reduction gene families (*asrABC*, *shyABCD* and *ttrABC*) and DMS transformation genes (*ddhABC*, *dmsABC*) (Figure 3). Taxonomic profiling of S cycling microbial communities showed that Proteobacteria was the dominant phylum of S cycling microbial communities in all four habitats (Figure S8), which is consistent with the copious representation of Proteobacteria in SCycDB (Table S4). At the genus level, the abundance of *Desulfallus*, *Desulfobacter*, *Desulfococcus*, *Desulfomonile*, *Desulfotomaculum* and *Syntrophobacter* for dissimilatory sulphate reduction, S reduction and disproportionation was higher in marine sediment than in the other three habitats (Figure 4). In contrast, the soil habitat showed high abundances

of *Halomonas*, *Pseudomonas*, *Rhodobacter*, *Roseobacter*, *Roseovarius*, *Ruegeria*, *Sagittula* and *Sulfitobacter*, which are related to DMSP production and degradation (Figure 4). The above results show that SCycDB is a powerful tool to facilitate analysis of shotgun metagenome sequencing data, enabled by the capacity for fast, comprehensive and accurate functional and taxonomic profiling of S cycling microbial communities in various environments.

## 4 | DISCUSSION

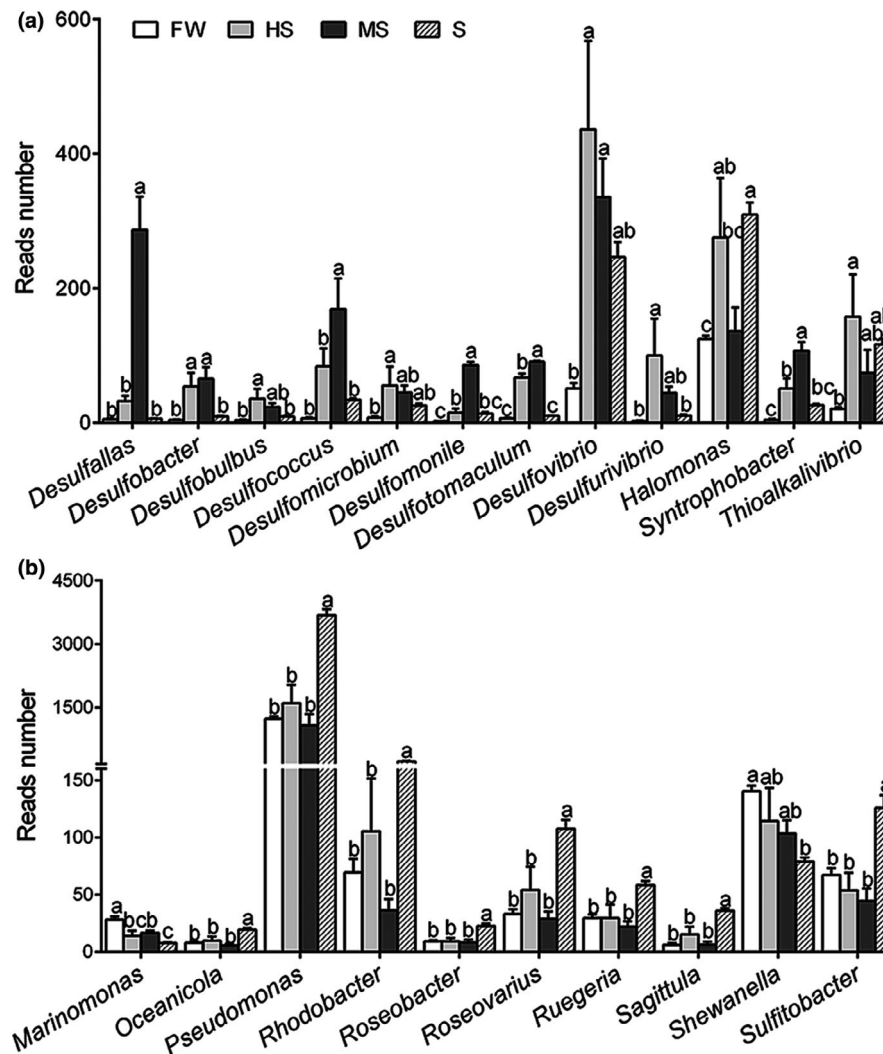
The S cycle is an important biogeochemical process largely driven by microorganisms, impacting the cycling of C and N as well as global change (Curson et al., 2011; Landa et al., 2019; Muyzer & Stams, 2008; Wasmund et al., 2017). Characterizing the function and taxonomy of microbial processes involved in S cycling is critical in providing a better understanding of the diversity of S cycling microbial populations and specific impacts on the environment. Here, we developed SCycDB for fast and accurate functional and taxonomic profiling of S cycling microbial communities, and subsequently applied this database for metagenome sequencing data analysis. The results demonstrate that SCycDB is a useful tool to profile S cycling microbial communities from different environments. To our knowledge, this is the first comprehensive, specific database for analysing both functional and taxonomic profiling of S cycling microbial communities.

Manually curated databases are of vital importance to improve the reliability and reproducibility during bioinformatics analysis of metagenomic data (Kanehisa et al., 2016; The Uniprot Consortium,



**FIGURE 3** Relative abundances of S cycling gene families annotated by SCycDB in four habitats. (a) Assimilatory sulphate reduction and dissimilatory S reduction and oxidation; (b) SOX systems, S reduction, oxidation and disproportionation; (c) organic S transformation and linkages between inorganic and organic S transformation. Functional profiling of those microbial communities identified 154–193 gene families and 112,417–213,847 sequences in those four habitats at a random subsampling of 4,710,299 sequences per sample. Data are presented as mean  $\pm$  SE (standard error,  $n = 6$ ). Different letters (“a,” “b” or “c”) indicate a statistically significant difference ( $p < .05$ ) of each gene family among four habitats. FW, freshwater; HS, hot spring; MS, marine sediment; S, soil





**FIGURE 4** Relative abundances of S cycling microbial communities annotated by SCycDB at the genus level. Taxonomic profiling of S cycling microbial communities identified 32–43 phyla and 692–1340 genera in the four habitats at a random subsampling of 4,710,299 sequences per sample. Data are presented as mean  $\pm$  SE ( $n = 6$ ). Different letters (“a,” “b” or “c”) indicate a statistically significant difference ( $p < .05$ ) of each genus among four habitats. FW, freshwater; HS, hot spring; MS, marine sediment; S, soil

2017). Automatically generated orthology databases, including arCOG, COG, eggNOG and KEGG, cover 62–152 gene families involved in microbial S cycling (Galperin et al., 2015; Huerta-Cepas et al., 2019). In comparison, SCycDB is much more comprehensive, covering 207 gene families with 585,055 representative sequences. The gene families included in SCycDB are retrieved manually based on publicly available databases and most up-to-date knowledge of S cycling. For instance, SCycDB covers gene families otherwise not included in existing databases, such as those involved in DMSP synthesis (*dsyB*) (Curson et al., 2017), acrylate utilization and detoxification (*acuNK* and *dddAC*) (Wang et al., 2017), and DMSP cleavage (*dddKQTYW*) (Li et al., 2017; Moran & Durham, 2019; Peng et al., 2019), enabling researchers to study these newly discovered gene families and metabolic pathways. These gene families have no clearly defined orthology groups in other publicly available databases, but play important roles in regulating marine S cycling and mediating the climate-active gas DMS (Moran & Durham, 2019). Also, SCycDB

includes not only commonly known gene families including *dsrAB*, *dsrC* and *dsrEFH*, but also other poorly known *dsr* gene families (e.g., *dsrMKJOP*, *dsrL*, *dsrN* and *dsrT*) for dissimilatory S reduction and oxidation (Hausmann et al., 2018; Löffler et al., 2020; Pires et al., 2006). In addition, to facilitate both functional annotations and taxonomic assignments that require more accurate sequences with taxonomic information (Tu et al., 2019), the NCBI RefSeq database has been integrated into SCycDB to increase the coverage of functional gene sequences and their associated taxonomic information. Therefore, SCycDB provides the much desired ability to explore questions of “who is there” and “what they are doing” in microbial ecology.

Accuracy is critical in metagenome sequencing data analysis, which is largely dependent on reference databases (Quince et al., 2017). The SCycDB ensures its annotation accuracy in three major aspects. First, gene families and annotations have one-to-one corresponding relationships. As automatically generated orthology databases identify orthologous groups based on species-aware

clustering algorithms (Huerta-Cepas et al., 2019), they could not clearly distinguish different homologous genes. For example, gene families *psrA* and *phsA* respectively encoding polysulphide reductase and thiosulphate reductase subunit are highly homologous, and thus they are always mis-annotated as a single orthology group in automatically generated orthologues databases. In SCycDB, we have carefully looked into this issue, and manually separated them into two orthology groups. Second, SCycDB reduces potential mis-annotations, which may occur in automatically generated orthology databases. For example, the *cysC* gene sequences are generally grouped with *sat* sequences, resulting in the possibility of mis-annotations. Such occurrences are not uncommon, as found in *cysP* vs. *sbp*, *metB* vs. *mccB*, and *sreA* vs. *soeA*. Particularly problematic is the observation that a sequence may be assigned to more than one orthologous group. Therefore, we have manually checked those sequences and carefully assigned them to the correct gene groups to reduce possible mis-annotations in SCycDB. Third, several databases for profiling specific gene families, such as ARDB (for antibiotic resistance genes) and NCycDB (for N cycling genes) were recently developed (Liu & Pop, 2009; Tu et al., 2019). False positives could be an issue arising from the relatively small size of these specialized databases (Tu et al., 2019). To solve such a "small database" issue, SCycDB deliberately includes S cycling-related homologous orthology groups identified from multiple publicly available orthology databases. Therefore, the accuracy of annotation has been considerably enhanced with the implementation of these features.

Unlike other orthology databases, SCycDB is specific to profile S cycling microbial communities, resulting in fast annotation of functional genes, pathways and taxonomy. As shotgun metagenome sequencing data increase exponentially, fast processing of metagenome data sets is critical for metagenomic studies (Kim et al., 2013; Scholz et al., 2012; Wood & Salzberg, 2014; Zhou et al., 2015). A study of the taxonomic classifier MetaPhyler showed that it was much faster than other tools (PhyML, MEGAN, WebCarma) as its reference database was smaller than a general reference database (Liu et al., 2011). Also, a specific database NCycDB provides a fast profiling platform to identify N cycling gene families (Tu et al., 2019). In our study, we used 370 G metagenome data sets and ran on 20 CPU threads, resulting in run times of ~8, 66 and 42 h for SCycDB, eggNOG and KEGG, respectively. Therefore, SCycDB is a much faster database for the annotation of S cycling microbial communities in metagenomic studies.

Functional and taxonomic profiles are important objectives in shotgun metagenome sequencing data analysis to understand microbial communities from different environments (Knight et al., 2018; Quince et al., 2017). Accurate functional profiling requires comprehensive sequence databases for specific metabolic pathways, which is frequently unavailable. Using S metabolism as an example, several previous metagenomic studies only focused on inorganic S cycling, especially dissimilatory sulphate reduction (Baker et al., 2015; Hausmann et al., 2018; Vavourakis et al., 2019), probably due to a lack of organic S cycling gene families in the reference

database. In this study, we included organic S cycling in SCycDB, and used it to analyse functional and taxonomic profiles of S cycling microbial communities from four types of environments, providing a full picture of microbial communities in natural ecosystems. Our results revealed a high diversity of S cycling gene families (154–193 gene families) and microorganisms (32–43 phyla and 692–1340 genera) in natural environments, especially for organic S transformation microbial communities. Also, we found significant variations in functional and taxonomic composition and structure of S cycling microbial communities among different environments. For instance, higher abundances of S reduction gene families and microorganisms in marine sediments were observed, probably linked to the importance of anaerobic respiration with S compounds as electron acceptors in the marine sediment (Jørgensen et al., 2019; Wasmund et al., 2017). Gene families and microorganisms involved in DMSP and DMS transformation were detected in all four environments, supporting the universal distribution of DMSP and DMS metabolism (Curson et al., 2011, 2017; Moran & Durham, 2019). Indeed, DMSP accounts for 10% of fixed carbon in marine environments and DMS plays an important role in S exchanges between the ocean and atmosphere (Curson et al., 2011; Landa et al., 2019; Todd et al., 2010). Consistently, the abundances of DMS transformation gene families were higher in marine sediment than in the other three environments. However, we identified a high abundance of DMSP biosynthesis and degradation gene families as well as associated microorganisms in the soil habitat, suggesting that DMSP transformation may also be an important process in soil. Therefore, these results demonstrate the vast diversity and importance of microbial S metabolisms in the environment that remain to be explored, which will be greatly facilitated by SCycDB developed in this study.

In summary, SCycDB is a manually curated, comprehensive database for fast and accurate functional and taxonomic analysis of S cycling microbial communities with shotgun metagenome sequencing data. By integrating multiple publicly available databases, the current SCycDB contains 207 gene families and 585,055 representative sequences as well as 20,761 homologous orthology groups to resolve the "small database" issue. Applied to profile S cycling microbial communities from various environments, SCycDB has demonstrated its utility for exploring the S cycling process and associated microbial communities in the environment. The SCycDB developed here provides a comprehensive and fast metagenomic analysis tool specialized for studying S metabolisms that will be periodically updated.

## ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos. 91951207, 31770539, 31971446, 31700427, 92051110), National Key Research and Development Program of China (Grant Nos. 2019YFA0606700, 2020YFA0607600, 2017YFA0604300) and Natural Science Foundation of Shandong Province (Grant No. ZR20191110287).

## CONFLICT OF INTEREST

The authors declare that they have no known competing interests.



## AUTHOR CONTRIBUTIONS

Q.T. and Z.H. designed the database structure. X.Y., J.Z., W.S. and M.X. searched and manually collected the sequences. Q.T. wrote the scripts for database construction. X.Y. constructed the database and drafted the manuscript. Q.H., Y.P., Y.T., C.W., L.S., S.W., Q.Y., J.L., Q.T. and Z.H. revised the manuscript. All authors read and approved the final manuscript.

## DATA AVAILABILITY STATEMENT

SCycDB database files are available at <https://github.com/qichaoc1984/SCycDB>.

## ORCID

Xiaoli Yu  <https://orcid.org/0000-0003-2714-2853>

Yun Tian  <https://orcid.org/0000-0003-3233-4470>

Longfei Shu  <https://orcid.org/0000-0001-9683-906X>

Qingyun Yan  <https://orcid.org/0000-0003-0053-892X>

Qichao Tu  <https://orcid.org/0000-0002-3245-7545>

Zhili He  <https://orcid.org/0000-0001-8225-7333>

## REFERENCES

- Anwar, M. Z., Lanzen, A., Bang-Andreasen, T., & Jacobsen, C. S. (2019). To assemble or not to resemble-A validated Comparative Metatranscriptomics Workflow (CoMW). *GigaScience*, 8(8), giz096. <https://doi.org/10.1093/gigascience/giz096>
- Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., Bengtsson-Palme, J., Anslan, S., Coelho, L. P., Harend, H., Huerta-Cepas, J., Medema, M. H., Maltz, M. R., Mundra, S., Olsson, P. A., Pent, M., Pölme, S., Sunagawa, S., Ryberg, M., ... Bork, P. (2018). Structure and function of the global topsoil microbiome. *Nature*, 560, 233–237. <https://doi.org/10.1038/s41586-018-0386-6>
- Baker, B. J., Lazar, C. S., Teske, A. P., & Dick, G. J. (2015). Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome*, 3, 14. <https://doi.org/10.1186/s40168-015-0077-6>
- Bilous, P. T., Cole, S. T., Anderson, W. F., & Weiner, J. H. (1988). Nucleotide sequence of the *dmsABC* operon encoding the anaerobic dimethylsulphoxide reductase of *Escherichia coli*. *Molecular Microbiology*, 2(6), 785–795. <https://doi.org/10.1111/j.1365-2958.1988.tb00090.x>
- Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1), 59–60. <https://doi.org/10.1038/nmeth.3176>
- Buongiorno, J., Herbert, L. C., Wehrmann, L. M., Michaud, A. B., Laufer, K., Røy, H., Jørgensen, B. B., Szykiewicz, A., Faiia, A., Yeager, K. M., Schindler, K., & Lloyd, K. G. (2019). Complex microbial communities drive iron and sulfur cycling in Arctic fjord sediments. *Applied and Environmental Microbiology*, 85(14), e00949–e919. <https://doi.org/10.1128/aem.00949-19>
- Byrne, A. M., Kukor, J. J., & Olsen, R. H. (1995). Sequence analysis of the gene cluster encoding toluene-3-monooxygenase from *Pseudomonas pickettii* PKO1. *Gene*, 154(1), 65–70. [https://doi.org/10.1016/0378-1119\(94\)00844-1](https://doi.org/10.1016/0378-1119(94)00844-1)
- Curson, A. R. J., Liu, J. I., Bermejo Martínez, A., Green, R. T., Chan, Y., Carrión, O., Williams, B. T., Zhang, S.-H., Yang, G.-P., Bulman Page, P. C., Zhang, X.-H., & Todd, J. D. (2017). Dimethylsulfoniopropionate biosynthesis in marine bacteria and identification of the key gene in this process. *Nature Microbiology*, 2, 17009. <https://doi.org/10.1038/nmicrobiol.2017.9>
- Curson, A. R. J., Todd, J. D., Sullivan, M. J., & Johnston, A. W. B. (2011). Catabolism of dimethylsulphoniopropionate: microorganisms, enzymes and genes. *Nature Reviews Microbiology*, 9(12), 849–859. <https://doi.org/10.1038/nrmicro2653>
- Curson, A. R. J., Williams, B. T., Pinchbeck, B. J., Sims, L. P., Martínez, A. B., Rivera, P. P. L., Kumaresan, D., Mercadé, E., Spurgin, L. G., Carrión, O., Moxon, S., Cattolico, R. A., Kuzhiumparambil, U., Guagliardo, P., Clode, P. L., Raina, J.-B., & Todd, J. D. (2018). DSYB catalyses the key step of dimethylsulfoniopropionate biosynthesis in many phytoplankton. *Nature Microbiology*, 3, 430–439. <https://doi.org/10.1038/s41564-018-0119-5>
- Dodsworth, J. A., Blainey, P. C., Murugapiran, S. K., Swingle, W. D., Ross, C. A., Tringe, S. G., Chain, P. S. G., Scholz, M. B., Lo, C.-C., Raymond, J., Quake, S. R., & Hedlund, B. P. (2013). Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nature Communications*, 4(1), 1854. <https://doi.org/10.1038/ncomms2884>
- Durham, B. P., Boysen, A. K., Carlson, L. T., Groussman, R. D., Heal, K. R., Cain, K. R., Morales, R. L., Coesel, S. N., Morris, R. M., Ingalls, A. E., & Armbrust, E. V. (2019). Sulfonate-based networks between eukaryotic phytoplankton and heterotrophic bacteria in the surface ocean. *Nature Microbiology*, 4(10), 1706–1715. <https://doi.org/10.1038/s41564-019-0507-5>
- Fike, D. A., Bradley, A. S., & Rose, C. V. (2015). Rethinking the ancient sulfur cycle. *Annual Review of Earth and Planetary Sciences*, 43, 593–622. <https://doi.org/10.1146/annurev-earth-060313-054802>
- Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>
- Galperin, M. Y., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, 43(D1), D261–D269. <https://doi.org/10.1093/nar/gku1223>
- Hausmann, B., Pelikan, C., Herbold, C. W., Köstlbacher, S., Albertsen, M., Eichorst, S. A., Glavina del Rio, T., Huemer, M., Nielsen, P. H., Rattei, T., Stingl, U., Tringe, S. G., Trojan, D., Wentrup, C., Woebken, D., Pester, M., & Loy, A. (2018). Peatland *Acidobacteria* with a dissimilatory sulfur metabolism. *The ISME Journal*, 12(7), 1729–1742. <https://doi.org/10.1038/s41396-018-0077-1>
- Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1), D309–D314. <https://doi.org/10.1093/nar/gky1085>
- Jørgensen, B. B., Findlay, A. J., & Pellerin, A. (2019). The biogeochemical sulfur cycle of marine sediments. *Frontiers in Microbiology*, 10, 849. <https://doi.org/10.3389/fmicb.2019.00849>
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1), D457–D462. <https://doi.org/10.1093/nar/gkv1070>
- Kim, M., Lee, K. H., Yoon, S. W., Kim, B. S., Chun, J., & Yi, H. (2013). Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics & Informatics*, 11(3), 102–113. <https://doi.org/10.5808/GI.2013.11.3.102>
- Knight, R., Vrbanc, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolk, T., McCall, L.-I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., ... Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7), 410–422. <https://doi.org/10.1038/s41579-018-0029-9>
- Koch, T., & Dahl, C. (2018). A novel bacterial sulfur oxidation pathway provides a new link between the cycles of organic and inorganic

- sulfur compounds. *The ISME Journal*, 12(10), 2479–2491. <https://doi.org/10.1038/s41396-018-0209-7>
- Landa, M., Burns, A. S., Durham, B. P., Esson, K., Nowinski, B., Sharma, S., Vorobev, A., Nielsen, T., Kiene, R. P., & Moran, M. A. (2019). Sulfur metabolites that facilitate oceanic phytoplankton–bacteria carbon flux. *The ISME Journal*, 13(10), 2536–2550. <https://doi.org/10.1038/s41396-019-0455-3>
- Li, C.-Y., Wei, T.-D., Zhang, S.-H., Chen, X.-L., Gao, X., Wang, P., Xie, B.-B., Su, H.-N., Qin, Q.-L., Zhang, X.-Y., Yu, J., Zhang, H.-H., Zhou, B.-C., Yang, G.-P., & Zhang, Y.-Z. (2014). Molecular insight into bacterial cleavage of oceanic dimethylsulfoniopropionate into dimethyl sulfide. *Proceedings of the National Academy of Sciences of the United States of America*, 111(3), 1026–1031. <https://doi.org/10.1073/pnas.1312354111>
- Li, C.-Y., Zhang, D., Chen, X.-L., Wang, P., Shi, W.-L., Li, P.-Y., Zhang, X.-Y., Qin, Q.-L., Todd, J. D., & Zhang, Y.-Z. (2017). Mechanistic insights into dimethylsulfoniopropionate Lyase DddY, a new member of the Cupin superfamily. *Journal of Molecular Biology*, 429(24), 3850–3862. <https://doi.org/10.1016/j.jmb.2017.10.022>
- Lidbury, I., Kröber, E., Zhang, Z., Zhu, Y., Murrell, J. C., Chen, Y., & Schäfer, H. (2016). A mechanism for bacterial transformation of dimethylsulfide to dimethylsulfoxide: a missing link in the marine organic sulfur cycle. *Environmental Microbiology*, 18(8), 2754–2766. <https://doi.org/10.1111/1462-2920.13354>
- Lin, K.-H., Liao, B.-Y., Chang, H.-W., Huang, S.-W., Chang, T.-Y., Yang, C.-Y., Wang, Y.-B., Lin, Y.-T., Wu, Y.-W., Tang, S.-L., & Yu, H.-T. (2015). Metabolic characteristics of dominant microbes and key rare species from an acidic hot spring in Taiwan revealed by metagenomics. *BMC Genomics*, 16, 1029. <https://doi.org/10.1186/s12864-015-2230-9>
- Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., & Pop, M. (2011). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics*, 12(Suppl 2), S4. <https://doi.org/10.1186/1471-2164-12-S2-S4>
- Liu, B., & Pop, M. (2009). ARDB—antibiotic resistance genes database. *Nucleic Acids Research*, 37(suppl\_1), D443–D447. <https://doi.org/10.1093/nar/gkn656>
- Löffler, M., Feldhues, J., Venceslau, S. S., Kammler, L., Grein, F., Pereira, I. A. C., & Dahl, C. (2020). DsrL mediates electron transfer between NADH and rDsrAB in *Allochromatium vinosum*. *Environmental Microbiology*, 22, 783–795. <https://doi.org/10.1111/1462-2920.14899>
- Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2015). Archaeal Clusters of Orthologous Genes (arCOGs): an update and application for analysis of shared features between *Thermococcales*, *Methanococcales*, and *Methanobacteriales*. *Life*, 5(1), 818–840. <https://doi.org/10.3390/life5010818>
- Marietou, A., Røy, H., Jørgensen, B. B., & Kjeldsen, K. U. (2018). Sulfate transporters in dissimilatory sulfate reducing microorganisms: a comparative genomics analysis. *Frontiers in Microbiology*, 9, 309. <https://doi.org/10.3389/fmicb.2018.00309>
- Mitchell, A. L., Scheremetjew, M., Denise, H., Potter, S., Tarkowska, A., Qureshi, M., Salazar, G. A., Pesseat, S., Boland, M. A., Hunter, F. M. I., ten Hoopen, P., Alako, B., Amid, C., Wilkinson, D. J., Curtis, T. P., Cochrane, G., & Finn, R. D. (2018). EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Research*, 46(D1), D726–D735. <https://doi.org/10.1093/nar/gkx967>
- Moran, M. A., & Durham, B. P. (2019). Sulfur metabolites in the pelagic ocean. *Nature Reviews Microbiology*, 17(11), 665–678. <https://doi.org/10.1038/s41579-019-0250-1>
- Moran, M. A., Reisch, C. R., Kiene, R. P., & Whitman, W. B. (2012). Genomic insights into bacterial DMSP transformations. *Annual Review of Marine Science*, 4, 523–542. <https://doi.org/10.1146/annurev-marine-120710-100827>
- Müller, A. L., Kjeldsen, K. U., Rattei, T., Pester, M., & Loy, A. (2015). Phylogenetic and environmental diversity of DsrAB-type dissimilatory (bi)sulfite reductases. *The ISME Journal*, 9(5), 1152–1165. <https://doi.org/10.1038/ismej.2014.208>
- Muyzer, G., & Stams, A. J. M. (2008). The ecology and biotechnology of sulphate-reducing bacteria. *Nature Reviews Microbiology*, 6, 441–454. <https://doi.org/10.1038/nrmicro1892>
- Nayfach, S., & Pollard, K. S. (2016). Toward accurate and quantitative comparative metagenomics. *Cell*, 166(5), 1103–1116. <https://doi.org/10.1016/j.cell.2016.08.007>
- O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., ... Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1), D733–D745. <https://doi.org/10.1093/nar/gkv1189>
- Pelikan, C., Herbold, C. W., Hausmann, B., Müller, A. L., Pester, M., & Loy, A. (2016). Diversity analysis of sulfite- and sulfate-reducing microorganisms by multiplex *dsrA* and *dsrB* amplicon sequencing using new primers and mock community-optimized bioinformatics. *Environmental Microbiology*, 18(9), 2994–3009. <https://doi.org/10.1111/1462-2920.13139>
- Peng, M., Chen, X.-L., Zhang, D., Wang, X.-J., Wang, N., Wang, P., Todd, J. D., Zhang, Y.-Z., & Li, C.-Y. (2019). Structure-function analysis indicates that an active-site water molecule participates in dimethylsulfoniopropionate cleavage by DddK. *Applied and Environmental Microbiology*, 85(8), e03127–e3118. <https://doi.org/10.1128/aem.03127-18>
- Pires, R. H., Venceslau, S. S., Morais, F., Teixeira, M., Xavier, A. V., & Pereira, I. A. C. (2006). Characterization of the *Desulfovibrio desulfuricans* ATCC 27774 DsrMKJOP complex—a membrane-bound redox complex involved in the sulfate respiratory pathway. *Biochemistry*, 45(1), 249–262. <https://doi.org/10.1021/bi0515265>
- Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, 35, 833–844. <https://doi.org/10.1038/nbt.3935>
- Rückert, C. (2016). Sulfate reduction in microorganisms—recent advances and biotechnological applications. *Current Opinion in Microbiology*, 33, 140–146. <https://doi.org/10.1016/j.mib.2016.07.007>
- Scholz, M. B., Lo, C. C., & Chain, P. S. G. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current Opinion in Biotechnology*, 23(1), 9–15. <https://doi.org/10.1016/j.copbio.2011.11.013>
- Seitz, K. W., Lazar, C. S., Hinrichs, K.-U., Teske, A. P., & Baker, B. J. (2016). Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *The ISME Journal*, 10(7), 1696–1705. <https://doi.org/10.1038/ismej.2015.233>
- Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5, 209. <https://doi.org/10.3389/fpls.2014.00209>
- Speciale, G., Jin, Y., Davies, G. J., Williams, S. J., & Goddard-Borger, E. D. (2016). YihQ is a sulfoquinovosidase that cleaves sulfoquinovosyl diacylglyceride sulfolipids. *Nature Chemical Biology*, 12, 215–217. <https://doi.org/10.1038/nchembio.2023>
- The Uniprot Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(suppl\_1), D158–D169. <https://doi.org/10.1093/nar/gkw1099>
- Todd, J. D., Curson, A. R. J., Nikolaidou-Katsaraidou, N., Brearley, C. A., Watmough, N. J., Chan, Y., Page, P. C. B., Sun, L., & Johnston, A. W. B. (2010). Molecular dissection of bacterial acrylate catabolism – unexpected links with dimethylsulfoniopropionate catabolism and dimethyl sulfide production. *Environmental Microbiology*, 12(2), 327–343. <https://doi.org/10.1111/j.1462-2920.2009.02071.x>
- Tran, P., Ramachandran, A., Khawasik, O., Beisner, B. E., Rautio, M., Huot, Y., & Walsh, D. A. (2018). Microbial life under ice: Metagenome

- diversity and in situ activity of Verrucomicrobia in seasonally ice-covered Lakes. *Environmental Microbiology*, 20(7), 2568–2584. <https://doi.org/10.1111/1462-2920.14283>
- Tu, Q., Lin, L., Cheng, L., Deng, Y., & He, Z. (2019). NCycDB: a curated integrative database for fast and accurate metagenomic profiling of nitrogen cycling genes. *Bioinformatics*, 35(6), 1040–1048. <https://doi.org/10.1093/bioinformatics/bty741>
- Vavourakis, C. D., Mehrshad, M., Balkema, C., van Hall, R., Andrei, A.-Ş., Ghai, R., Sorokin, D. Y., & Muyzer, G. (2019). Metagenomes and metatranscriptomes shed new light on the microbial-mediated sulfur cycle in a Siberian soda lake. *BMC Biology*, 17, 69. <https://doi.org/10.1186/s12915-019-0688-7>
- Vigneron, A., Cruaud, P., Alsop, E., de Rezende, J. R., Head, I. M., & Tsesmetzis, N. (2018). Beyond the tip of the iceberg; a new view of the diversity of sulfite- and sulfate-reducing microorganisms. *The ISME Journal*, 12(8), 2096–2099. <https://doi.org/10.1038/s41396-018-0155-4>
- Wang, P., Cao, H.-Y., Chen, X.-L., Li, C.-Y., Li, P.-Y., Zhang, X.-Y., Qin, Q.-L., Todd, J. D., & Zhang, Y.-Z. (2017). Mechanistic insight into acrylate metabolism and detoxification in marine dimethylsulfoniopropionate-catabolizing bacteria. *Molecular Microbiology*, 105(5), 674–688. <https://doi.org/10.1111/mmi.13727>
- Wasmund, K., Mußmann, M., & Loy, A. (2017). The life sulfuric: microbial ecology of sulfur cycling in marine sediments. *Environmental Microbiology Reports*, 9(4), 323–344. <https://doi.org/10.1111/1758-2229.12538>
- Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20, 257. <https://doi.org/10.1186/s13059-019-1891-0>
- Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3), R46. <https://doi.org/10.1186/gb-2014-15-3-r46>
- Zhang, J., Buhe, C., Yu, D., Zhong, H., & Wei, Y. (2020). Ammonia stress reduces antibiotic efflux but enriches horizontal gene transfer of antibiotic resistance genes in anaerobic digestion. *Bioresource Technology*, 295, 122191. <https://doi.org/10.1016/j.biortech.2019.122191>
- Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30(5), 614–620. <https://doi.org/10.1093/bioinformatics/btt593>
- Zhou, J., He, Z., Yang, Y., Deng, Y., Tringe, S. G., & Alvarez-Cohen, L. (2015). High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. *MBio*, 6(1), e02288–e2214. <https://doi.org/10.1128/mBio.02288-14>
- Zhu, J., He, Y., Zhu, Y., Huang, M., & Zhang, Y. (2018). Biogeochemical sulfur cycling coupling with dissimilatory nitrate reduction processes in freshwater sediments. *Environmental Reviews*, 26(2), 121–132. <https://doi.org/10.1139/er-2017-0047>

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Yu X, Zhou J, Song W, et al. SCycDB: A curated functional gene database for metagenomic profiling of sulphur cycling pathways. *Mol Ecol Resour*. 2021;21:924–940. <https://doi.org/10.1111/1755-0998.13306>