Published online 1 November 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 4 **1**
https://doi.org/10.1093/nargab/lqac080

AsgeneDB: a curated orthology arsenic metabolism gene database and computational tool for metagenome annotation

Xinwei Song[1,2,3], Yiqun Li[1,2,3], Erinne Stirling[1,2,3], Kankan Zhao[1,2,3], Binhao Wang[1,2,3], Yongguan Zhu[4], Yongming Luo[5], Jianming Xu[1,2] and Bin Ma[1,2,3,*]

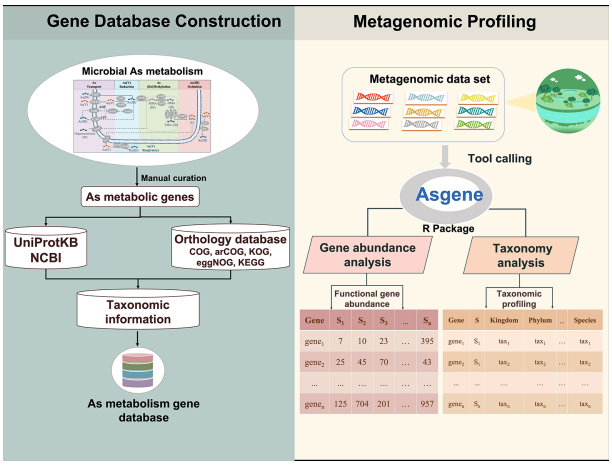
[1]Institute of Soil and Water Resources and Environmental Science, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310000, China, [2]Zhejiang Provincial Key Laboratory of Agricultural Resources and Environment, Zhejiang University, Hangzhou 310000, China, [3]Hangzhou Innovation Center, Zhejiang University, Hangzhou 311200, China, [4]State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100000, China and [5]Key Laboratory of Soil Environment and Pollution Remediation, Institute of Soil Science, Chinese Academy of Science, Nanjing 210000, China

Received April 13, 2022; Revised August 02, 2022; Editorial Decision September 26, 2022; Accepted October 27, 2022

ABSTRACT

Arsenic (As) is the most ubiquitous toxic metalloid in nature. Microbe-mediated As metabolism plays an important role in global As biogeochemical processes, greatly changing its toxicity and bioavailability. While metagenomic sequencing may advance our understanding of the As metabolism capacity of microbial communities in different environments, accurate metagenomic profiling of As metabolism remains challenging due to low coverage and inaccurate definitions of As metabolism gene families in public orthology databases. Here we developed a manually curated As metabolism gene database (AsgeneDB) comprising 400 242 representative sequences from 59 As metabolism gene families, which are affiliated with 1653 microbial genera from 46 phyla. AsgeneDB achieved 100% annotation sensitivity and 99.96% annotation accuracy for an artificial gene dataset. We then applied AsgeneDB for functional and taxonomic profiling of As metabolism in metagenomes from various habitats (freshwater, hot spring, marine sediment and soil). The results showed that AsgeneDB substantially improved the mapping ratio of short reads in metagenomes from various environments. Compared with other databases, AsgeneDB provides more accurate, more comprehensive and faster analysis of As metabolic genes. In addition, we developed an R package, Asgene, to facilitate the analysis of metagenome sequencing data. Therefore, AsgeneDB and the associated Asgene package will greatly promote the study of As metabolism in microbial communities in various environments.

GRAPHICAL ABSTRACT



INTRODUCTION

Arsenic (As) is classified as a group I carcinogen by the International Agency for Research on Cancer, known as both 'the king of poisons' and 'the poison of kings' (1). As has therefore been a prime focus of ecology and environmental sciences (1–3). Once elemental As is released from mineral deposits by geological, agricultural and industrial processes, the element's toxicity and mobility can be greatly altered by microbial metabolism (4,5). These metabolic processes play a major role in the global As cycle through mi-

*To whom correspondence should be addressed. Tel: +86 13282198979; Email: bma@zju.edu.cn

crobial oxidation, respiration, reduction and methylation (6), and are mediated by a variety of genes. It has been reported that almost all microorganisms have As resistance and metabolism genes (7). For example, As redox genes encoding cytoplasmic arsenate [As(V)] reductase (*arsC*), periplasmic As(V) respiratory reductase (*arrAB*) and arsenite [As(III)] oxidase (*aioAB*/*arxA*) affect species transformation between As(V) and As(III) (8–10), while As(III) *S*-adenosylmethionine (SAM) methyltransferase (*arsM*) and non-heme iron-dependent dioxygenase (*arsI*) with C–As lyase activity catalyze As methylation and demethylation (11,12). Mechanisms involved in As metabolism can also be co-opted from other processes, with As(III) and As(V) acting as analogs of glycerol and phosphate, allowing microbial uptake through glycerol transporters (*GlpF*) and phosphate transporters (*Pit*/*Pst*) (13,14). As these processes greatly change the toxicity and bioavailability of As, the study of microbial As metabolism genes is of great importance for understanding the process of environmental As metabolism and microbial remediation potential.

Although the mechanisms of microbial As metabolism are well documented and characterized, the distribution and diversity of As metabolic genes in microbial communities is still unclear due to the large proportion of uncultured microorganisms in environmental samples. Previous works investigating the distribution and diversity of several genes have typically used targeted primer sets to conduct analyses such as polymerase chain reaction (PCR), cloning, denaturing gradient gel electrophoresis (DGGE), microarray-based metagenomic techniques (e.g. GeoChip) and quantitative PCR (qPCR) (5,15–17). These methods are limited by their low throughput that only targets one or several specific genes and also by non-specific amplification introduced by the primers. In addition, as primers cannot be designed for unknown nucleic acid sequences, the inability to detect unknown microorganisms is the biggest obstacle to this kind of technology. Characterization of microbial-induced As metabolism at gene- and species-level resolution has become an important method to better understand microbial As metabolism in the current metagenomic era. In contrast, high-throughput sequencing techniques target all genes and do not rely on the specificity and coverage of primers. Shotgun metagenomic sequencing technology can probe the function of unknown microbiomes and enable us to have a detailed understanding of As metabolism in a complex microbiome, so that microbiome metabolism can be used to address environmental issues (2,18). However, metagenomic data analysis requires comprehensive and reliable orthology databases for accurate metagenomic profiling of functional gene families. An undesired observation is that the results of metagenomic analysis are substantially affected by the orthology database (19).

Orthology databases such as arCOG (Archaeal Clusters of Orthologous Genes) (20), COG (Clusters of Orthologous Groups) (21), eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) (22) and KEGG (Kyoto Encyclopedia of Genes and Genomes) (23) have been developed to date and are 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 (20), COG and eggNOG for annotation of orthologous groups (21,22) and KEGG for linking genes with pathways (23). When As metabolism is considered, analytical limitations encountered in these databases include low coverage of As metabolic genes, difficulty in distinguishing homologous genes and long database search times (24,25). Therefore, the development of a comprehensive and accurate database of As metabolism genes is essential for efficient analysis of As metabolism function in microbial communities.

To understand the microbial community of As metabolism in the environment, we developed a manually curated As metabolism gene database (AsgeneDB). We identified 59 As metabolic gene families from five As metabolic pathways (transport, respiration, reduction, oxidation and methylation/demethylation processes) to construct AsgeneDB. AsgeneDB integrates multiple publicly orthology databases and the NCBI RefSeq database, and then manual checks and corrections were made. We demonstrate that AsgeneDB enables researchers to directly study newly discovered As metabolic pathways and gene families, allowing high specificity, comprehensiveness, sensitivity and accuracy. By analyzing metagenomic sequencing data from four habitats (freshwater, hot spring, marine sediment and soil), our results show that AsgeneDB can detect more As metabolism genes and their abundance in environmental microorganisms. Moreover, the vast diversity and importance of microbial As metabolism in the environment remain to be explored. To facilitate metagenomic data comparison and statistics, we developed an R package, Asgene, that can be used to automatically provide statistical results of gene family abundance and functional community composition at different classification levels in different environments. AsgeneDB and the Asgene package will become a convenient tool for comprehensive and accurate metagenomic analysis of arsenic metabolism, greatly promoting research in this area.

MATERIALS AND METHODS

Core database construction

An improved pipeline based on previous research was used to build AsgeneDB (24,25). As the coverage of As metabolic genes in orthology databases is very limited, we mainly retrieved new and more comprehensive genetic information of As metabolism through a literature search (7,16,17,26). As metabolic genes in KEGG were also referenced (23). In conclusion, the genes mentioned in the literature and clearly described by KEGG that play a role in As metabolism were selected as our target genes. Target protein sequences of genes were downloaded from the Swiss-Prot and TrEMBL databases (27) by creating and refining keywords for each gene family involved in As metabolic pathways (including gene and protein names). To ensure the accuracy of AsgeneDB, the seed sequences of each gene family were checked manually based on their annotations and similarity to other sequences, especially for sequences with no reference sequence in Swiss-Prot. For each gene family, protein sequences from the TrEMBL database were searched against the Swiss-Prot database and were clustered into different groups using USEARCH (version 11.0) at a 30%

global identity cut-off. A nearest neighbor clustering procedure was then carried out to cluster sequences into groups. The outlier groups were then checked again to confirm their annotation information in the TrEMBL database and to remove abnormal sequences. The remaining sequences were then retained as the core database for As metabolic gene families (Figure 1A).

Full database construction

After the core database was created, orthology databases including COG, arCOG, KOG, eggNOG and KEGG were searched against the core database. There were two purposes for comparing the databases. The first was to increase the comprehensiveness of the core database. The second was to identify homologous gene families and include them in the full database, thereby reducing false positives in database searching (24). In addition, corresponding sequences (As metabolic gene families) from the NCBI RefSeq database (Identical Protein Groups) of bacteria, archaea and eukarya were identified, extracted and merged. The coverage of As-metabolizing functional species in AsgeneDB was determined by comparing the core database against NCBI RefSeq (options: -evalue 1e-6 -id 60). Complete taxonomic-level information of sequences was determined used TaxonKit (28,29). Finally, the sequence ID and genes were matched with taxonomic information to generate the taxonomy file. Sequences of both As metabolic gene families and homologous gene families were clustered by cd-hit (30) at 100% identity. All representative sequences and related information were checked and used to construct AsgeneDB (Figure 1B).

Database sources

We used the UniProt database (http://www.uniprot.org) to retrieve seed sequences and construct the core database (27). The orthology databases used for database merging and homologous gene identification in this study were arCOG (ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/arCOG/, version ar14) (20), COG (ftp://ftp.ncbi.nih.gov/pub/COG/COG2020/, version COG2020) (21), eggNOG (http://eggnogdb.embl.de/download/ eggnog_5.0/, version 5.0) (22) and KEGG (http://www.genome.jp/kegg/) (23). The microbial NCBI RefSeq database (31) was used to enrich AsgeneDB (https://www.ncbi.nlm.nih.gov/) and for taxonomically classifying microbial communities of As metabolism (ftp://ftp.ncbi.nih.gov/pub/taxonomy/).

Metagenomic profiling of As metabolic genes

To facilitate the analysis of metagenome sequencing data, an R package (Asgene) is provided for metagenomic alignment (nucleic acid or protein sequence), gene abundance standardization and statistics of all samples. The database AsgeneDB was built into the R package Asgene. Therefore, users only need to input several parameters (e.g. search tools, working path, search parameters of tool and filetype) to automatically analyze and output statistical results. Asgene provides example datasets for analysis as input and output to help users better understand the usage of the package. Users can select gene abundance statistics (option: abundance) to normalize read counts per kilobase per million reads (RPKM) to eliminate differences in sequencing depth and reference sequence length between samples. In addition, if the user selects functional species statistics (option: taxonomy), the statistical results of the driving species of each As metabolism gene at different classification levels in the sample can be generated automatically (Figure 1C). The Asgene package is available on github (https://github.com/XinweiSong/Asgene). Our work can be used to analyze metagenomic data, providing functional profiles at the gene family level and composition of the functional microbial community at various classification levels in different environments.

Sensitivity, accuracy and run-time assessment of AsgeneDB

An artificial dataset, including 81 631 As metabolism gene sequences and 54 403 sequences highly similar to As metabolic genes, from the NCBI GeneBank database was used to assess the sensitivity, accuracy and run-time of AsgeneDB. The 41 As metabolic genes and 10 homologous genes were contained in the artificial dataset to calculate the false-positive and false-negative rates. Homologous gene sequences annotated as As metabolic gene or As metabolic gene sequence assigned to the incorrect gene gene family were considered as false-positive annotations. The sequences belonging to As metabolic genes but not assigned were counted as false-negative annotations. The artificial dataset was searched against KEGG, COG, arCOG, KOG and AsgeneDB using DIAMOND (32) with an e-value of $\leq 10^{-4}$ and identity >30%, and against the eggNOG database using eggNOG-mapper with an e-value of $\leq 10^{-4}$ to compare the accuracy of these databases for annotation (22). Each query sequence only output one hits result with the best matching degree (option: –max-target-seqs 1). All searches specified one thread (option: −p 1 or –cpu 1) to calculate the time of annotation the dataset.

Case study

We applied AsgeneDB and the orthology databases (KEGG, eggNOG, COG, arCOG and KOG) to analyze microbial As metabolism from four distinct habitats: freshwater, hot spring, marine sediment and soil. Forty metagenome sequencing data files were downloaded from the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra) (Supplementary Table S2). Raw reads were quality controlled using Trimmomatic v2.39 (33) to trim adaptors and primers, and to filter short (<50 bp) and low-quality reads (<20 bases). The forward and reverse quality-controlled reads were merged by the program idba (34). Merged shotgun metagenome sequences were searched against KEGG, eggNOG, COG, arCOG, KOG and AsgeneDB databases using DIAMOND (parameters: -k 1 1e-10 -p 20 –query-cover 80 –id 50) (32). Subsequent standardization of gene abundance between samples and statistics of gene abundance and As metabolic microbial communities were performed with R studio. We assessed significant differences for the number and abundance (RPKM) of key As metabolic gene families in environmental samples detected

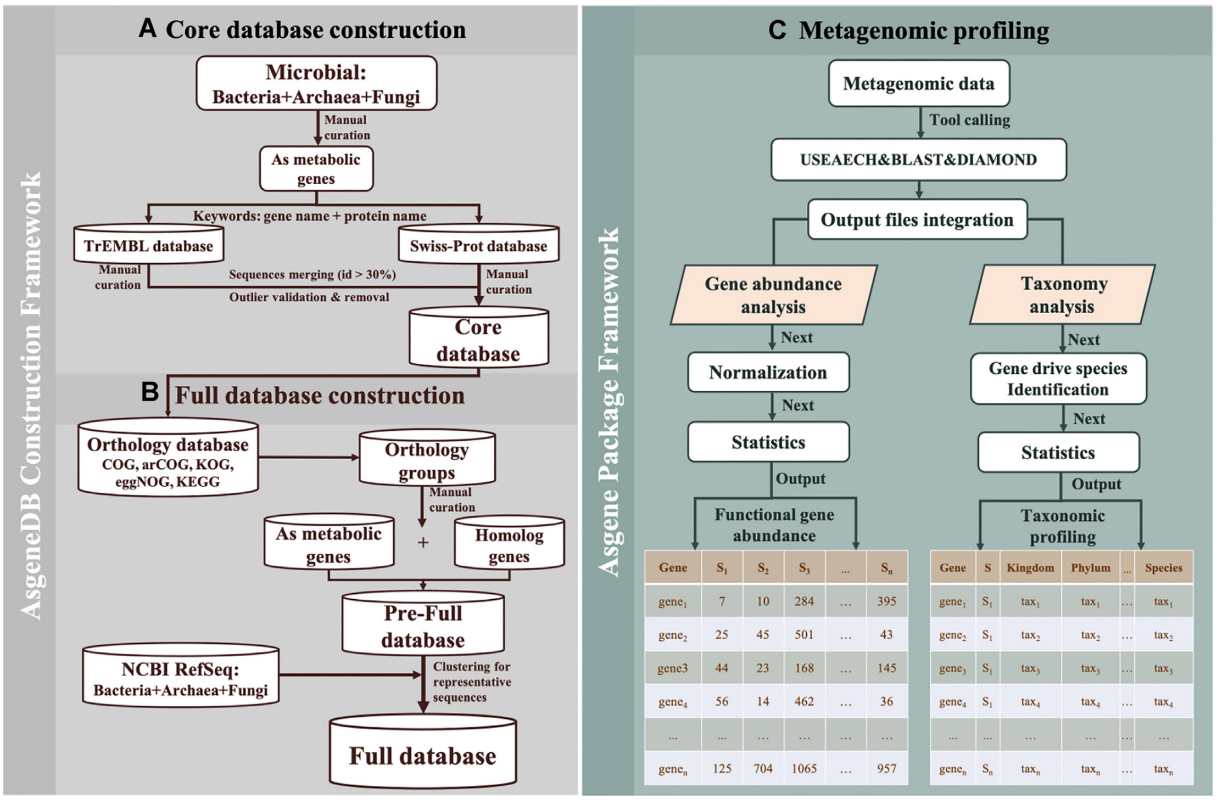


Figure 1. Framework of AsgeneDB construction. (**A**) Core database construction: a core database was constructed for selected gene (sub)families by retrieving protein sequences from UniProt databases using keywords (Swiss-Prot database and TrEMBL database). Sequences that failed to cluster at 30% identity were manually checked again to remove outlier sequences. (**B**) Full database construction: As metabolic 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: the Asgene package generates both gene abundance and taxonomic profiles of environmental samples.

by KEGG, eggNOG, COG, arCOG, KOG and AsgeneDB using one-way analysis of variance (ANOVA) and Tukey's honest significant difference (HSD).

RESULTS

Advantages of the AsgeneDB over other orthology databases

High coverage of As metabolic gene (sub)families. To show the need for construction of a manually curated As metabolism gene database, we compared the coverage of As metabolism genes (subfamily; Figure 2) in AsgeneDB with the main public orthology databases. Of the 59 gene subfamilies recruited to AsgeneDB, fewer than a third were found in any other single database, with the largest proportion found in KEGG (20 gene subfamilies), followed by COG (16 gene subfamilies), eggNOG (13 gene subfamilies), arCOG (6 gene subfamilies) and KOG (2 gene subfamilies). AsgeneDB further contains several key As metabolic gene families that are missing in the five common orthology databases, including As(V) respiratory reductase (*arrA* and *arrB*), organic As efferent osmotic enzyme (*arsJ* and *arsP*), pentavalent As(V) reductase (*GstB*) and trivalent As(III) oxidase (*aioR*, *arxR*, *arxA* and *arxB*; Supplementary Figure S1). In addition to containing more genes, the families defined by AsgeneDB were considered one homologous group in the five public orthology databases. For example, both *arsB* and *acr3* are involved in arsenite efflux even though they belong to two different phylogenetic clades (5,15,35). However, in KEGG, COG and eggNOG databases, arsB and *ACR3* are mixed into one orthology group (Supplementary Table S3). Similarly vague definitions of As metabolic genes were improved in AsgeneDB. We analyzed the phylogenetic evolution of *ACR3* and *arsB* in AsgeneDB, and this illustrated that the sequences of the two genes were obviously distinct in AsgeneDB (Supplementary Figure S3). AsgeneDB is therefore a superior database for determining gene families related to As metabolism and has obvious advantages over existing resources in terms of coverage, completeness and clear definition.

More highly sensitive, accurate and rapid annotation of As metabolic gene families

The sensitivity, accuracy (the rate of false positives and false negatives) and running time of database annotations can be calculated by constructing artificial microbial communities with or without As metabolism genes. We further assessed AsgeneDB, KEGG, COG, arCOG, KOG and eggNOG by the artificial dataset (Figure 3; Supplementary Tables S7 and S9). The results of sensitivity assessment showed that all the sequences belonging to As metabolism in the artificial dataset were assigned by AsgeneDB (100% sensitiv-

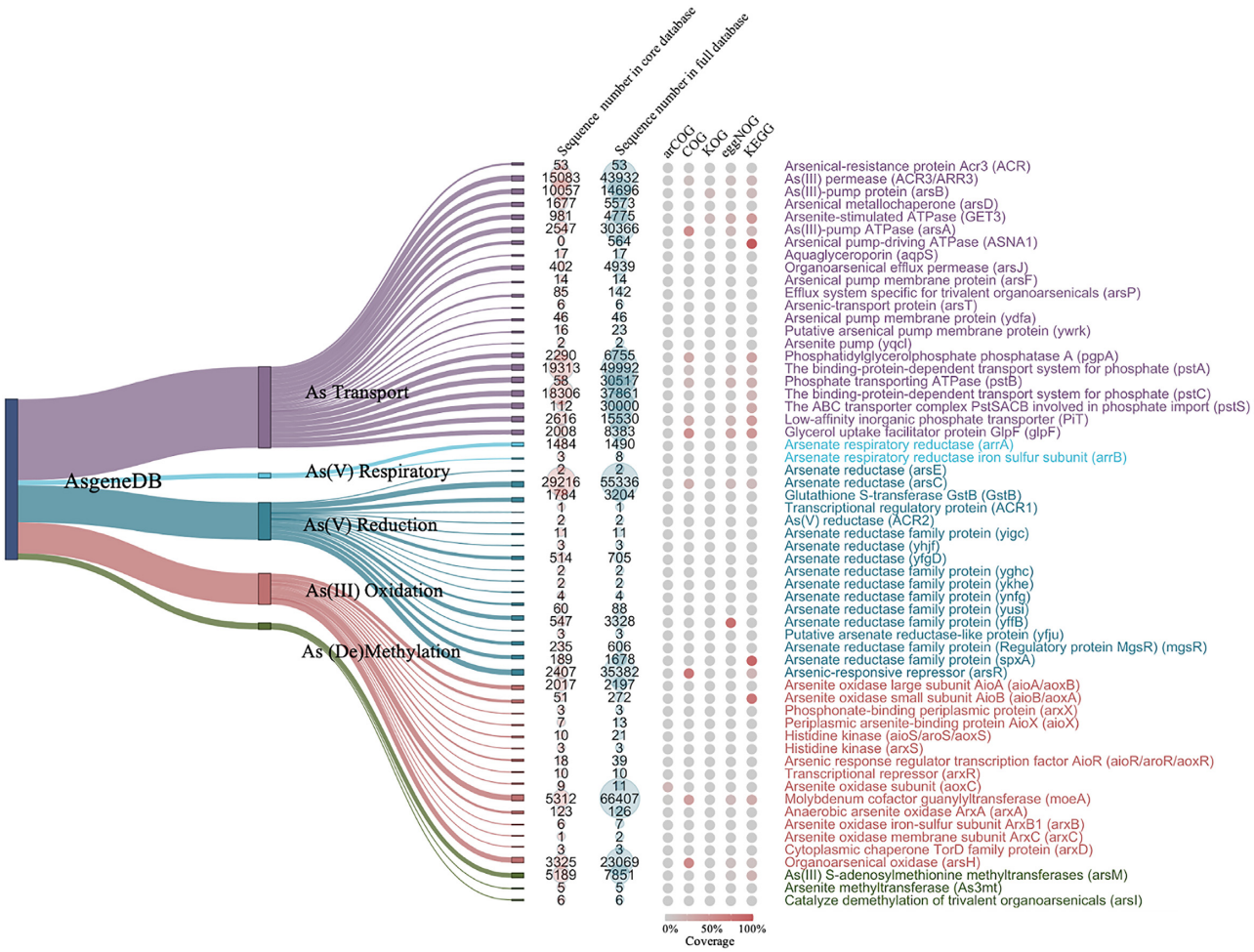


Figure 2. Summary of As metabolic gene families with representative sequences and comparison of As metabolic gene families in AsgeneDB with other public orthology databases. The heatmap represents coverage of the selected As metabolic gene families in corresponding orthology databases. AsgeneDB was used as a reference for the comparison. Gray indicates the absence of this gene family in the public orthology databases.

ity). Furthermore, almost all As metabolism gene sequences were annotated to the correct As metabolism gene family, while other non-As metabolism sequences were not assigned (99.96% accuracy) (Figure 3). In addition, the running time of AsgeneDB was less than that of KEGG, COG and eggNOG databases (Supplementary Table S7).

However, public orthology databases were not as good as AsgeneDB for annotation of As metabolic genes (Figure 3). For KEGG, the accuracy for annotation was the highest of the public orthology databases (89.61%), but a 5.22% false-positive rate and 5.18% false-negative rate were observed. Also, sequences of As metabolism detected by the eggNOG database accounted for 86.18% of the total As metabolic sequences in the artificial dataset. Even though the COG database was highly sensitive to As metabolic sequences (99.97%), many homologous sequences were misassigned as As metabolic genes (false-positive rate 34.07%). KOG and arCOG for fungi and archaea showed very low sensitivity to the As metabolic sequences (3.96% and 0.37% sensitivity). Therefore, the AsgeneDB represents higher sensitivity, accuracy and rapidity for As metabolic gene analysis than other orthology databases.

Summary of gene families and pathways in AsgeneDB

The 59 gene subfamilies in AsgeneDB target five As metabolic pathways (Figure 2), i.e. As transport, As(V) respiration, As(V) reduction, As(III) oxidation and As (de)methylation pathways.

As transport pathway. The As transport pathway includes a total of 22 gene families with 284 186 representative sequences and 386 homologous orthology groups (Figure 2). Among these, the genes responsible for glycerol and phosphate transporters (*glpF*, *PiT*, *pstA*, *pstB*, *pstC* and *pstS*) can absorb As(III) and As(V) as their analogs into microorganisms. Gene families including *arsA*, *arsB*, *aqpS*, *acr3*, *arsF*, *arsT*, *GET3* and *ASNA1* participate in As(III) efflux (36,37). As(III) efflux systems have been intensively studied in both microbes and higher organisms (7,38,39). In particular, the *acr3* gene family is most common in bacteria (40). In addition, the gene family *arsJ* encodes an organoarsenical efflux permease, in which organic As is decomposed into As(V) and 3-phosphoglycerate when excreted from cells. The net reaction is effectively As(V) extrusion, which is the only known efflux pathway for As(V) (41).

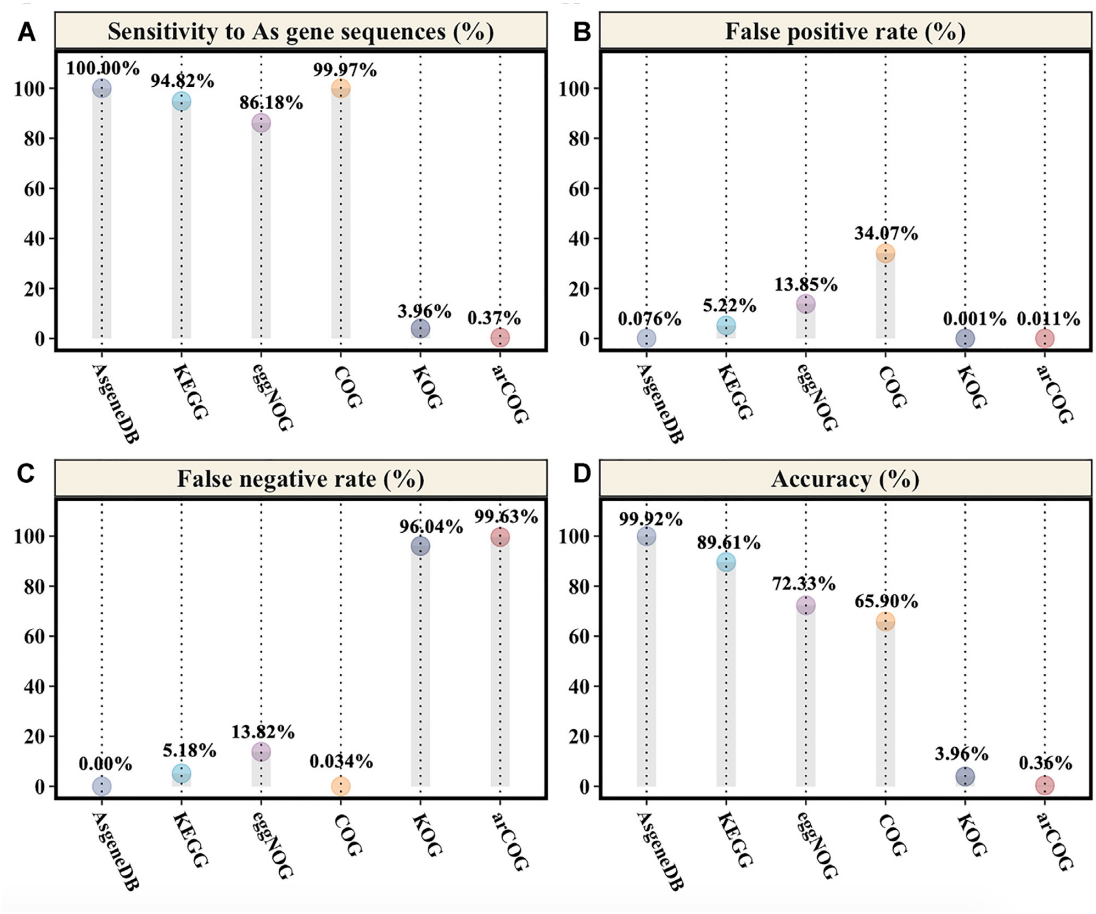


Figure 3. The accuracy assessment for KEGG, eggNOG, COG, arCOG, KOG and AsgeneDB. (**A**) The sensitivity of annotation to As gene sequences. (**B**) The false-positive rate (the rate of mismatched or incorrect sequence annotation). (**C**) The false-negative rate (the rate of unassigned sequences). (**D**) The accuracy rate (the rate at which As metabolic gene sequences were correctly assigned to the correct As metabolic gene families).

Meanwhile, the gene family *arsP* has been demonstrated to be an efflux system specific for trivalent organoarsenicals (42). Since As(III) and As(V) act as analogs of glycerol and phosphate, they can enter microbial cells via glycerol transporters (*GlpF*) and phosphate transporters (*Pit*/*Pst*), respectively.

As(V) respiratory pathway. The As respiratory pathway contains *arrA* and *arrB* gene families with 1498 representative sequences encoding arsenate respiratory reductase (Figure 2; Supplementary Table S5). The large catalytic subunit (ArrA) and small subunit (ArrB) can form a heterodimer (ArrAB) (43,44). Dissimilatory As(V)-respiring prokaryotes (DARPs) have evolved pathways to take advantage of As(V) as a terminal electron acceptor. This energy-generating respiratory chain uses the respiratory As(V) reductase ArrAB, which reduces the less toxic As(V) to the more toxic and potentially more mobile As(III) (45,46,47). It is noteworthy that As(V) respiration and As(III) oxidation functions mainly occur in the periplasm whereas As(V) reduction and As(III) methylation mainly occur in the cytoplasm (46).

As(V) reduction pathway. Gene families such as *arsC*, *acr2* and *GstB* are included for this pathway with 100 357 sequences and 84 homologous orthology groups (Figure 2; Supplementary Table S5). Nearly every extant microbe has ArsB or Acr3 efflux permeases for As(III) detoxification (7). When As(V) became the predominant soluble species, all cells had to do was to reduce As(V) to As(III), the substrate of ArsB or Acr3, and they would become resistant to As(V) (48). However, ArsC, Acr2, GstB, etc. located in the cytoplasm can reduce As(V) in the cytoplasmic membrane and then excrete As(III) through the ArsB or Acr3 efflux pump (49–51). The transcriptional repressor (ArsR) controls these *ars* operons (52,53).

As(III) oxidation pathway. There are 15 gene families responsible for As(III) oxidation, with a total of 92 183 sequences and 39 homologous orthology groups (Figure 3; Supplementary Table S5). As(III)-oxidizing microorganisms exist widely in nature and include both heterotrophic and chemo/photosynthetic autotrophic microorganisms (54). During early life, As(III) oxidation by anaerobes would have produced As(V) in the absence of an oxygen-containing atmosphere, which opened a niche for As(V)-respiring microbes prior to the Great Oxidation Event (GOE) (55). As(III) oxidation is catalyzed by the enzyme As(III) oxidase. This enzyme is composed of two subunits, a large subunit (α) having molybdopterin

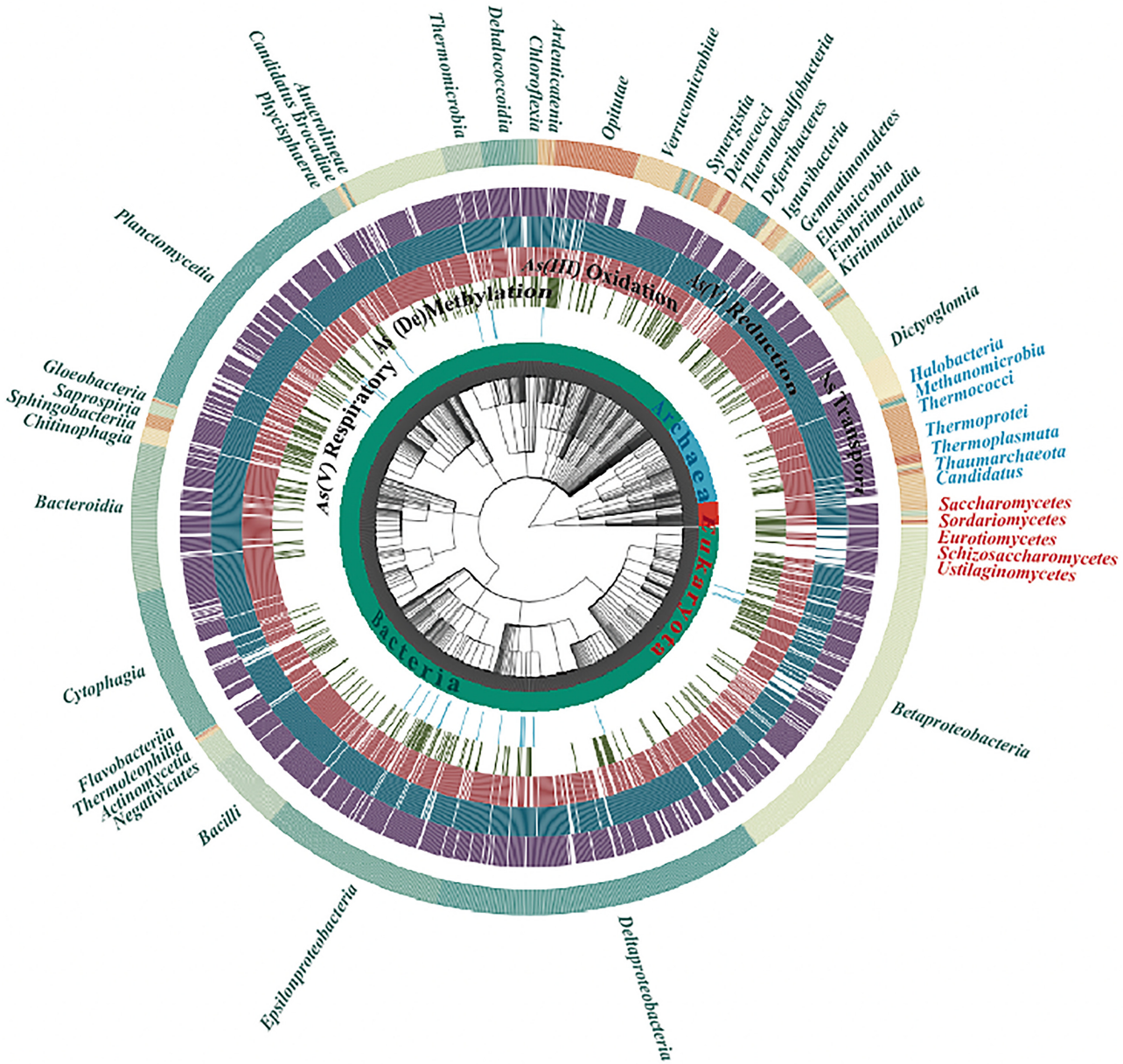


Figure 4. Phylogenetic tree of As metabolic pathways in AsgeneDB. The outermost circle shows the classification of microorganisms in AsgeneDB at the class level.

and a [3Fe–4S] cluster (AioA) and a smaller subunit (β) incorporating a Rieske-type [2Fe–2S] cluster (AioB) (54). Both *aioS*/*aroS*/*aoxS* (sensor histidine kinase) and *aioR*/aroR/*aoxR* (transcriptional regulator) can regulate expression of *aio* genes via recognizing As(III) (56). The operon sometimes has a *aioX*/*arxX* gene that encodes an As(III)-binding protein involved in As(III)-based signaling and regulation of As(III) oxidation, or a *moeA* gene encoding MoeA protein that synthesizes the molybdenum cofactor of AioAB oxidase (56). A new type of As(III) oxidase (*arxA*) has been discovered with both As(V) reductase and As(III) oxidase activities *in vitro* (57). In addition to *arxA*, *arxB*, *arxC*, *arxD* and *arxH* code for As(III) oxidation coupled to photosynthesis (58). An adjacent and divergent gene cluster, *arxXSR*, encodes putative regulatory proteins, a periplasmic substrate-binding protein specific for phosphate (ArxX), a two-component histidine kinase sensor (ArxS) and a response regulator (ArxR) (58). In addition, methylarsenite-specific oxidase ArsH can oxidize methylarsenite to methylarsenate (59,60).

As (de)methylation pathway. Three gene families, inamely *arsM*, *As3mt* and *arsI*, are involved in As methylation and demethylation pathways with 7862 sequences and 24 homologous orthology groups (Figure 3; Supplementary Table S5). More recent reports of methylated As show that As methylation is widespread in the environment (16,60,61). Methylation is catalyzed by the enzyme As(III) SAM

methyltransferase, designated as AS3MT in animals and as ArsM in microorganisms. The gene *arsI*, whose product catalyzes demethylation of organic As(III), was identified and characterized from the environmental isolate bacterium *Bacillus* sp. MD1 (12) and from the cyanobacterium *Nostoc* sp. 7120 (62). ArsI, a non-heme iron-dependent dioxygenase with C–As lyase activity, cleaves the C–As bond in MAs(III), trivalent roxarsone and other trivalent aromatic Asals (63). Putative ArsI orthologs were found only in bacterial species, suggesting that alternative pathways of organoarsenical demethylation might exist in other organisms (7,12).

Taxonomic composition of As metabolic genes and pathways in AsgeneDB

To understand the taxonomic composition of As metabolism genes and pathways in AsgeneDB, we mapped sequences targeting As metabolism genes and pathways to reference genomes from NCBI RefSeq. The results indicate that AsgeneDB covers 46 phyla and 1653 genera of bacteria, archaea and fungi (Supplementary Table S1). In the As transport pathway, AsgeneDB covered 33 phyla and 1141 genera of bacteria, among which the dominant phyla were *Proteobacteria*, *Actinobacteria*, *Firmicutes* and *Bacteroidetes* (Supplementary Table S6). *Euryarchaeota* was the dominant phyla in six phyla of archaea. The predominant Eukaryotaes were *Sordariomycetes*, *Eurotiomycetes* and *Saccharomycetes* in *Ascomycota*, and *Ustilaginomycetes* in *Basidiomycota*. In addition, *Halobacteria* of *Euryarchaeota*, *Betaproteobacteria*, *Deltaproteobacteria* and *Gammaproteobacteria* class of *Proteobacteria*, *Clostridia* in *Firmicutes* and *Deferribacteres* in *Deferribacteres* drove the As(V) respiratory pathway. For the As(V) reduction pathway, AsgeneDB covered 34 bacterial phyla, mainly *Proteobacteria*, *Actinobacteria*, *Firmicutes* and *Bacteroidetes*. It covers six archaea, mainly *Euryarchaeota*, *Candidatus Thermoplasmatota* and *Thaumarchaeota*. *Saccharomycetes* and *Eurotiomycetes* of *Ascomycota* were the dominant Eukaryotaes. The target sequence of the As(III) oxidation pathway covers 29 phyla of bacteria, six phyla of archaea and one phylum of Eukaryotae. For bacteria, *Proteobacteria*, *Actinobacteria*, *Firmicutesand* and *Bacteroidetes* represented the dominant phyla, which were consistent with the results of previous studies (16,64). *Halobacteria* of *Euryarchaeota* and *Sordariomycetes* of *Ascomycota* were the dominant class of bacteria and eukaryotae, respectively. The functional sequences of As methylation and demethylation include 20 phyla of bacteria, four phyla of archaea and two phyla of fungi. The bacteria mainly belonged to *Rhodopseudomonas* in *Proteobacteria*, *Symbiobacterium* in *Firmicutes*, *Dehalogenimonas* in *Chloroflexi* and *Streptomyces* in *Actinobacteria*. The dominant archaea were the class *Methanomicrobia* and *Halobacteria* of *Euryarchaeota*. *Saccharomycetes* in *Ascomycota* were the dominant fungi, which also fit with previous research (2,11). These results suggest that AsgeneDB covers a high diversity of microorganisms involved in As metabolism, providing a useful platform for searching and annotating As metabolic genetic pathways and related key microorganisms in the environment.

Application of AsgeneDB for functional and taxonomic profiling of metagenomes

We applied AsgeneDB and five other orthology databases (KEGG, eggNOG, COG, arCOG and KOG) for taxonomic and functional profiling of As metabolism in metagenomes from freshwater, hot spring, marine sediment and soil (Figures 5 and 6). The number of As metabolic gene families detected by searching sample data against AsgeneDB ranged from 13 to 46 in the four habitats, which was significantly greater (HSD, $P < 0.001$) than the other five databases (1–13 in KEGG, 1–4 in eggNOG, 4–8 in COG, one in arCOG and one in KOG) (Figure 5A). Moreover, AsgeneDB substantially increased the metagenomic mapping rates compared with the other five databases (Figure 5B).

The abundance of As metabolic genes can be affected by both ecosystem and geographic location (2,16), and our results indicated the differences in the biogeographic distribution of As metabolic microbial communities (Figure 5C). Among the five metabolic pathways, the most abundant pathway was As transport and the least abundant was As(V) respiration. Within the four habitats, the As metabolism microbiomes were most similar between marine sediment and soil. Freshwater samples had the lowest diversity in their As metabolism-driven microbiomes. A wide variety of organisms that belong to certain pathways were identified within the samples. Organisms that drive As(III) oxidation, such as *Candidatus Korarchaeota*, *Balneolaeota*, *Chlorobi*, *Spirochaetes*, *Ignavibacteriae*, *Chlamydiae*, *Thermodesulfobacteria* and *Thermotogae*, were found in all habitats except freshwater. *Candidatus Omnitrophica* and *Synergistetes* drove the oxidation of As(III) in marine sediment and soil. *Deferribacteres*, which oxidize As(III), were found only in hot spring and marine sediment. *Synergistetes*, *Chlorobi* and *Candidatus Lokiarchaeota* drove As methylation in sediment and soil, while only *Fusobacteria* drove As methylation in marine sediment. *Candidatus Bipolaricaulot* drove As methylation in all tested environments except freshwater. *Calditrichaeotazai* drove As transport and reduction in hot spring, marine sediment and soil, but only drove As transport in freshwater. *Dictyoglomi* had extensive As(V) reduction functions in hot spring, marine sediment and soil, but was not detected in freshwater. Microbes associated with As(V) respiration were the least diverse, with only *Chrysiogenetes Deferribacteres*, *Firmicutes* and *Proteobacteria* in bacteria and *Euryarchaeota* in archaea detected (Figure 6). In contrast, microorganisms with As transport genes were the most diverse, correlating with the gene abundance of various metabolic pathways in the environment (Figure 5C).

DISCUSSION

Combined with metagenomic methods, the identification of microbial As metabolism pathways and corresponding driving microbes can provide a comprehensive perspective for understanding the complexity of microbial As metabolism in the environment (65–67). This study develops AsgeneDB, a manually curated orthology As metabolism gene database, for fast and accurate annotation of As metabolic genes in shotgun metagenome sequence

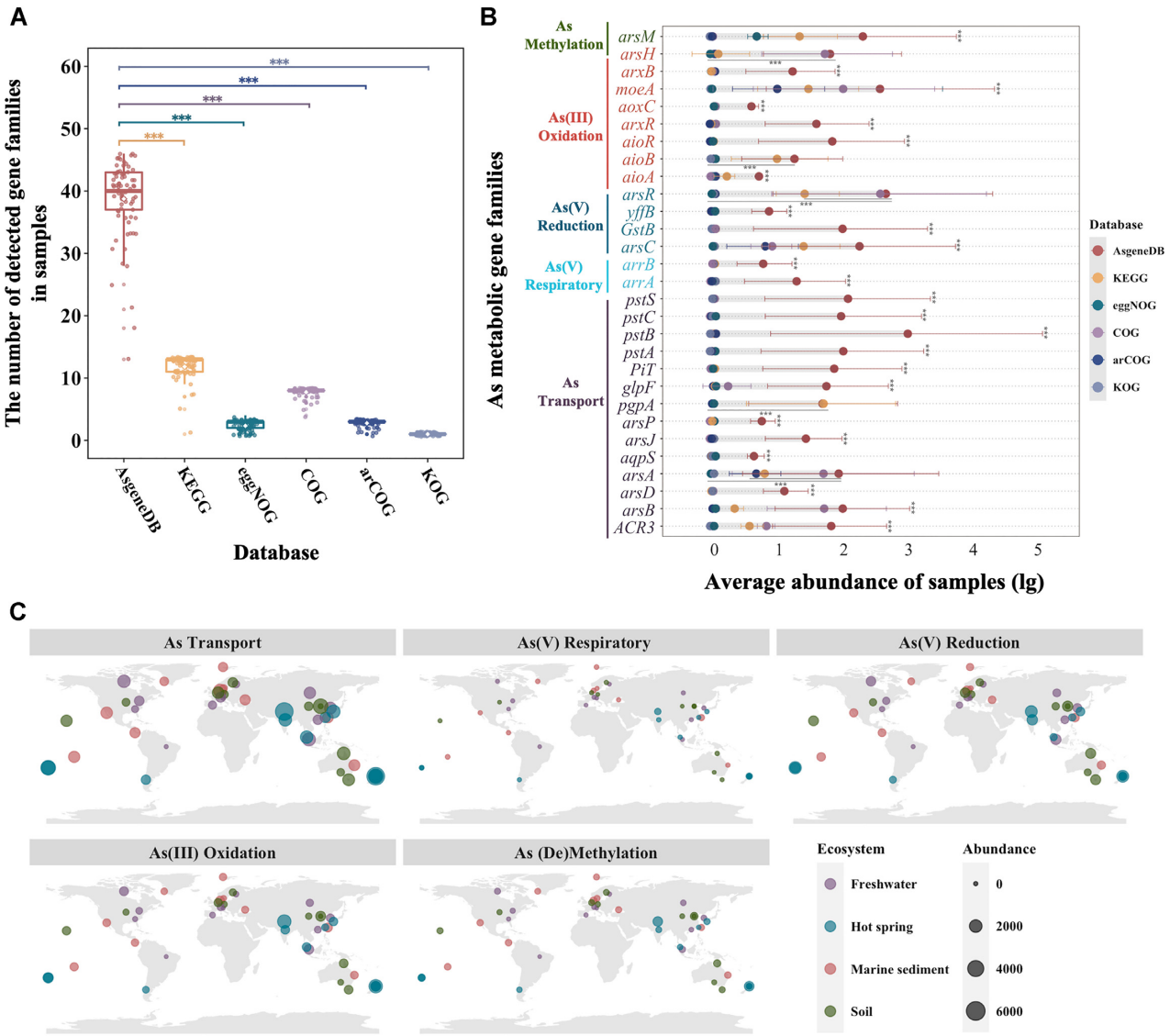


Figure 5. AsgeneDB for functional profiling of As metabolism in metagenomes from freshwater, hot spring, marine sediment and soil. (**A**) Comparison of the number of As metabolism gene families detected using KEGG, eggNOG, COG, arCOG, KOG and AsgeneDB in environmental samples. '***' indicates that the use of AsgeneDB is significantly different from the use of the other five databases (P <0.001). (**B**) Abundances (RPKM) of key As metabolic gene families in environmental samples among KEGG, eggNOG, COG, arCOG, KOG and AsgeneDB. Data are presented as the mean ± SE of all samples (n = 43). '***' indicates that the use of AsgeneDB is significantly different from the use of the other five databases (P <0.001). (**C**) Abundances of As metabolic gene families annotated by AsgeneDB in four different habitats.

data. AsgeneDB has three major advantages over automatically generated orthology databases: accurate annotation, comprehensive information and rapid automated analysis of metagenomic data.

Firstly, it has the precise definition of As metabolic gene families, which were manually inspected and retrieved using keywords combined with sequence similarity, unlike other databases that automatically generate orthology groups based on sequence similarities or sharing of functional domains (21–23). AsgeneDB does not have equivocal annotations for As metabolism genes like public orthology databases (Supplementary Table S3). A typical example is *arsB* and *ACR3*, which belong to two different phylogenetic branches evolutionarily (68). Previous studies have demonstrated that *ACR3* and *arsB* have complementary environmental abundances (68), but they are rarely separated in large databases (5,15). In the AsgeneDB, a clear separation of *the ACR3* gene from *the arsB* gene was observed by the analysis of phylogenetic evolution (Supplementary Figure S3). In addition, precise definitions prevent the misattribution of genes to incorrect families. By comparing with artificial microbial communities with or without As metabolism genes, the results showed that AsgeneDB could annotate As metabolism genes with 99.96% accuracy (Figure 3).

Secondly, both the public orthology databases and existing specialized databases for microbial As metabolism cover between only two and 20 gene families (Supplementary Figure S2; Supplementary Table S8) (21–23,68). As-

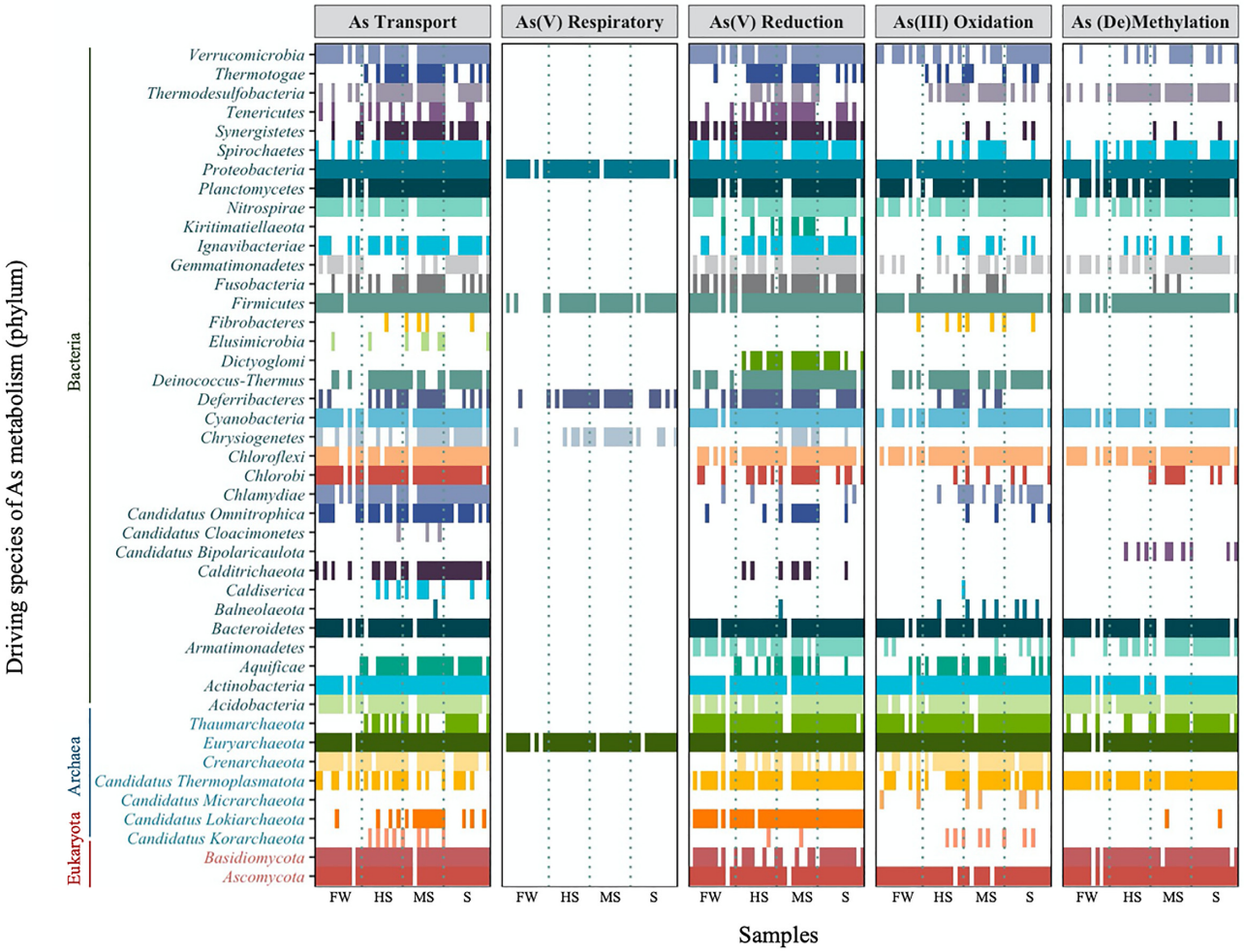


Figure 6. Microbial species driving As metabolism in microbial communities in environmental samples as annotated by AsgeneDB. FW, freshwater; HS, hot spring; MS, marine sediment; S, soil.

geneDB covers 59 gene families with 400 242 representative sequences. The lack of a large number of As metabolic genes in existing database may hinder our understanding of various As metabolism pathways of microbiome in the real environment. Examples include *arsP* (41), a gene family that encodes trivalent organoarsenical [MAs(III)] effluents (42), *GstB*, a newly discovered alternative pathway to arsenate resistance in bacteria (49) and those that encode trivalent As oxidases: *aioR, arxR*, arxA and *arxB* (52,58,69). These gene families have not been clearly defined in other publicly available databases, but play important roles in microbial metabolism of environmental As (7,70). Moreover, AsgeneDB can also be used in the analysis of metabolic genes of As homolog metals, such as antimony (Sb). Many studies have confirmed that microorganisms can metabolize As and Sb through through the same biological processes (71–73). For example, Sb(III) is transported by the As(III) transporter Acr3 or ArsB (71,72), and Sb(V) can be reduced through an arsenate respiratory reductase encoded by arrAB (73). Therefore, AsgeneDB, with its comprehensive and complete information on microbial As metabolism, can contribute a lot of novel information useful to the microbial community.

Thirdly, AsgeneDB itself is relatively small; the Asgene package and database allow researchers to quickly determine 'who has As metabolism' and 'what can they do' in microbiome analyses. Unlike other orthology databases, AsgeneDB achieved 100% sensitivity to As metabolic gene sequences while allowing fast profiling in artificial microbial communities (Figure 3; Supplementary Table S7). Without huge computational cost or output file sizes, the Asgene package will automate metagenomic alignment and results statistics. AsgeneDB takes the 'small database' issue observed in genes into account (24) and addresses it by including homologous gene families from multiple orthology databases (Supplementary Table S4). Therefore, AsgeneDB has the lowest incorrect rate of annotations (false-positive and false-negative rates) compared with the existing databases (Figure 3; Supplementary Table S8).

Finally, metagenomic samples were selected from the natural environment to analyze As metabolism genes and functional species. AsgeneDB significantly increased the average detected numbers and mapping rates of As metabolic genes in all environmental metagenomic data. Moreover, our results also demonstrate that As metabolism genes *aioA*, *arrA* and *arxA* are phylogenetically conserved (68).

The *aioA* gene is limited to *Proteobacteria*: *Alphaproteobacteria*, *Gammaproteobacteria* and *Betaproteobacteria*. *arrA* was detected in *Proteobacteria*, *Firmicutes* and *Eurycota*, while *arxA* was only detected in *Proteobacteria* and *Eurycota* (Supplementary Table S6). Furthermore, microorganisms extensively metabolize As in natural ecosystems (74). Functional genes of different As metabolic pathways could be identified in all environmental samples, and As transport genes are the most abundant and As respiratory genes are the least abundant in environmental samples (Figure 4C). Previous work has also shown that detoxification genes (As transport genes) are more abundant in the microbial communities than As metabolism genes [As(V) respiration, methylation and demethylation genes, etc.] in order to adapt to a wide range of As stress environments (68,75). In addition to the species previously shown to have As(III) oxidation function (16), we find that *Chlamydiae*, *Thermotogae*, *Ignavibacteriae* and *Aquificae* also have As(III) oxidation functions in specific ecosystems. In addition to previous studies [such as (2,11,16)], *Verrucomicrobia*, *Spirochaetes*, *Ignavibacteriae* and *Candidatus Bipolaricaulota* were found to have an As methylation function (Figure 6). There are significant differences in the functional species composition of As metabolism in microbial communities of different ecosystems. *Dictyoglomi*, for example, has As(V) reduction properties in hot spring, marine sediment and soil that are not present in freshwater. Therefore, these results demonstrate the vast diversity and importance of microbial As metabolism functions in the environment that remain to be explored, and which will be greatly facilitated by AsgeneDB.

While genetic migration and limited genetic diversification can be achieved through horizontal gene transfer (HGT) or vertical transfer (68), many As metabolism genes, including *ACR3*, *arsB*, *arsD*, *arsM* and *aioA*, have regional dispersal limitations (68,76). However, the distribution and diversity of large-scale As metabolism genes remain to be further explored. AsgeneDB and the Asgene package are powerful tools for facilitating the analysis of shotgun metagenomic sequencing data, enabling rapid, comprehensive and accurate functional analysis of As-metabolizing microbial communities in a variety of environments. They will greatly promote large-scale genetic research on As metabolism and be updated periodically.

DATA AVAILABILITY

The Asgene package is available on github (https://github.com/XinweiSong/Asgene). AsgeneDB files can be downloaded from cyverse (https://data.cyverse.org/dav-anon/iplant/home/xinwei/AsgeneDB/AsgeneDB.zip).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

This worked was supported by the National Foundation of China [41991334 and 42277283]; and the Zhejiang Natural Science Foundation [LD19D060001].

Conflict of interest statement. None declared.

REFERENCES

1. Zheng,Y. (2020) Global solutions to a silent poison. *Science.*, **368**, 818–819.
2. Zhang,S.-Y., Su,J.-Q., Sun,G.-X., Yang,Y., Zhao,Y., Ding,J., Chen,Y.-S., Shen,Y., Zhu,G., Rensing,C. *et al.* (2017) Land scale biogeography of As biotransformation genes in estuarine wetland. *Environ. Microbiol.*, **19**, 2468–2482.
3. Moon,K.A., Guallar,E., Umans,J.G., Devereux,R.B., Best,L.G., Francesconi,K.A., Goessler,W., Pollak,J., Silbergeld,E.K., Howard,B.V. *et al.* (2013) Association between low to moderate As exposure and incident cardiovascular disease. A prospective cohort study. *Ann. Intern. Med.*, **159**, 649–659.
4. Oremland,R.S. and Stolz,J.F. (2003) The ecology of As. *Science*, **300**, 939–944.
5. Achour,A.R., Bauda,P. and Billard,P. (2007) Diversity of arsenite transporter genes from As-resistant soil bacteria. *Res. Microbiol.*, **158**, 128–137.
6. Mukhopadhyay,R., Rosen,B.P., Phung,L.T. and Silver,S. (2002) Microbial As: from geocycles to genes and enzymes. *FEMS Microbiol. Rev.*, **26**, 311–325.
7. Zhu,Y.-G., Xue,X.-M., Kappler,A., Rosen,B.P. and Meharg,A.A. (2017) Linking genes to microbial biogeochemical cycling: lessons from As. *Environ. Sci. Technol.*, **51**, 7326–7339.
8. Rosenstein,R., Peschel,A., Wieland,B. and Götz,F. (1992) Expression and regulation of the antimonite, arsenite, and arsenate resistance operon of *Staphylococcus xylosus* plasmid pSX267. *J. Bacteriol.*, **174**, 3676–3683.
9. Malasarn,&D. (2004) arrA is a reliable marker for As(V) respiration. *Science*, **306**, 455–455.
10. Sultana,M., Vogler,S., Zargar,K., Schmidt,A.-C., Saltikov,C., Seifert,J. and Schlömann,M. (2012) New clusters of arsenite oxidase and unusual bacterial groups in enrichments from As-contaminated soil. *Arch. Microbiol.*, **194**, 623–635.
11. Jia,Y., Huang,H., Zhong,M., Wang,F.-H., Zhang,L.-M. and Zhu,Y.-G. (2013) Microbial As methylation in soil and rice rhizosphere. *Environ. Sci. Technol.*, **47**, 3141–3148.
12. Yoshinaga,M. and Rosen,B.P. (2014) A C·As lyase for degradation of environmental organoarsenical herbicides and animal husbandry growth promoters. *Proc. Natl Acad. Sci. USA*, **111**, 7701–7706.
13. Borgnia,M., Nielsen,S., Engel,A. and Agre,P. (1999) Cellular and molecular biology of the aquaporin water channels. *Annu. Rev. Biochem.*, **68**, 425–458.
14. Wysocki,R., Chéry,C.C., Wawrzycka,D., Van Hulle,M., Cornelis,R., Thevelein,J.M. and Tamás,M.J. (2001) The glycerol channel fps1p mediates the uptake of arsenite and antimonite in *Saccharomyces cerevisiae*. *Mol. Microbiol.*, **40**, 1391–1401.
15. Cai,L., Liu,G., Rensing,C. and Wang,G. (2009) Genes involved in As transformation and resistance associated with different levels of As-contaminated soils. *BMC Microbiol.*, **9**, 4.
16. Zhang,C., Xiao,X., Zhao,Y., Zhou,J., Sun,B. and Liang,Y. (2021) Patterns of microbial As detoxification genes in low-As continental paddy soils. *Environ. Res.*, **201**, 111584.
17. Wang,H.-T., Zhu,D., Li,G., Zheng,F., Ding,J., O'Connor,P.J., Zhu,Y.-G. and Xue,X.-M. (2019) Effects of As on gut microbiota and its biotransformation genes in earthworm *Metaphiresieboldi*. *Environ. Sci. Technol.*, **53**, 3841–3849.
18. Xiao,K.-Q., Li,L.-G., Ma,L.-P., Zhang,S.-Y., Bao,P., Zhang,T. and Zhu,Y.-G. (2016) Metagenomic analysis revealed highly diverse microbial As metabolism genes in paddy soils with low-As contents. *Environ. Pollut.*, **211**, 1–8.
19. Nayfach,S. and Pollard,K.S. (2016) Toward accurate and quantitative comparative metagenomics. *Cell*, **166**, 1103–1116.
20. Makarova,K.S., Wolf,Y.I. and 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**, 818–840.
21. Galperin,M.Y., Wolf,Y.I., Makarova,K.S., Vera Alvarez,R., Landsman,D. and Koonin,E.V. (2021) COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.*, **49**, D274–D281.

22. 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. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.
23. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
24. Tu,Q., Lin,L., Cheng,L., Deng,Y. and He,Z. (2019) NCycDB: a curated integrative database for fast and accurate metagenomic profiling of nitrogen cycling genes. *Bioinformatics*, **35**, 1040–1048.
25. Yu,X., Zhou,J., Song,W., Xu,M., He,Q., Peng,Y., Tian,Y., Wang,C., Shu,L., Wang,S. *et al.* (2021) SCycDB: a curated functional gene database for metagenomic profiling of sulphur cycling pathways. *Mol. Ecol. Resour.*, **21**, 924–940.
26. Chen,S.-C., Sun,G.-X., Yan,Y., Konstantinidis,K.T., Zhang,S.-Y., Deng,Y., Li,X.-M., Cui,H.-L., Musat,F., Popp,D. *et al.* (2020) The great oxidation event expanded the genetic repertoire of As metabolism and cycling. *Proc. Natl Acad. Sci. USA*, **117**, 10414–10421.
27. The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*,**49**, D480–D489.
28. Edgar,RC. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460.
29. Shen,W. and Xiong,J. (2021) TaxonKit: a practical and efficient NCBI taxonomy toolkit. *J. Genet. Genomics*, **48**, 844–850.
30. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
31. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
32. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
33. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
34. Peng,Y., Leung,H.C.M., Yiu,S.M. and Chin,F.Y.L. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420–1428.
35. Rosen,B.P. (2002) Biochemistry of As detoxification. *FEBS Lett.*, **529**, 86–92.
36. Hemmingsson,O., Zhang,Y., Still,M. and Naredi,P. (2009) ASNA1, an ATPase targeting tail-anchored proteins, regulates melanoma cell growth and sensitivity to cisplatin and arsenite. *Cancer Chemother. Pharmacol.*, **63**, 491–499.
37. Kurdi-Haidar,B., Aebi,S., Heath,D., Enns,R.E., Naredi,P., Hom,D.K. and Howell,S.B. (1996) Isolation of the ATP-binding human homolog of the arsA component of the bacterial arsenite transporter. *Genomics*, **36**, 486–491.
38. Tamaki,S. and Frankenberger,W.T. (1992) Environmental biochemistry of As. In: Ware,G.W. (ed). *Reviews of Environmental Contamination and Toxicology: Continuation of Residue Reviews*. Springer, NY, pp. 79–110.
39. Ali,W., Isayenkov,S.V., Zhao,F.-J. and Maathuis,F.J.M. (2009) Arsenite transport in plants. *Cell. Mol. Life Sci.*, **66**, 2329–2339.
40. Bobrowicz,P., Wysocki,R., Owsianik,G., Goffeau,A. and Ułaszewski,S. (1997) Isolation of three contiguous genes, ACR1, ACR2 and ACR3, involved in resistance to As compounds in the yeast *Saccharomyces cerevisiae*. *Yeast*, **13**, 819–828.
41. Chen,J., Yoshinaga,M., Garbinski,L.D. and Rosen,B.P. (2016) Synergistic interaction of glyceraldehydes-3-phosphate dehydrogenase and ArsJ, a novel organoarsenical efflux permease, confers arsenate resistance. *Mol. Microbiol.*, **100**, 945–953.
42. Chen,J., Madegowda,M., Bhattacharjee,H. and Rosen,B.P. (2015) ArsP: a methylarsenite efflux permease. *Mol. Microbiol.*, **98**, 625–635.
43. Afkar,E., Lisak,J., Saltikov,C., Basu,P., Oremland,R.S. and Stolz,J.F. (2003) The respiratory arsenate reductase from *Bacillus selenitireducens* strain MLS10. *FEMS Microbiol. Lett.*, **226**, 107–112.
44. Krafft,T. and Macy,J.M. (1998) Purification and characterization of the respiratory arsenate reductase of *Chrysiogenes arsenatis*. *Eur. J. Biochem.*, **255**, 647–653.
45. Héry,M., Gault,A.G., Rowland,H.A.L., Lear,G., Polya,D.A. and Lloyd,J.R. (2008) Molecular and cultivation-dependent analysis of metal-reducing bacteria implicated in As mobilisation in south-east Asian aquifers. *Appl. Geochem.*, **23**, 3215–3223.
46. Basu,P., Stolz,J.F. and Oremland,R.S. (2010) Microbial As metabolism: new twists on an old poison: during the early anoxic phase on earth, some microbes depended on As to respire. *Microbe Mag.*, **5**, 53–59.
47. Eman,A., Joy,L., Chad,S., Partha,B., Oremland,R.S. and Stolz,J.F. (2010) The respiratory arsenate reductase from *Bacillus selenitireducens* strain mls10. *FEMS Microbiol. Lett.*, **226**, 107–112.
48. Mukhopadhyay,R. and Rosen,B.P. (2002) Arsenate reductases in prokaryotes and eukaryotes. *Environ. Health Perspect.*, **110**, 745–748.
49. Chrysostomou,C., Quandt,E.M., Marshall,N.M., Stone,E. and Georgiou,G. (2015) An alternate pathway of arsenate resistance in *E. coli* mediated by the glutathione S-transferase gstB. *ACS Chem. Biol.*, **10**, 875–882.
50. Bhattacharjee,H., Sheng,J., Ajees,A.A., Mukhopadhyay,R. and Rosen,B.P. (2010) Adventitious arsenate reductase activity of the catalytic domain of the human Cdc25B and Cdc25C phosphatases. *Biochemistry*, **49**, 802–809.
51. A,R., Dc,P., U,G. and H,M. (2012) The modulator of the general stress response, MgsR, of *Bacillus subtilis* is subject to multiple and complex control mechanisms. *Environ. Microbiol.*, **14**, 2838–2850.
52. Qin,J., Fu,H.-L., Ye,J., Bencze,K.Z., Stemmler,T.L., Rawlings,D.E. and Rosen,B.P. (2007) Convergent evolution of a new As binding site in the ArsR/SmtB family of metalloregulators. *J. Biol. Chem.*, **282**, 34346–34355.
53. Chen,J., Nadar,V.S. and Rosen,B.P. (2017) A novel MAs(III)-selective ArsR transcriptional repressor. *Mol. Microbiol.*, **106**, 469–478.
54. Hamamura,N., Macur,R.E., Korf,S., Ackerman,G., Taylor,W.P., Kozubal,M., Reysenbach,A.-L. and Inskeep,W.P. (2009) Linking microbial oxidation of As with detection and phylogenetic analysis of arsenite oxidase genes in diverse geothermal environments. *Environ. Microbiol.*, **11**, 421–431.
55. Kulp,T.R. (2014) As and primordial life. *Nat. Geosci.*, **7**, 785–786.
56. Sardiwal,S., Santini,J.M., Osborne,T.H. and Djordjevic,S. (2010) Characterization of a two-component signal transduction system that controls arsenite oxidation in the chemolithoautotroph NT-26. *FEMS Microbiol. Lett.*, **313**, 20–28.
57. Zargar,K., Hoeft,S., Oremland,R. and Saltikov,C.W. (2010) Identification of a novel arsenite oxidase gene, arxA, in the haloalkaliphilic, arsenite-oxidizing bacterium *Alkalilimnicola ehrlichii* strain MLHE-1. *J. Bacteriol.*, **192**, 3755–3762.
58. Zargar,K., Conrad,A., Bernick,D.L., Lowe,T.M., Stolc,V., Hoeft,S., Oremland,R.S., Stolz,J. and Saltikov,C.W. (2012) ArxA, a new clade of arsenite oxidase within the DMSO reductase family of molybdenum oxidoreductases. *Environ. Microbiol.*, **14**, 1635–1645.
59. Qin,J., Rosen,B.P., Zhang,Y., Wang,G., Franke,S. and Rensing,C. (2006) As detoxification and evolution of trimethylarsine gas by a microbial arsenite S-adenosylmethionine methyltransferase. *Proc. Natl Acad. Sci. USA*, **103**, 2075–2080.
60. Chen,J., Bhattacharjee,H. and Rosen,B.P. (2015) ArsH is an organoarseical oxidase that confers resistance to trivalent forms of the herbicide monosodium methylarsenate and the poultry growth promoter roxarsone. *Mol. Microbiol.*, **96**, 1042–1052.
61. Wang,P., Sun,G., Jia,Y., Meharg,A.A. and Zhu,Y. (2014) A review on completing As biogeochemical cycle: microbial volatilization of arsines in environment. *J. Environ. Sci.*, **26**, 371–381.
62. Yan,Y., Ye,J., Xue,X.-M. and Zhu,Y.-G. (2015) As demethylation by a C·As lyase in *Cyanobacterium nostoc* sp. PCC 7120. *Environ. Sci. Technol.*, **49**, 14350–14358.
63. Yoshinaga,M., Cai,Y. and Rosen,B.P. (2011) Demethylation of methylarsonic acid by a microbial community. *Environ. Microbiol.*, **13**, 1205–1215.
64. Xu,R., Huang,D., Sun,X., Zhang,M., Wang,D., Yang,Z., Jiang,F., Gao,P., Li,B. and Sun,W. (2021) Diversity and metabolic potentials of As(III)-oxidizing bacteria in activated sludge. *Appl. Environ. Microbiol.*, **87**, e0176921.
65. Cai,L., Yu,K., Yang,Y., Chen,B., Li,X. and Zhang,T. (2013) Metagenomic exploration reveals high levels of microbial As metabolism genes in activated sludge and coastal sediments. *Appl. Microbiol. Biotechnol.*, **97**, 9579–9588.

66. Zhu,Y.-G., Yoshinaga,M., Zhao,F.-J. and Rosen,B.P. (2014) Earth abides As biotransformations. *Annu. Rev. Earth Planet. Sci.*, **42**, 443–467.
67. Quince,C., Walker,A.W., Simpson,J.T., Loman,N.J. and Segata,N. (2017) Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.*, **35**, 833–844.
68. Dunivin,T.K., Yeh,S.Y. and Shade,A. (2019) A global survey of As-related genes in soil microbiomes. *BMC Biology*, **17**, 45.
69. Liu,G., Liu,M., Kim,E.-H., Maaty,W.S., Bothner,B., Lei,B., Rensing,C., Wang,G. and McDermott,T.R. (2012) A periplasmic arsenite-binding protein involved in regulating arsenite oxidation: arsenite-binding protein. *Environ. Microbiol.*, **14**, 1624–1634.
70. Heinrich-Salmeron,A., Cordi,A., Brochier-Armanet,C., Halter,D., Pagnout,C., Abbaszadeh-Fard,E., Montaut,D., Seby,F., Bertin,P.N. and Bauda,P. (2011) Unsuspected diversity of arsenite-oxidizing bacteria revealed by a widespread distribution of the aoxb gene in prokaryotes. *Appl. Environ. Microbiol.*, **77**, 4685–4692.
71. Andrewes,P., Cullen,W.R. and Polishchuk,E. (2000) Arsenic and antimony biomethylation by *Scopulariopsis brevicaulis*: interaction of arsenic and antimony compounds. *Environ. Sci.Technol.,***34**, 2249–2253.
72. Li,J., Wang,Q., Oremland,R.S., Kulp,T.R., Rensing,C. and Wang,G. (2016) Microbial antimony biogeochemistry: enzymes, regulation, and related metabolic pathways. *Appl. Environ. Microbiol.*, **82**, 5482–5495.
73. Sun,W., Sun,X., Haggblom,M.M., Kolton,M., Lan,L. and Li,B. (2021) Identification of antimonate reducing bacteria and their potential metabolic traits by the combination of stable isotope probing and metagenomic-pangenomic analysis. *Environ. Sci. Technol.*, **55**, 13902–13912.
74. 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. *et al.* (2018) Structure and function of the global topsoil microbiome. *Nature*, **560**, 233–237.
75. Saltikov,C.W. and Newman,D.K. (2003) Genetic identification of a respiratory arsenate reductase. *Proc. Natl Acad. Sci. USA*, **100**, 10983–10988.
76. Fahy,A., Giloteaux,L., Bertin,P., Le Paslier,D., Médigue,C., Weissenbach,J., Duran,R. and Lauga,B. (2015) 16S rRNA and As-related functional diversity: contrasting fingerprints in As-rich sediments from an acid mine drainage. *Microb. Ecol.*, **70**, 154–167.