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

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

[1]Zhejiang University
[2]Zhejiang University College of Environmental and Resource Sciences
[3]Research Centre for Eco-Environmental Sciences Chinese Academy of Sciences
[4]Affiliation not available

April 12, 2022

Abstract

Arsenic (As) is the most ubiquitous toxic metalloid in nature. Microbe mediated As metabolism plays an important role in the 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 414,773 representative sequences from 59 As metabolism gene families, which are affiliated with 1,653 microbial genera from 46 phyla. We then applied AsgeneDB for functional and taxonomic profiling of As metabolism in metagenomes from various habitats (freshwater, hot spring, marine sediment, and soil). Compared with other databases, AsgeneDB substantially improved the mapping ratio of short read in metagenomes from various environments. Our results indicate that the diversity and importance of microbial arsenic metabolism in the environment remains to be explored. In addition, we developed an R package Asgene to facilitate the analysis and statistical of metagenomic data. AsgeneDB and the associated R Package Asgene will greatly promote the study of arsenic metabolism in microbial communities in various environments.

AsgeneDB:A curated orthology arsenic metabolism gene database and computational tool for metagenomeannotation

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], Bin Ma[1,2,3*]

[1]Institute of Soil and Water Resources and Environmental Science, College of Environmental and Resource Sciences, Zhejiang University 310000, China

[2]Zhejiang Provincial Key Laboratory of Agricultural Resources and Environment, Zhejiang University 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

[5]Key Laboratory of Soil Environment and Pollution Remediation, Institute of Soil Science, Chinese Academy of Science, Nanjing 210000, China

*Corresponding author: Tel: +8613282198979; Email: *bma@zju.edu.cn*

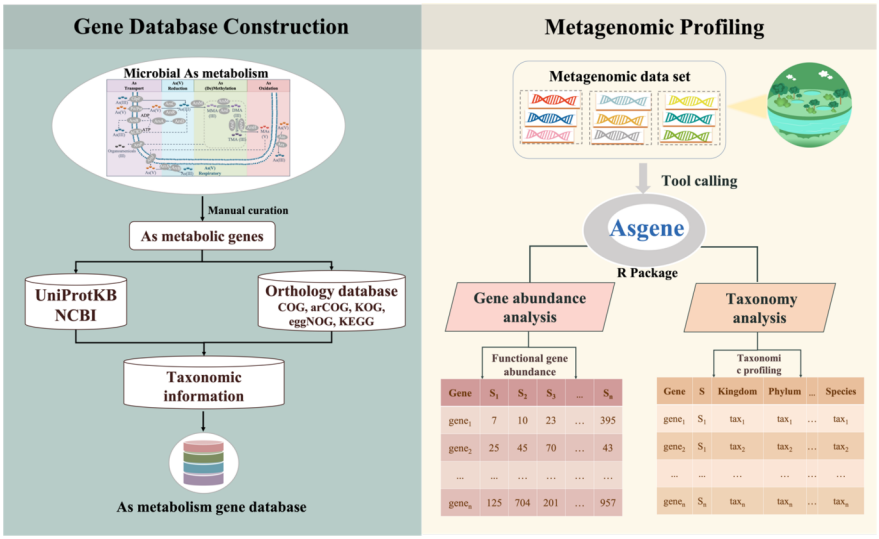
Abstract

Arsenic (As) is the most ubiquitous toxic metalloid in nature. Microbe mediated As metabolism plays an important role in the 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 414,773 representative sequences from 59 As metabolism gene families, which are affiliated with 1,653 microbial genera from 46 phyla. We then applied AsgeneDB for functional and taxonomic profiling of As metabolism in metagenomes from various habitats (freshwater, hot spring, marine sediment, and soil). Compared with other databases, AsgeneDB substantially improved the mapping ratio of short read in metagenomes from various environments. Our results indicate that the diversity and importance of microbial arsenic metabolism in the environment remains to be explored. In addition, we developed an R package *Asgene* to facilitate the analysis and statistical of metagenomic data. AsgeneDB and the associated R Package *Asgene* will greatly promote the study of arsenic metabolism in microbial communities in various environments.

Keywords

Microbiome; Arsenic metabolism; Gene database; R package; Metagenome annotation

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" (Zheng, 2020). As has therefore been a prime focus of ecology and environmental sciences (S.-Y. Zhang et al., 2017; Zheng, 2020). 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 (Achour, Bauda, & Billard, 2007; Oremland & Stolz, 2003). These metabolic processes play a major role in the global As cycle through microbial oxidation, respiration, reduction, and methylation (Mukhopadhyay, Rosen, Phung, & Silver, 2002) and are mediated by a variety of genes. It has been reported that almost all microorganisms have As resistance and metabolism genes (Zhu, Xue, Kappler, Rosen, & Meharg, 2017). 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) [7–9] while As(III) S-adenosine methionine methyltransferase (*arsM)* and nonheme iron-dependent dioxygenase (*arsI)* with C-As lyase activity catalyze As methylation and demethylation (Jia et al., 2013a; Yoshinaga & Rosen, 2014). Mechanisms involved in As metabolism can also be coopted from other processes with As(III) and As(V) acting as analogues of glycerol and phosphate, allowing microbial uptake through glycerol transporters (*GlpF*) and phosphate transporters (*Pit/Pst*) (Borgnia, Nielsen, Engel, & Agre, 1999; Wysocki et al., 2001). 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) (Achour et al., 2007; Cai, Liu, Rensing, & Wang, 2009; H.-T. Wang et al., 2019; C. Zhang et al., 2021). These methods are limited by their low throughput that only targets one or several specific genes and also by nonspecific 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 microbiome 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 (Xiao et al., 2016; S.-Y. Zhang et al., 2017). 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 orthology database (Nayfach & Pollard, 2016).

Orthology databases such as arCOG (Archaeal Clusters of Orthologous Genes) (Nayfach & Pollard, 2016), COG (Clusters of Orthologous Groups) (Galperin et al., 2021), eggNOG (Huerta-Cepas et al., 2019) and KEGG (Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016) 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 (Makarova, Wolf, & Koonin, 2015), COG and eggNOG for annotation of orthologous groups (Galperin et al., 2021; Huerta-Cepas et al., 2019) and KEGG for linking genes with pathways (Kanehisa et al., 2016). 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 (Tu, Lin, Cheng, Deng, & He, 2019; Yu et al., 2021). 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.

In this study, to understand the microbial community of As metabolism in the environment, we developed a manually curated As metabolism gene database (AsgeneDB), which covers five As metabolic pathways (transport, respiratory, reduction, oxidative and methylation/demethylation processes), 59 As metabolism gene families and 414,773 representative sequences. AsgeneDB integrates multiple lineal homology databases, including 46 phyla and 1,653 genera of bacteria, archaea and fungi. AsgeneDB enables researchers to directly study newly discovered arsenic metabolic pathways and gene families, allowing high specificity, comprehensiveness, representativeness, and accuracy. 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 was compared with five other orthology databases by analyzing metagenomic sequencing data from four habitats (freshwater, hot spring, marine sediment and soil). Our results show that AsgeneDB could detect more As metabolism genes and abundance in environmental microorganisms. In addition, there were significant differences in the abundance of As functional genes and functional driving species in microbial communities among different habitats. As transport genes are more abundant in the microbial communities than As(V) respiration, methylation, and demethylation genes. These results demonstrate the vast diversity and importance of microbial As metabolism functions in the environment that remain to be explored. Therefore, AsgeneDB and *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 (Tu et al., 2019; Yu et al., 2021). Firstly, the core database was manually constructed based on the current knowledge and literature of As metabolism (S.-C. Chen et al., 2020; H.-T. Wang et al., 2019; C. Zhang et al., 2021; Zhu et al., 2017). As metabolic genes in KEGG were also referenced (Kanehisa et al., 2016). Target sequences were downloaded from the Swiss-Prot and TrEMBL databases (The UniProt Consortium, 2017) 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, a self-vs.-self usearch (version 11.0, 30% global identity cutoff) was then performed to generate a distance matrix between different sequences. A nearest neighbor clustering procedure was then carried out to cluster sequences into groups. The outlier sequences were then checked again to confirm their annotation information in Swiss-Prot and TrEMBL 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 (Tu et al., 2019). In addition, corresponding sequences (As metabolic gene families) from 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 full database against NCBI RefSeq (options: -evalue 1e-6 -id 60). Complete taxonomic level information of sequences was determined used TaxonKit (Shen & Xiong, 2019). 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 (Fu, Niu, Zhu, Wu, & Li, 2012) 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 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., 2021), eggNOG (Huerta-Cepas et al., 2019) and KEGG (Kanehisa et al., 2016). The microbial NCBI RefSeq database (O'Leary et al., 2016) was used to enrich AsgeneDB and for taxonomically classifying microbial communities of As metabolism.

Metagenomic profiling of Asmetabolic genes

To facilitate user operation, an R Package (*Asgene*) is provided for metagenomic alignment (nucleic acid or protein sequence), subsequent gene family abundance statistics, and sample abundance standardization. The *Asgene* Package is available on github (*https://github.com/XinweiSong/Asgene*). Users only need to

choose a database search tool according to their needs (e.g., USEARCH, BLAST or DIAMOND) and input several parameters (e.g., working path, search parameters of tool and filetype) to automatically analyze statistics and output statistical results. 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). Our work can be used to analyze metagenomic data, providing functional profiles at the gene family level and composition of functional microbial community at various classification levels in different environments.

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) (Table S2). Raw reads were quality-controlled using Trimmomatic v2.39 (Bolger, Lohse, & Usadel, 2014) 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 (Peng, Leung, Yiu, & Chin, 2012). 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) (Buchfink, Xie, & Huson, 2015). 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 by KEGG, eggNOG, COG, arCOG, KOG and AsgeneDB using one-way analysis of variance (ANOVA) and Tukey's Honest Significant Difference (HSD).

Results

Comparing AsgeneDB against established orthology databases

To show the necessity of building a manually managed As metabolism gene database, we compared the coverage of As metabolism genes (subfamily; Figure 2) in AsgeneDB to 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 (16 gene subfamilies), followed by COC (13 gene subfamilies), eggnog (10 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 four 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*). In addition to containing more genes, the families defined by AsgeneDB were considered one homologous group in the four publicly available homologous databases. For example, both *arsB* and *acr3* are involved in arsenite efflux even though they belong to two different phylogenetic clades (Achour et al., 2007; Cai et al., 2009; Rosen, 2002). However, in KEGG, COG and eggNOG databases, arsB and *ACR3* are mixed into one orthology group (Table S3). Similarly, *arsA* , *ASNA1* and*GET3* are homologous genes (Hemmingsson, Zhang, Still, & Naredi, 2009; Kurdi-Haidar et al., 1996) that have no clear distinction in COG, KEGG and KOG. 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, representativeness and accuracy.

Summary of gene families and pathways in AsgeneDB

The 59 gene subfamilies in AsgeneDB target five As metabolic pathways (Figure 2), including 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 3). 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 analogues into microorganisms. Gene families including *arsA* , *arsB* , *aqpS* , *acr3* ,*arsF* , *arsT* , *GET3* and *ASNA1* participate in As(III) efflux. As(III) efflux systems have been intensively studied in both microbes and higher organisms (Ali, Isayenkov, Zhao, & Maathuis, 2009; Tamaki & Frankenberger, 1992; Zhu et al., 2017). In particular, the *acr3* gene family is most common in bacteria (Bobrowicz, Wysocki, Owsianik, Goffeau, & Ułaszewski, 1997). 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) (J. Chen, Yoshinaga, Garbinski, & Rosen, 2016). Meanwhile, the gene family *arsP* has been demonstrated to be an efflux system specific for trivalent organoarsenicals (J. Chen, Madegowda, Bhattacharjee, & Rosen, 2015). Since As(III) and As(V) act as analogues 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 1,498 representative sequences encoding arsenate respiratory reductase (Figure 3; Table S5). The large catalytic subunit (ArrA) and small subunit (ArrB) can form a heterodimer (ArrAB) (Afkar et al., 2003; Krafft & Macy, 1998). 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)(Afkar et al., 2003; Basu, Stolz, & Oremland, 2010). It is noteworthy that As(V) respiration and As(III) oxidation functions mainly occur the periplasm whereas As(V) reduction and As(III) methylation mainly occur in the cytoplasm (Basu et al., 2010).

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 3; Table S5). Nearly every extant microbe has ArsB or Acr3 efflux permeases for As(III) detoxification (Zhu et al., 2017). 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) (Mukhopadhyay & Rosen, 2002). 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 pum (Bhattacharjee, Sheng, Ajees, Mukhopadhyay, & Rosen, 2010; Chrysostomou, Quandt, Marshall, Stone, & Georgiou, 2015). The transcriptional repressor (ArsR) controls these ars-operons (J. Chen, Nadar, & Rosen, 2017; Qin et al., 2007).

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; Table S5). As(III) oxidizing microorganisms exist widely in nature and include both heterotrophic and chemo/photosynthetic autotrophic microorganisms (Hamamura et al., 2009). 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) (Kulp, 2014). As(III) oxidation is catalyzed by the enzyme As(III) oxidase. This enzyme is composed of two subunits, a large subunit (α) having molybdopterin and a [3Fe-4S] cluster (AioA) and a smaller subunit (β) incorporating a Rieske-type [2Fe-2S] cluster (AioB) (Hamamura et al., 2009). Both*aioS* /*aroS* /*aoxS* (sensor histidine kinase) and*aioR* /aroR/*aoxR* (transcriptional regulator) can regulate expression of aio genes via recognizing As(III) (Sardiwal, Santini, Osborne, & Djordjevic, 2010). 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 (Sardiwal et al., 2010). A new type of As(III) oxidase (*arxA*) has been discovered with both As(V) reductase and As(III) oxidase activities *in vitro* (Zargar, Hoeft, Oremland, & Saltikov, 2010). In addition to *arxA* , *arxB* , *arxC* , *arxD* and*arxH* encode for As(III) oxidation

coupled to photosynthesis (Zargar et al., 2012). 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) (Zargar et al., 2012). In addition, methylarsenite-specific oxidase ArsH can oxidize methylarsenite to methylarsenate (J. Chen, Bhattacharjee, & Rosen, 2015; Qin et al., 2006).

As (De)Methylation Pathway

Three gene families, including *arsM* , *As3mt* and *arsI* are involved in As methylation and demethylation pathways with 7,862 sequences and 24 homologous orthology groups (Figure 3; Table S5). More recent reports of methylated As show that As methylation is widespread in the environment (J. Chen, Bhattacharjee, et al., 2015; P. Wang, Sun, Jia, Meharg, & Zhu, 2014; C. Zhang et al., 2021). Methylation is catalyzed by the enzyme As(III) S-adenosylmethionine (SAM) methyltransferase (ArsM), designated as AS3MT in animals and as ArsM in microorganisms. The gene *arsI* , which catalyzes demethylation of organic As(III), was identified and characterized from the environmental isolate bacterium *Bacillus*sp. MD1 (Yoshinaga & Rosen, 2014) and from the cyanobacterium*Nostoc* sp. 7120 (Yan, Ye, Xue, & Zhu, 2015). ArsI, a nonheme iron-dependent dioxygenase with C–As lyase activity, cleaves the C–As bond in MAs(III), trivalent roxarsone, and other trivalent aromatic Asals (Yoshinaga, Cai, & Rosen, 2011). Putative ArsI orthologs were found only in bacterial species, suggesting that alternate pathways of organoarsenical demethylation might exist in other organisms (Yoshinaga & Rosen, 2014; Zhu et al., 2017).

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 1,653 genera of bacteria, archaea and fungi (Table S1). In the As transport pathway, AsgeneDB covered 33 phyla and 1,141 genera of bacteria, among which the dominant phyla were *Proteobacteria* , *Actinobacteria* ,*Firmicutes* and *Bacteroidetes (* Table S6).*Euryarchaeota*was the dominant phyla in 6 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 6 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, 6 phyla of archaea and 1 phylum of Eukaryotae. For bacteria, *Proteobacteria* , *Actinobacteria* ,*Firmicutesand* and *Bacteroidetes*represented the dominant phyla, which were consistent with the results of previous studies (Xu et al., 2021; C. Zhang et al., 2021).*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, 4 phyla of archaea and 2 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* was the dominant fungi, which also fit with previous research (Jia et al., 2013a; S.-Y. Zhang et al., 2017). 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 ofAsgeneDB 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 4 and 5). 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 four databases (1-13 in KEGG, 1-4 in eggNOG, 4-8 in COG, one in arCOG, and one in KOG) (Figure 4a). Moreover, AsgeneDB substantially increased the metagenomic mapping rates than other five databases (Figure 4b).

Arsenic metabolic functional genes and pathways varied widely in different habitats (Figure 4c). Among the five metabolic pathways, the most abundant pathway was As transport and the least abundant was As(V) respiration. Gene abundance also varied by ecosystem and by ecosystem geographical location, indicating differences in the biogeographical distribution of microbial communities (C. Zhang et al., 2021; S.-Y. Zhang et al., 2017). 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* , *Proteobacteria* in bacteria and*Euryarchaeota*in archaea detected (Figure 5). In contrast, microorganisms with As transport genes were the most diverse, correlating with the gene abundance of various metabolic pathways in the environment (Figure 4c).

D iscussion

Combined with metagenomic methods, the identification of microbial arsenic metabolism pathways and corresponding driving microbes can provide a comprehensive perspective for understanding the complexity of microbial arsenic metabolism in the environment. This study develops AsgeneDB, a manually curated orthology As metabolism gene database, for fast and accurate annotating As metabolic genes in shotgun metagenome sequence data. AsgeneDB has three major advantages over automatically generated orthology databases: precise definitions, comprehensive gene families, 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 (Galperin et al., 2021; Huerta-Cepas et al., 2019; Kanehisa et al., 2016). Precise definitions prevent the misattribution of genes to incorrect families. A typical example is *arsB* and *ACR3* , which belong to two different phylogenetic branches evolutionarily. Previous studies have demonstrated that*ACR3*and *arsB* have complementary environmental abundances (Dunivin, Yeh, & Shade, 2019), but they are rarely separated in large databases (Achour et al., 2007; Cai et al., 2009).

Secondly, the automatically generated orthology databases cover between only 2-20 gene families involved in microbial As metabolism (Figure S2) (Galperin et al., 2021; Huerta-Cepas et al., 2019; Kanehisa et al., 2016), whereas AsgeneDB covers 59 gene families with 414,773 representative sequences. AsgeneDB reflects the latest knowledge and research progress of the As metabolism research and covers gene families not included in existing databases, for example, *arsJ* , a gene family that encodes organoarsenicals resistance in microorganism (J. Chen et al., 2016), *arsP* , a gene family that encodes trivalent organoarsenicals (MAs(III)) effluents (J. Chen, Madegowda, et al., 2015), and*GstB* , a newly discovered alternative pathway to arsenate resistance in bacteria (Chrysostomou et al., 2015) and those that encode trivalent As oxidases:*aioR,arxR* , arxA and *arxB* (Liu et al., 2012; Qin et al., 2007; Zargar et al., 2012). These gene families have not been

clearly defined in other publicly available databases, but play important roles in microbial metabolism of environmental As (Zhu et al., 2017). AsgeneDB enables researchers to directly study these newly discovered gene families and metabolic pathways.

Thirdly, as the NCBI RefSeq database has been integrated into AsgeneDB (Yu et al., 2021) and 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 allows fast profiling of As metabolic microbial communities, without huge computational cost or output file size, no matter which database searching tool is used. AsgeneDB takes the 'small database' issue observed in genes into account and presents a solution for this bioinformatics problem (Tu et al., 2019) and addresses it by including homologous gene families from multiple orthology databases, thereby reducing false positives introduced by homologs (Table S4).

In this study, we used the AsgeneDB to analyze microbial As metabolism functional genes and functional species in four environments. Our results show that AsgeneDB has obvious advantages over current comprehensive databases in the detection of As functional gene families and abundance in environmental metagenomic data. Moreover, our results also demonstrate that As metabolism genes *aioA* ,*arrA* , and *arxA* are phylogenetically conserved (Dunivin et al., 2019). *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* (Table S6). Furthermore, microorganisms extensively metabolize As in natural ecosystems. 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 (Dunivin et al., 2019). The genes *arrA* and *arrB* encode arsenate reductases that function in anaerobic environments (Saltikov & Newman, 2003), so they are more abundant in water and marine sediment.

In addition to the species previously shown to have As(III) oxidation function (C. Zhang et al., 2021), we find that *Chlamydiae* , *Thermotogae* ,*Ignavibacteriae* , and *Aquificae* also have As(III) oxidation functions in specific ecosystems. In addition to previous studies [such as (Jia et al., 2013b; C. Zhang et al., 2021; S.-Y. Zhang et al., 2017)], *Verrucomicrobia* , *Spirochaetes* ,*Ignavibacteriae* and*Candidatus Bipolaricaulota* were found to have an As methylation function (Figure 5). Our results also show that there are significant differences in the abundance of functional genes and 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 (Dunivin et al., 2019), many As metabolism genes, including *ACR3* , *arsB* ,*arsD* , *arsM* and*aioA* , have regional dispersal limitations (Dunivin et al., 2019; Fahy et al., 2015). However, the distribution and diversity of large-scale As metabolism genes remain to be further explored. AsgeneDB and 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. AsgeneDB and *Asgene* Package include comprehensive information on microbial As metabolism and will be updated periodically.

Availability of data and materials

*Asgene*Package are available on the github (https://github.com/XinweiSong/Asgene). AsgeneDB files can be downloaded from cyverse (https://de.cyverse.org/data/ds /iplant/home/xinwei/AsgeneDB/AsgeneDB.zip).

Abbreviations

As(arsenic)

Acknowledgement

Not applicable.

Funding Sources

This worked was supported by the National Foundation of China (419931334, 42090060) and the Zhejiang Natural Science Foundation (LD19D060001).

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Author Contributions

XS and BM conceptualized and designed the study. XS wrote the R script and make the R package. XS ran the test data and improved the As database and R package. Yl and KZ provided a part of the R code. BW, YZ and YL provided some ideas to construction the As database. XS, ES and BM wrote the manuscript, and all authors contributed and approved the final edition of the manuscript.

References

Achour, A. R., Bauda, P., & Billard, P. (2007). Diversity of arsenite transporter genes from arsenic-resistant soil bacteria. *Research in Microbiology* , *158* (2), 128–137. doi: 10.1016/j.resmic.2006.11.006

Afkar, E., Lisak, J., Saltikov, C., Basu, P., Oremland, R. S., & Stolz, J. F. (2003). The respiratory arsenate reductase from Bacillus selenitireducens strain MLS10. *FEMS Microbiology Letters* ,*226* (1), 107–112. doi: 10.1016/S0378-1097(03)00609-8

Ali, W., Isayenkov, S. V., Zhao, F.-J., & Maathuis, F. J. M. (2009). Arsenite transport in plants. *Cellular and Molecular Life Sciences: CMLS* , *66* (14), 2329–2339. doi: 10.1007/s00018-009-0021-7

Basu, P., Stolz, J. F., & Oremland, R. S. (2010). Microbial Arsenic Metabolism: New Twists on an Old Poison: During the early anoxic phase on Earth, some microbes depended on arsenic to respire. *Microbe Magazine* , *5* (2), 53–59. doi: 10.1128/microbe.5.53.1

Bhattacharjee, H., Sheng, J., Ajees, A. A., Mukhopadhyay, R., & Rosen, B. P. (2010). Adventitious Arsenate Reductase Activity of the Catalytic Domain of the Human Cdc25B and Cdc25C Phosphatases. *Biochemistry* ,*49* (4), 802–809. doi: 10.1021/bi9019127

Bobrowicz, P., Wysocki, R., Owsianik, G., Goffeau, A., & Ułaszewski, S. (1997). Isolation of Three Contiguous Genes, ACR1, ACR2 and ACR3, Involved in Resistance to Arsenic Compounds in the Yeast Saccharomyces cerevisiae. *Yeast* , *13* (9), 819–828. doi: 10.1002/(SICI)1097-0061(199707)13:9<819::AID-YEA142>3.0.CO;2-Y

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* ,*30* (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Borgnia, M., Nielsen, S., Engel, A., & Agre, P. (1999). Cellular and Molecular Biology of the Aquaporin Water Channels. *Annual Review of Biochemistry* , *68* (1), 425–458. doi: 10.1146/annurev.biochem.68.1.425

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods* , *12* (1), 59–60. doi: 10.1038/nmeth.3176

Cai, L., Liu, G., Rensing, C., & Wang, G. (2009). Genes involved in arsenic transformation and resistance associated with different levels of arsenic-contaminated soils. *BMC Microbiology* , *9* (1), 4. doi: 10.1186/1471-2180-9-4

Chen, J., Bhattacharjee, H., & Rosen, B. P. (2015). ArsH is an organoarsenical oxidase that confers resistance to trivalent forms of the herbicide monosodium methylarsenate and the poultry growth promoter roxarsone. *Molecular Microbiology* , *96* (5), 1042–1052. doi: 10.1111/mmi.12988

Chen, J., Madegowda, M., Bhattacharjee, H., & Rosen, B. P. (2015). ArsP: A methylarsenite efflux permease. *Molecular Microbiology* ,*98* (4), 625–635. doi: 10.1111/mmi.13145

Chen, J., Nadar, V. S., & Rosen, B. P. (2017). A novel MAs(III)-selective ArsR transcriptional repressor. *Molecular Microbiology* , *106* (3), 469–478. doi: 10.1111/mmi.13826

Chen, J., Yoshinaga, M., Garbinski, L. D., & Rosen, B. P. (2016). Synergistic interaction of glyceraldehydes-3-phosphate dehydrogenase and ArsJ, a novel organoarsenical efflux permease, confers arsenate resistance. *Molecular Microbiology* , *100* (6), 945–953. doi: 10.1111/mmi.13371

Chen, S.-C., Sun, G.-X., Yan, Y., Konstantinidis, K. T., Zhang, S.-Y., Deng, Y., ... Zhu, Y.-G. (2020). The Great Oxidation Event expanded the genetic repertoire of arsenic metabolism and cycling. *Proceedings of the National Academy of Sciences* , *117* (19), 10414–10421. doi: 10.1073/pnas.2001063117

Chrysostomou, C., Quandt, E. M., Marshall, N. M., Stone, E., & Georgiou, G. (2015). An Alternate Pathway of Arsenate Resistance in E. coli Mediated by the Glutathione S-Transferase GstB. *ACS Chemical Biology* , *10* (3), 875–882. doi: 10.1021/cb500755j

Dunivin, T. K., Yeh, S. Y., & Shade, A. (2019). A global survey of arsenic-related genes in soil microbiomes. *BMC Biology* ,*17* (1), 45. doi: 10.1186/s12915-019-0661-5

Fahy, A., Giloteaux, L., Bertin, P., Le Paslier, D., Médigue, C., Weissenbach, J., ... Lauga, B. (2015). 16S rRNA and As-Related Functional Diversity: Contrasting Fingerprints in Arsenic-Rich Sediments from an Acid Mine Drainage. *Microbial Ecology* , *70* (1), 154–167. doi: 10.1007/s00248-014-0558-3

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. doi: 10.1093/bioinformatics/bts565

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., & Koonin, E. V. (2021). COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research* , *49* (D1), D274–D281. doi: 10.1093/nar/gkaa1018

Hamamura, N., Macur, R. E., Korf, S., Ackerman, G., Taylor, W. P., Kozubal, M., ... Inskeep, W. P. (2009). Linking microbial oxidation of arsenic with detection and phylogenetic analysis of arsenite oxidase genes in diverse geothermal environments. *Environmental Microbiology* , *11* (2), 421–431. doi: 10.1111/j.1462-2920.2008.01781.x

Hemmingsson, O., Zhang, Y., Still, M., & Naredi, P. (2009). ASNA1, an ATPase targeting tail-anchored proteins, regulates melanoma cell growth and sensitivity to cisplatin and arsenite. *Cancer Chemotherapy and Pharmacology* , *63* (3), 491–499. doi: 10.1007/s00280-008-0762-2

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., ... 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. doi: 10.1093/nar/gky1085

Jia, Y., Huang, H., Zhong, M., Wang, F.-H., Zhang, L.-M., & Zhu, Y.-G. (2013a). Microbial Arsenic Methylation in Soil and Rice Rhizosphere.*Environmental Science & Technology* , *47* (7), 3141–3148. doi: 10.1021/es303649v

Jia, Y., Huang, H., Zhong, M., Wang, F.-H., Zhang, L.-M., & Zhu, Y.-G. (2013b). Microbial Arsenic Methylation in Soil and Rice Rhizosphere.*Environmental Science & Technology* , *47* (7), 3141–3148. doi: 10.1021/es303649v

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. doi: 10.1093/nar/gkv1070

Krafft, T., & Macy, J. M. (1998). Purification and characterization of the respiratory arsenate reductase of Chrysiogenes arsenatis.*European Journal of Biochemistry* , *255* (3), 647–653. doi: 10.1046/j.1432-1327.1998.2550647.x

Kulp, T. R. (2014). Arsenic and primordial life. *Nature Geoscience* , *7* (11), 785–786. doi: 10.1038/ngeo2275

Kurdi-Haidar, B., Aebi, S., Heath, D., Enns, R. E., Naredi, P., Hom, D. K., & Howell, S. B. (1996). Isolation of the ATP-binding human homolog of the arsA component of the bacterial arsenite transporter. *Genomics* , *36* (3), 486–491. doi: 10.1006/geno.1996.0494

Liu, G., Liu, M., Kim, E.-H., Maaty, W. S., Bothner, B., Lei, B., ... McDermott, T. R. (2012). A periplasmic arsenite-binding protein involved in regulating arsenite oxidation: Arsenite-binding protein. *Environmental Microbiology* , *14* (7), 1624–1634. doi: 10.1111/j.1462-2920.2011.02672.x

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. doi: 10.3390/life5010818

Mukhopadhyay, R., & Rosen, B. P. (2002). Arsenate reductases in prokaryotes and eukaryotes. *Environmental Health Perspectives* ,*110 Suppl 5* , 745–748. doi: 10.1289/ehp.02110s5745

Mukhopadhyay, R., Rosen, B. P., Phung, L. T., & Silver, S. (2002). Microbial arsenic: From geocycles to genes and enzymes. *FEMS Microbiology Reviews* , *26* (3), 311–325. doi: 10.1111/j.1574-6976.2002.tb00617.x

Nayfach, S., & Pollard, K. S. (2016). Toward Accurate and Quantitative Comparative Metagenomics. *Cell* , *166* (5), 1103–1116. doi: 10.1016/j.cell.2016.08.007

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., ... Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* , *44* (D1), D733–D745. doi: 10.1093/nar/gkv1189

Oremland, R. S., & Stolz, J. F. (2003). The ecology of arsenic.*Science (New York, N.Y.)* , *300* (5621), 939–944. doi: 10.1126/science.1081903

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* , *28* (11), 1420–1428. doi: 10.1093/bioinformatics/bts174

Qin, J., Fu, H.-L., Ye, J., Bencze, K. Z., Stemmler, T. L., Rawlings, D. E., & Rosen, B. P. (2007). Convergent Evolution of a New Arsenic Binding Site in the ArsR/SmtB Family of Metalloregulators. *Journal of Biological Chemistry* , *282* (47), 34346–34355. doi: 10.1074/jbc.M706565200

Qin, J., Rosen, B. P., Zhang, Y., Wang, G., Franke, S., & Rensing, C. (2006). Arsenic detoxification and evolution of trimethylarsine gas by a microbial arsenite S-adenosylmethionine methyltransferase. *Proceedings*

of the National Academy of Sciences , *103* (7), 2075–2080. doi: 10.1073/pnas.0506836103

Rosen, B. P. (2002). Biochemistry of arsenic detoxification. *FEBS Letters* , *529* (1), 86–92. doi: 10.1016/S0014-5793(02)03186-1

Saltikov, C. W., & Newman, D. K. (2003). Genetic identification of a respiratory arsenate reductase. *Proceedings of the National Academy of Sciences* , *100* (19), 10983–10988. doi: 10.1073/pnas.1834303100

Sardiwal, S., Santini, J. M., Osborne, T. H., & Djordjevic, S. (2010). Characterization of a two-component signal transduction system that controls arsenite oxidation in the chemolithoautotroph NT-26. *FEMS Microbiology Letters* , *313* (1), 20–28. doi: 10.1111/j.1574-6968.2010.02121.x

Shen, W., & Xiong, J. (2019). *TaxonKit: A cross-platform and efficient NCBI taxonomy toolkit* [Preprint]. Bioinformatics. doi: 10.1101/513523

Tamaki, S., & Frankenberger, W. T. (1992). Environmental Biochemistry of Arsenic. In G. W. Ware (Ed.), *Reviews of Environmental Contamination and Toxicology: Continuation of Residue Reviews* (pp. 79–110). New York, NY: Springer. doi: 10.1007/978-1-4612-2864-6_4

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. doi: 10.1093/bioinformatics/bty741

The UniProt Consortium. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research* , *45* (D1), D158–D169. doi: 10.1093/nar/gkw1099

Wang, H.-T., Zhu, D., Li, G., Zheng, F., Ding, J., O'Connor, P. J., … Xue, X.-M. (2019). Effects of Arsenic on Gut Microbiota and Its Biotransformation Genes in Earthworm Metaphire sieboldi.*Environmental Science & Technology* , *53* (7), 3841–3849. doi: 10.1021/acs.est.8b06695

Wang, P., Sun, G., Jia, Y., Meharg, A. A., & Zhu, Y. (2014). A review on completing arsenic biogeochemical cycle: Microbial volatilization of arsines in environment. *Journal of Environmental Sciences* ,*26* (2), 371–381. doi: 10.1016/S1001-0742(13)60432-5

Wysocki, R., Chéry, C. C., Wawrzycka, D., Van Hulle, M., Cornelis, R., Thevelein, J. M., & Tamás, M. J. (2001). The glycerol channel Fps1p mediates the uptake of arsenite and antimonite in Saccharomyces cerevisiae. *Molecular Microbiology* , *40* (6), 1391–1401. doi: 10.1046/j.1365-2958.2001.02485.x

Xiao, K.-Q., Li, L.-G., Ma, L.-P., Zhang, S.-Y., Bao, P., Zhang, T., & Zhu, Y.-G. (2016). Metagenomic analysis revealed highly diverse microbial arsenic metabolism genes in paddy soils with low-arsenic contents. *Environmental Pollution* , *211* , 1–8. doi: 10.1016/j.envpol.2015.12.023

Xu, R., Huang, D., Sun, X., Zhang, M., Wang, D., Yang, Z., … Sun, W. (2021). Diversity and metabolic potentials of As(III)-oxidizing bacteria in activated sludge. *Applied and Environmental Microbiology* . doi: 10.1128/AEM.01769-21

Yan, Y., Ye, J., Xue, X.-M., & Zhu, Y.-G. (2015). Arsenic Demethylation by a C·As Lyase in Cyanobacterium Nostoc sp. PCC 7120.*Environmental Science & Technology* , *49* (24), 14350–14358. doi: 10.1021/acs.est.5b03357

Yoshinaga, M., Cai, Y., & Rosen, B. P. (2011). Demethylation of methylarsonic acid by a microbial community. *Environmental Microbiology* , *13* (5), 1205–1215. doi: 10.1111/j.1462-2920.2010.02420.x

Yoshinaga, M., & Rosen, B. P. (2014). A C[?]As lyase for degradation of environmental organoarsenical herbicides and animal husbandry growth promoters. *Proceedings of the National Academy of Sciences* ,*111* (21), 7701–7706. doi: 10.1073/pnas.1403057111

Yu, X., Zhou, J., Song, W., Xu, M., He, Q., Peng, Y., … He, Z. (2021). SCycDB: A curated functional gene database for metagenomic profiling of sulphur cycling pathways. *Molecular Ecology Resources* , *21* (3),

924–940. doi: 10.1111/1755-0998.13306

Zargar, K., Conrad, A., Bernick, D. L., Lowe, T. M., Stolc, V., Hoeft, S., ... Saltikov, C. W. (2012). ArxA, a new clade of arsenite oxidase within the DMSO reductase family of molybdenum oxidoreductases.*Environmental Microbiology* , *14* (7), 1635–1645. doi: 10.1111/j.1462-2920.2012.02722.x

Zargar, K., Hoeft, S., Oremland, R., & Saltikov, C. W. (2010). Identification of a novel arsenite oxidase gene, arxA, in the haloalkaliphilic, arsenite-oxidizing bacterium Alkalilimnicola ehrlichii strain MLHE-1. *Journal of Bacteriology* , *192* (14), 3755–3762. doi: 10.1128/JB.00244-10

Zhang, C., Xiao, X., Zhao, Y., Zhou, J., Sun, B., & Liang, Y. (2021). Patterns of microbial arsenic detoxification genes in low-arsenic continental paddy soils. *Environmental Research* , *201* , 111584. doi: 10.1016/j.envres.2021.111584

Zhang, S.-Y., Su, J.-Q., Sun, G.-X., Yang, Y., Zhao, Y., Ding, J., ... Zhu, Y.-G. (2017). Land scale biogeography of arsenic biotransformation genes in estuarine wetland. *Environmental Microbiology* , *19* (6), 2468–2482. doi: 10.1111/1462-2920.13775

Zheng, Y. (2020). Global solutions to a silent poison. *Science* ,*368* (6493), 818–819. doi: 10.1126/science.abb9746

Zhu, Y.-G., Xue, X.-M., Kappler, A., Rosen, B. P., & Meharg, A. A. (2017). Linking Genes to Microbial Biogeochemical Cycling: Lessons from Arsenic. *Environmental Science & Technology* , *51* (13), 7326–7339. doi: 10.1021/acs.est.7b00689

Figures

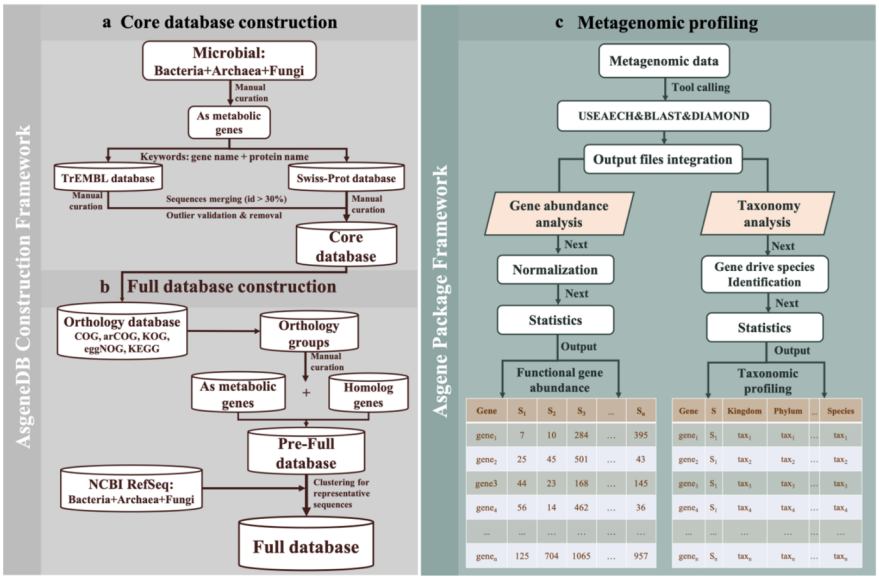


Figure 1. Framework of AsgeneDB construction. (a) Core database construction: a core database was constructed for selected gene (sub)families by retrieving protein sequences from UniProt databases using keywords (Swiss-Prot database & TrEMBL database). Sequences that failed to cluster at 30% identity 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:*Asgene*Package generates both gene abundance and taxonomic profiles of environmental samples.

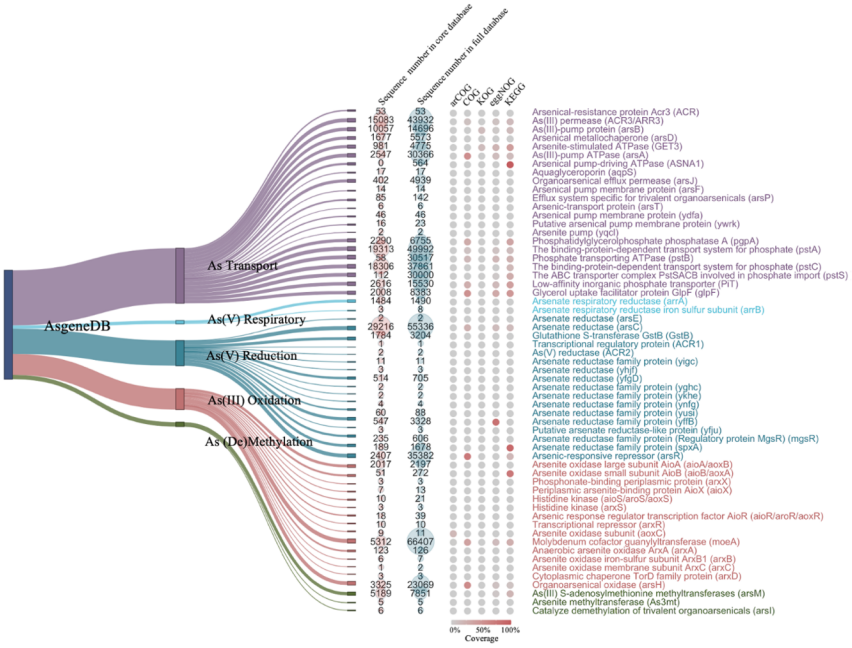


Figure 2. Summary of As metabolic gene families with representative sequences and comparison of As metabolic gene families in AsgeneDB with other public orthology databases. The heatmap represent coverage of the selected As metabolic gene families in corresponding orthology databases. AsgeneDB was used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

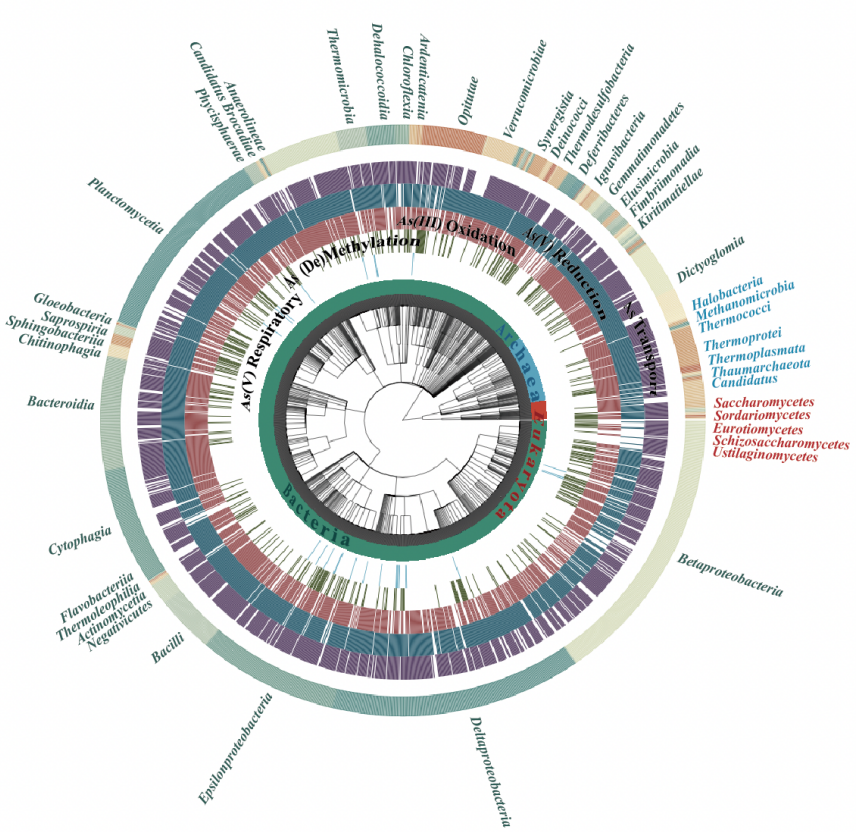


Figure 3. Phylogenetic tree of As metabolic pathways in AsgeneDB. The outermost circle shows the classification of microorganisms in AsgeneDB at the class level.

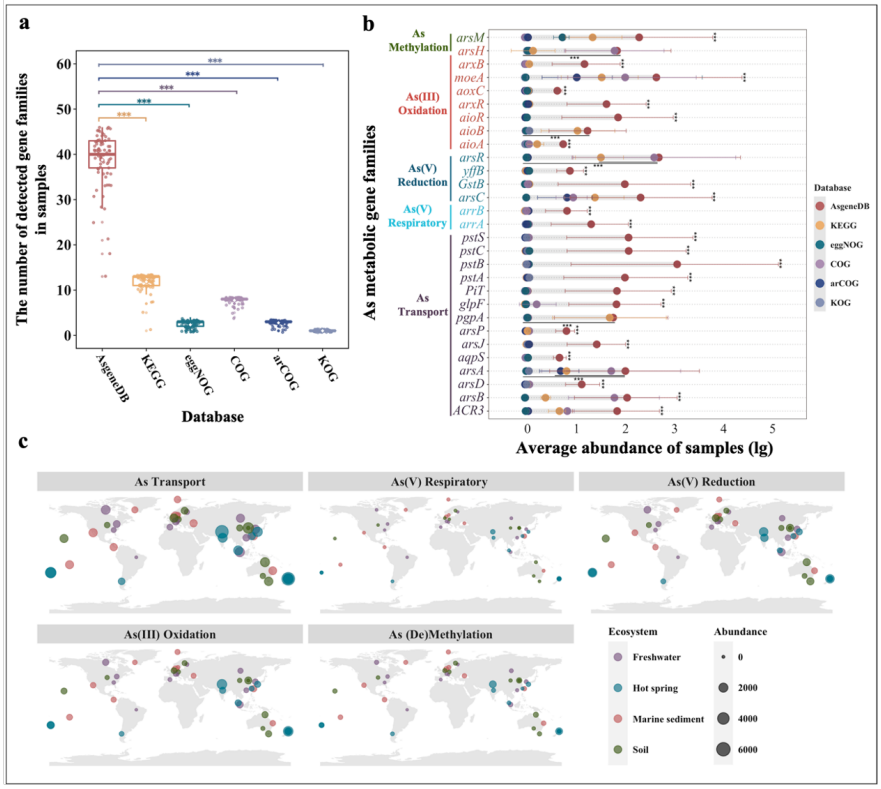


Figure 4. AsgeneDB for functional profiling of As metabolism in metagenomes from freshwater, hot spring, marine sediment, and soil (a) Comparison of the number of As metabolism gene families detected using KEGG, eggNOG, COG, arCOG, KOG and AsgeneDB in environmental samples. "***" indicates that the use of AsgeneDB is significantly different to 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 mean ± SE of all samples (standard error, n = 43). "***" indicates that the use of AsgeneDB is significantly different to the use of the other five databases ($p < 0.001$). (c) Abundances of As metabolic gene families annotated by AsgeneDB in four different habitats.

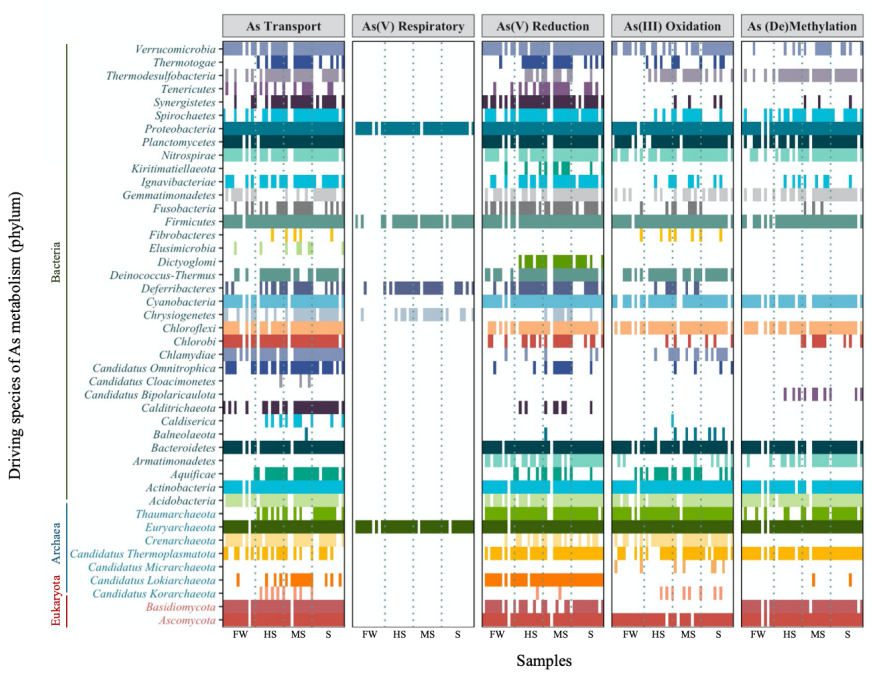


Figure 5. Microbial species driving As metabolism in microbial communities in environmental samples as annotated by AsgeneDB. FW: Freshwater; HS: Hot Spring; MS: Marine Sediment; S: Soil.

Supplemental Information

Figure S1. Microbial mediates arsenic metabolism.

Figure S2. The number of arsenic metabolism gene (sub)families detected in different databases.

Table S1. Summary of As metabolic pathways and the number of taxa covered at different taxonomic levels in AsgeneDB.

Table S2 Summary of shotgun metagenome sequencing data used in this study.

Table S3 Summary of gene families belonging to the same orthology group in public orthology databases.

Table S4 Summary of homologous gene families that reduces false positives introduced.

Table S5 Summary of As metabolic gene families with orthology groups.

Table S6 Taxonomic composition of As metabolic pathways and the number of taxa covered at different taxonomic levels in AsgeneDB.