Title:

Targeting microbial arsenic resistance genes: a new bioinformatic toolkit informs arsenic ecology and evolution in soil genomes and metagenomes

Authors

TK Dunivin[1,2], SY Yeh[3], and A Shade[1,4–7*]

Affiliations

[1]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing MI 48824

[2]Environmental and Integrative Toxicological Sciences Doctoral Program, Michigan State University, East Lansing, MI 48824, USA

[3] Institute for Cyber-Enabled Research, Michigan State University, East Lansing, MI 48824, USA

[4]Program in Ecology, Evolutionary Biology and Behavior, Michigan State University, East Lansing, MI 48824, USA

[5]Department of Plant, Soil, and Microbial Sciences, Michigan State University, East Lansing, MI 48824, USA

[6]Plant Resilience Institute, Michigan State University, East Lansing, MI, 48834, USA

[7]DOE Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI, 48834, USA

*correspondence: shadeash@msu.edu

Keywords:

functional gene, bioinformatics, targeted gene assembly, horizontal gene transfer, biogeography, phylogeny

Running title:

Targeting microbial arsenic resistance genes

Abstract

Microbial communities impact the biogeochemical cycling of arsenic, a ubiquitous and toxic metalloid that impacts human and animal health. The ecology of arsenic resistance genes in the environment is not well understood, especially within the context of diverse microbial communities. We developed an open-access bioinformatic toolkit for testing different sequencing datasets for nine arsenic resistance genes: *acr3*, *aioA*, *arsB*, *arsC* (grx), *arsC* (trx), *arsD*, *arsM*, *arrA*, and *arxA*. Our toolkit includes BLAST databases, hidden markov models and resources for gene-targeted assembly of arsenic resistance genes. We used this toolkit to examine the phylogenetic diversity, genomic location (chromosome or plasmid), and biogeography of arsenic resistance genes in genomes and metagenomes from soil, providing a synthesis of arsenic resistance from the most microbially diverse and complex environment. We found that arsenic resistance genes were common, though not universally detected, in genomes and metagenomes from soils, contradicting the common conjecture that all organisms have arsenic resistance genes. From soil genomes, we defined major clades and explored the potential for horizontal and vertical transfer of each gene family, informing their evolutionary histories. Microbial communities differed in their potential to impact arsenic biogeochemical cycling because different types of resistances were detected in different proportions, in different soils. Arsenic resistance genes were globally distributed but particular sequence variants were highly endemic (e.g., *acr3*), suggesting dispersal limitation. The gene encoding arsenic methylation *arsM* was abundant in soil metagenomes (median 48%), suggesting that it plays a previously understated and prominent role in arsenic biogeochemistry globally. Use of this flexible toolkit to understand arsenic resistance in a variety of environments will inform the foundations of microbial arsenic ecology and evolution, and provide insights into environmental potentials for arsenic biogeochemical cycling and bioremediation.

Introduction

Microbial communities drive global biogeochemical cycles through diverse functions. The biogeography of functional genes can help to predict and manage the influence of microbial communities on biogeochemical cycling [1]. These trait-based analyses require that the functional genes are well-characterized from both evolutionary and genetic perspectives [2]. An example of a suite of well-characterized functional genes that have consequence for biogeochemistry is arsenic resistance genes. Arsenic is a toxic metalloid that, upon exposure, can have negative effects for all life, including humans, livestock, and microorganisms. Microbial transformations of arsenic can have consequences for arsenic speciation and, therefore, the toxicity of arsenic in the environment.

Arsenic biogeochemical cycling by microbial communities is both an ancient [3, 4] and a contemporary [5, 6] phenomenon. Changes to the methylation or oxidation state of arsenic alters biogeochemical cycling of arsenic, and microbes have evolved a variety of mechanisms to carry out these functions. Arsenic resistance genes are generally separated into two categories: detoxification and metabolism [7]. Arsenic detoxification is encoded by the *ars* operon [8]. This includes arsenite efflux (ArsB, Acr3) which is potentially precluded by cytoplasmic arsenate reduction with either glutaredoxin (ArsC (grx)) or thioredoxin (ArsC (trx)) [8]. Arsenic metabolisms include methylation (ArsM), oxidation (AioAB, ArxAB), and dissimilatory reduction (ArrAB) [7]. While these genetic determinants of arsenic resistance are well-characterized, the full scope of arsenic resistance gene distribution, diversity, and interspecies transfer is unknown [9–11].

Microbial arsenic resistance reportedly is widespread in the environment. Arsenic resistant organisms have been found in sites with low arsenic concentrations (< 7 ppm) [12, 13], and it has been speculated that nearly all organisms have arsenic resistance genes [14]. While the number of identified microorganisms with arsenic resistance genes continues to grow [7], the number of microorganisms without arsenic resistance genes is unclear. Furthermore, though the complete arsenic biogeochemical cycle has been detected in the environment [3], the relative contributions of genes encoding detoxification and metabolism remains unknown [4]. A global, biogeographic perspective of environmental arsenic resistance genes would improve understanding of

their ecology. This would expand foundational knowledge of arsenic resistance, including local and global abundances, gene diversity, dispersal across different environments, and representations over the microbial tree of life.

Knowledge gaps concerning the ecology of microbial arsenic resistance genes are driven, in part, by numerous inconsistencies in nomenclature and detection methods. Though public microbial metagenome and genome data continue to surge, there are several practical hurdles to achieving a robust, global assessment of microbial arsenic resistance genes from this wealth of data. First, tools to detect arsenic resistance genes rely on imperfect annotation [9] and widely vary in nomenclature [15]. Next, the use of different reference databases [5, 16–19] and normalization techniques [19, 20] complicates comparisons between studies. To overcome these hurdles, we developed an open-access toolkit to examine arsenic resistance genes in microbial sequence datasets. This toolkit allowed us probe genomic and metagenomic datasets simultaneously to investigate arsenic resistance genes in soil. We first asked whether arsenic resistance genes are universal in soil-associated microorganisms. Next we asked whether genes encoding arsenic detoxification were more abundant than those encoding arsenic metabolism. Third, we asked whether estimations of arsenic resistance gene abundance are biased by cultivation efforts, as cultivation is often a research emphasis because cultivable, arsenic resistant microorganisms can be used in bioremediation [11]. Finally, we tested whether particular sequence variants of arsenic resistance genes were endemic or cosmopolitan.

Materials and Methods

Gene Selection and Functional Gene (FunGene) Database Construction

Marker genes can be used to estimate their potential to influence the arsenic biogeochemical cycle [15, 19], so we selected nine well-characterized genes: *acr3*, *aioA*, *arsB*, *arsC* (grx), *arsC* (trx), *arsD*, *arsM*, *arrA*, and *arxA*. FunGene databases [21] were constructed for the following arsenic resistance genes: *arsB*, *arsC* (grx), *arsC* (trx), *acr3*, *aioA*, *arrA*, and *arxA*. The *arxA* database was constructed with seed sequences from [5]. For all other genes, UniProt [22] was used to obtain full length, reviewed

sequences when possible. NCBI clusters of orthologous groups (COG) [23] for each gene were examined for evidence of function in the literature. All COG and UniProt sequences were aligned using MUSCLE [24]. Aligned sequences were included in a maximum likelihood tree with 50 bootstrap replications made with MEGA (v7.0,[25]). Sequences that did not cluster with known sequences and had no evidence of function were removed. A final FASTA file for each gene was submitted to the Ribosomal Database Project (RDP) to construct a FunGene database [21]. All arsenic resistance gene databases are freely available on FunGene (http://fungene.cme.msu.edu/).

Arsenic resistance genes in cultivable soil microorganisms

The RefSoil+ database [26] was used to obtain high-quality genomes (chromosomes and plasmids) from soil microorganisms in the Genomes OnLine (GOLD) database [27]. RefSoil+ chromosomes and plasmids were searched with hmmsearch [28] using HMMs from FunGene with an e-value cutoff of 10^{-10}. The top hits were analyzed in R [29]. For each gene, scores and percent alignments were plotted to determine quality cutoffs. Only hits with scores > 100, percent alignment > 90% were included. Hits against *aioA*, *arrA*, and *arxA* were further quality filtered to have scores > 1,000. When one open reading frame (ORF) contained multiple arsenic resistance gene hits, the hit with a lower score was removed. Taxonomy was assigned using the RefSoil database [30], and the relative abundance of arsenic resistance genes within phyla were examined. A 16S rRNA gene maximum likelihood tree of RefSoil+ bacterial strains was with RAxML (v.8.0.6 [31]) based on the Whelan and Goldman (WAG) model with 100 bootstrap replicates ("-m PROTGAMMAWAG -p 12345 -f a -k -x 12345 -# 100"). Based on accession numbers, arsenic resistance gene hits were extracted from RefSoil+ sequences and used to construct arsenic resistance gene maximum likelihood trees.

Reference Database Construction

Reference gene databases of diverse, near full length sequences were constructed using limited sequences from FunGene databases [21] for the following genes: *acr3*, *aioA*, *arrA*, *arsB*, *arsC* (grx), *arsC* (trx), *arsD*, *arsM*, and *arxA*. Seed

sequences and hidden markov models (HMMs) for each gene were downloaded from FunGene, and diverse protein and corresponding nucleotide sequences were selected with gene-specific search parameters (**Supplementary Table S1**). Briefly, minimum size amino acid was set to 70% of the HMM length; minimum HMM coverage was set to 80% as is recommended by Xander software for targeted gene assembly; and a score cutoff was manually selected based on a drop off point. Sequences were de-replicated before being used in subsequent analysis, and final sequence counts are included in **Supplementary Table S1**. Reference databases were converted to publicly available BLAST databases using BLAST+ [32]. Reference and BLAST databases are publicly available on GitHub (https://github.com/ShadeLab/meta_arsenic).

Sample collection and preparation

A soil surface core (20 cm depth and 5.1 cm diameter) was collected in October 2014 from Centralia, Pennsylvania (GPS coordinates: 40 48.070, 076 20.574). For cultivation-dependent work, a soil slurry was made by vortexing 5 g soil with 25 mL phosphate-buffered saline (PBS) for 1 m. Remaining soil was stored at -80°C until DNA extractions. The soil slurry was allowed to settle for 2 m. 100 µL of the slurry was then removed and serial diluted using PBS to a 10^{-2} dilution. 100 µL of the solution was added to 50% trypticase soy agar (TSA50) with 200 µg/ml cycloheximide to prevent fungal growth. Plates were incubated at 60°C for 72 h. Lawns of growth were extracted by adding 600 µL trypticase soy broth with 25% glycerol to plates. The plate scrapings were stored at -80°C until DNA extraction.

DNA extraction and metagenome sequencing

DNA for cultivation-independent analysis was manually extracted from soil using a phenol chloroform extraction [33] and the MoBio DNEasy PowerSoil Kit (MoBio, Solana Beach, CA, USA) according the manufacturer's instructions. DNA extraction for cultivation-dependent analysis was performed in triplicate from 200 µL of plate scrapings using the E.Z.N.A. Bacterial DNA Kit according to the manufacturer's instructions. All DNA was quantified using a Qubit dsDNA BR Assay Kit (Life Technologies, NY, USA) and was submitted for NGS library prep and sequencing at the

Michigan State University Genomics Core sequencing facility (East Lansing, MI, USA). Libraries were prepared using the Illumina TruSeq Nano DNA Library Preparation Kit. After QC and quantitation, the libraries were pooled and loaded on one lane of an Illumina HiSeq 2500 Rapid Run flow cell (v1). Sequencing was performed in a 2 x 150 bp paired end format using Rapid SBS reagents. Base calling was performed by Illumina Real Time Analysis (RTA) v1.18.61 and output of RTA was demultiplexed and converted to FastQ format with Illumina Bcl2Fastq v1.8.4.

Public soil metagenome acquisition

In total, 38 soil metagenomes were obtained for this work (**Supplementary Table S2**). Datasets from Centralia, PA were generated in our research group. All other metagenome data sets were obtained from MG-RAST ((http://metagenomics.anl.gov/). The MG-RAST database was searched on May 15, 2017 with the following criteria: material = soil, sequence type = shotgun, public = true. The resulting list of metagenome data sets were ordered by number of base pairs (bp). Metagenomic data sets with the most bp were only included if they were sequenced using Illumina to standardize sequencing errors, had an available FASTQ file for internal quality control, and contained < 30% low quality as determined by MG-RAST. Within high quality Illumina samples, priority for inclusion was given to projects with multiple samples. When a project had multiple samples, data sets with the greatest bp were selected. This search yielded 26 data sets from 12 locations and five countries (**Supplementary Table S2**).

Soil metagenome processing and gene targeted assembly

Sequences from MG-RAST data sets as well as Centralia sample Cen13 were quality controlled using the FASTX toolkit (fastq_quality_filter, "-Q33 -q 30 -p 50"). Twelve datasets from Centralia, PA, were obtained from the Joint Genome Institute and quality filtered as described previously [34]. Quality filtered sequences were used in all downstream analyses. For each data set, a gene targeted metagenome assembler [35] was used to assemble each gene of interest. For each gene of interest, seed sequences, HMMs, and reference gene databases described above were included. For

rplB, reference gene database, seed sequences, and HMMs from the Xander package were used. In most instances, default assembly parameters were used except to incorporate differences in protein length (i.e. protein is shorter than default 150 aa) or to improve quality (i.e. maximum length is increased to improve specificity) (**Supplementary Table S1**). While the assembler includes chimera removal, additional quality control steps were added. Final assembled sequences (operational taxonomic units, OTUs) were searched against the reference gene database as well as the non-redundant database (nr) from NCBI (August 28, 2017) using BLAST [32]. Genes were re-examined if the top hit had an e-value > 10^{-5} or if top hit descriptors were not the target gene. Genes with low quality results were re-assembled with adjusted parameters.

Soil metagenome comparison

To compare assembled sequences between samples, gene-based OTU tables were constructed. Aligned sequences from each sample were dereplicated and clustered at 90 amino acid identity using the RDP Classifier [36]. Dereplicated, clustered sequences were converted into OTU tables with coverage-adjusted abundance. These tables were subsequently analyzed in R [37]. RplB OTUs were used to compare community structure. The six most abundant phyla were extracted to include at least 75% of each community; the full community structure is available. To compare the abundance of arsenic resistance genes among data sets, total counts of *rplB* were used to normalize the abundance of each OTU. Relative abundance of arsenic resistance genes was also calculated for each sample.

Accession

The full arsenic resistance gene toolkit (BLAST databases, hidden markov models, and gene resources for Xander) is publicly available on GitHub (https://github.com/ShadeLab/meta_arsenic). Cultivation dependent and cultivation independent Centralia metagenomes from this study are available on NCBI under BioProject PRJNA492298. All RefSoil+ and metagenome analyses from this work are

also available on GitHub in the HMM_search and gene_targeted_assembly directories respectively.

Results and Discussion

A bioinformatic toolkit for detecting and quantifying arsenic resistance genes

We developed a toolkit to improve investigations of microbial arsenic resistance genes (**Figure 1AB**)[8, 38–42]. We selected these nine genes because they are markers of arsenic resistance [15, 19] and because their genetic underpinnings are well established. Seed sequences (high quality and full length sequences) for each gene of interest were collected and used to construct BLAST databases [43], functional gene (FunGene) databases [21], hidden markov models (HMMs; Eddy 1996), and gene resources for gene-targeted assembly (Xander; Wang *et al.* 2015) (**Figure 1A**). Altogether, this toolkit relies on consistent references and nomenclature and can search both amino acid and nucleotide sequence data.

This toolkit supports a variety of applications (**Figure 1A**): arsenic resistance genes can be detected in amino acid sequences from completed genomes (HMMs [28], BLAST [43]), nucleotide sequences in draft genomes (BLAST), as well as metagenomes and metatranscriptomes (Xander [35]). Additionally, FunGene [21] can be used for primer design and sequence browsing. Because each tool relies on the same seed sequences, there is consistency and opportunity for comparison between sequence datasets that were generated from different sources. Additionally, FunGene is frequently synced with GenBank [45], so the most recently deposited arsenic resistance genes are readily available [21].

The toolkit is scalable for additional mechanisms for arsenic resistance and other functional genes of interest (e.g., methylarsenite oxidase (ArsH), C-As lyase (ArsI), trivalent organoarsenical efflux permease (ArsP), organoarsenical efflux permease (ArsJ) [14]; or redox transformations of metals involved in arsenic biogeochemical cycling [6, 14]). This toolkit serves as both a resource and an example workflow for developing similar toolkits to examine functional gene(s), beyond arsenic resistance genes, in microbial sequence datasets.

To demonstrate the utility of our toolkit, we performed an analysis of arsenic resistance genes in soil-associated genomes and metagenomes. We used HMMs for arsenic resistance marker genes to search RefSoil+ genomes, a set of complete chromosomes and plasmids from cultivable soil microorganisms [26]. Additionally, we used a gene-targeted assembler [35] to test 38 public soil metagenomes from Brazil, Canada, Malaysia, Russia, and the United States for arsenic resistance genes (**Supplementary Table S2**). Ultimately, these data serve as a broad baseline of arsenic resistance genes in soil.

Phylogenetic distributions and genomic locations of arsenic resistance genes

It has been conjectured that nearly all organisms have arsenic resistance genes [14], and though this assumption has propagated in the literature, it had never been explicitly quantified. We asked whether arsenic resistance genes were universal in RefSoil+ organisms [26]. RefSoil+ is a reference database of complete genomes (chromosomes and plasmids) from cultivable soil organisms [26]. Of the 922 RefSoil+ genomes spanning 25 phyla (**Figure 2B**; **Supplementary Table S3**), 14.3% (132) did not contain any tested arsenic resistance genes. It is possible for these 132 organisms to have untested or novel arsenic resistance genes; however, these nine well-characterized, arsenic resistance genes were not universally detected.

Of the 25 phyla in RefSoil+, two phyla (Chlamydiae and Crenarchaeota) did not have any of these genes. These phyla, however, had few RefSoil+ representatives (three and nine, respectively), so other members of these phyla may have arsenic resistance genes. Supporting this, a Crenarchaeota isolate was previously reported to oxidize arsenic [46]. Nonetheless, these data suggest that arsenic resistance marker genes are widespread but not universal, even among cultivable soil organisms (**Figure 2**).

We next asked whether 16S rRNA gene phylogeny was predictive of arsenic resistance genotypes using a test for phylogenetic signal (Bloomberg's K [47]). No phylogenetic signal was observed for plasmid-borne sequences or genes encoding arsenic metabolisms (*aioA*, *arrA*, *arxA*); however, relatively few RefSoil+ microorganisms tested positive for these genes. Despite their phylogenetic breadth (41,

42; **Supplementary Figure S1 - S5**), chromosomally-encoded *acr3*, *arsB*, *arsC* (grx), *arsC* (trx), and *arsM* were similar between phylogenetically related organisms (false discovery rate adjusted $p < 0.01$; **Figure 2A**). This suggests that taxonomy is predictive of arsenic resistance genes despite documented potential for HGT [13, 49–52]. This could be explained by ancient rather than contemporary HGT, as seen with *arsM* [49] and *arsC* (grx) [50]. Therefore, we next assessed evidence for HGT by examining the phylogenetic congruence of arsenic resistance gene sequences.

Phylogenetic diversity of arsenic resistance genes: insights into vertical and horizontal transfer

Horizontal transfer of arsenic resistance genes has been well documented [13, 49–54] and is an important consideration for understanding the propagation and taxonomic identity of arsenic resistance genes. We examined the phylogenetic diversity of arsenic resistance genes in RefSoil+ microorganisms, including plasmids and chromosomes, and compared them with the 16S rRNA gene taxonomy.

<u>Arsenite efflux pumps</u>. We examined the phylogenetic diversity of distinct genes encoding arsenite efflux pumps, *acr3* and *arsB*, for soil-associated microorganisms (**Figure 3, Supplementary Figure S1, S2**). Gene *acr3* is separated into two clades: *acr3*(1) and *acr3*(2) [51]. Clade *acr3*(1) is typically composed of Proteobacterial sequences while *acr3*(2) is typically composed of Firmicutes and Actinobacterial sequences [15, 51, 55]. Though RefSoil+ genomes were mostly composed of *acr3*(2) sequences from Proteobacteria (**Figure 3A; Supplementary Figure S1**), we observed greater taxonomic diversity observed than previously reported for this clade [15, 51, 55]. Surprisingly, there were deep branches in *acr3*(2) that belonged to Bacteroidetes, Euryarchaeota, Firmicutes, Fusobacteria, and Verrucomicrobia. Similarly, *acr3*(1) contained closely related *acr3* sequences present in a diverse array of phyla (10 out of 25). Both clades had sequences present on plasmids (6.1%) suggesting a potential for transfer across unrelated taxa. Therefore, studies assigning taxonomy to *acr3* in the absence of host information should consider the clade precisely and proceed with caution.

Despite their functional redundancy as arsenite efflux pumps, *acr3* and *arsB* have very distinctive diversity. As compared with *acr3*, *arsB* was less diverse and more phylogenetically conserved (**Figure 3B**; **Supplementary Figure S2**). This is in agreement with previous reports comparing the diversity of *arsB* to *acr3* [51, 55]. Plasmid-borne *arsB* sequences were only present in Proteobacteria and Deinococcus-Thermus strains. Sequences from Actinobacteria, Proteobacteria, and Firmicutes were each present in two distinct phylogenetic groups, and previous studies also observe separation of *arsB* sequences based on phylum [51, 55]. Interestingly, our genome-centric analysis revealed that microorganisms with multiple copies of *arsB* did not harbor identical copies. For example, seven *Bacillus subtilis* subsp. *subtilis* strains had two copies of *arsB*, with one from each of the two clades (**Supplementary Figure S2**). This could be due to an early gene duplication and subsequent diversification or to an early transfer event. Therefore, despite relatively lower sequence variation, this *arsB* phylogeny suggests an interesting evolutionary history that could be investigated further.

<u>Cytoplasmic arsenate reductases.</u> Cytoplasmic arsenate reductase (ArsC (trx)) was phylogenetically widespread in RefSoil+ microorganisms (**Figure 4A**; **Supplementary Figure S3**). While some *arsC* (trx) sequences were plasmid-borne, the majority were chromosomally encoded. This suggests vertical transfer of plasmid-encoded *arsB* in the phylum rather than multiple transfer events. *arsC* (trx) was present in both Bacteria and Archaea, and sequences from the two Domains formed two distinct clades. This supports an early evolutionary origin for *arsC* (trx), as previously suggested [56]. Thus, *arsC* (trx) appears to be an evolutionarily old enzyme that is phylogenetically conserved despite its presence on plasmids and potential for HGT.

Unlike *arsC* (trx), the other gene encoding a cytoplasmic arsenate reductase *arsC* (grx) had no known homologs in archaea, suggesting it evolved after the divergence of Bacteria and Archaea (**Figure 4B**; **Supplementary Figure S4**). Plasmid encoded *arsC* (grx) made up 4.6% of RefSoil+ hits. This highlights a contemporary potential for HGT that has been documented in soil [50]. Notably, several Proteobacteria strains have multiple copies of *arsC* (grx) with distinct sequences. It is

possible that this is the result of an early gene duplication event or HGT of a second *arsC* (grx).

Arsenic metabolisms. The evolutionary history of the gene encoding arsenite S-adenosylmethionine methyltransferase, *arsM*, was recently investigated [49, 52]. Both studies independently determined that arsM evolved billions of years ago and was subject to HGT [49, 52]. However, *arsM* was relatively uncommon in RefSoil+ microorganisms (5.2%) (**Figure 2**). In the RefSoil+ database, *arsM* was observed in Euryarchaeota as well as several bacterial phyla Acidobacteria, Actinobacteria, Armatimonadetes, Bacteroidetes, Chloroflexi, Cyanobacteria, Firmicutes, Gemmatimonadetes, Nitrospirae, Proteobacteria, Verrucomicrobia (**Figure 5**; **Supplementary Figure S5**). Notably, only one RefSoil+ microorganism, *Rubrobacter radiotolerans* (NZ_CP007516.1), had a plasmid-borne *arsM*. However, *arsM* sequences from Euryarchaeota were dispersed throughout the arsM phylogeny, supporting the potential for inter-kingdom transfer events that were recently suggested [49, 52].

Arsenic metabolism genes *aioA*, *arrA*, and *arxA* were phylogenetically conserved (**Figure 6**). Genes encoding arsenite oxidases *aioA* and *arxA* were restricted to Proteobacteria. *aioA* sequences clustered into two clades based on class-level taxonomy: all Alphaproteobacteria sequences cluster separately from Gamma- and Betaproteobacteria sequences. The gene encoding dissimilatory arsenate reduction *arrA* was also phylogenetically conserved in RefSoil+ strains, with strains from Proteobacteria clustering separate from Firmicutes (**Figure 6**).

Cultivation bias and environmental distributions of arsenic resistance genes

To gain a cultivation-dependent perspective of the abundances of arsenic resistance genes in soils, we used inferred environmental abundances of RefSoil microorganisms [30, 57] to estimate the environmental abundances of arsenic resistance genes from the cultivated bacteria in RefSoil+. These data suggest that arsenic metabolism genes are less common in the environment compared with arsenic detoxification (**Figure 7A**). Despite similar distributions of *acr3* and *arsB* in RefSoil+ (**Figure 2B**), *acr3* was more abundant in most soil orders (**Figure 7A**). A previous study

also reported an abundance of *acr3* over *arsB* in soil and attributed this to the greater phylogenetic distribution of *acr3* compared with *arsB* [55]. Additionally, *arsC* (grx) was more abundant than *arsC* (trx) in soils. It has been posed that *arsC* (trx) is more efficient than *arsC* (grx) and that high local arsenic concentrations result in a relatively greater abundance of *arsC* (trx) [15, 58]. Our cultivation-dependent abundances suggest that *acr3* and *arsC* (grx), rather than *arsB* and *arsC* (trx), predominantly comprise the arsenic detoxification pathway in soils.

To gain a cultivation-independent perspective of the abundances of arsenic resistance genes, we examined their normalized abundance from soil metagenomes. An undetected gene does not confirm absence, so we present a conservative estimate that only includes metagenomes testing positive for a gene. As predicted by cultivable organisms, bioenergetic arsenic resistance genes (*aioA*, *arrA*, *arxA*) were generally in low abundance while *acr3* and *arsC* (grx) were in high abundance. Estimates of genes encoding arsenic detoxification (*acr3*, *arsB*, *arsD*, *arsC* (grx), *arsC* (trx)) were considerably lower in these cultivation-independent samples. This could be due, in part, to the large number of RefSoil+ microorganisms with multiple copies of these genes (**Supplementary Figure S6**). Cultivation-independent genomes (e.g., single-cell amplified genomes and metagenome-assembled genomes) could provide greater context about the environmental distributions of arsenic resistance gene copy numbers.

Notably, *arsM* was abundant in soil (median 48%), which greatly exceeds cultivation-dependent estimations. Due to the early phylogenetic origins of *arsM* and its independent functionality [49], this abundance of *arsM* in soil metagenomes is not unexpected. *arsM* is typically studied in paddy soils [59–61], but these metagenomes suggest it is an important component of the arsenic biogeochemical cycle in a variety of soils. We explored cultivation bias against detection of *arsM* with a case study comparing cultivation-dependent (lawn growth on the standard medium TSA50) and -independent communities from the same soil. Genes in the *ars* operon (*acr3*, *arsB*, *arsD*, and *arsC* (trx)) were elevated in the cultivation-dependent metagenome (**Figure 7C**). Additionally, arsenic metabolism genes were not detected (*aioA*, *arrA*, *arxA*) or in low abundance (*arsM*) in the cultivation-dependent sample; however, all four of these arsenic metabolism genes were detected in the cultivation-independent sample. Though

this is a single case study of cultivation-dependent and independent methods, these results recapitulate the general discrepancies between RefSoil+ genomes and soil metagenomes (**Figure 7B**). This bias has important implications for studies focusing on arsenic bioremediation because cultivation-dependent studies could misestimate the potential of microbiomes for arsenic detoxification and metabolism *in situ*.

Arsenic resistance gene endemism

Arsenic resistance genes are globally distributed, but their biogeography is poorly understood. Broadly, arsenic resistance genes have comparable abundance among different soils (**Figure 7AB**). The relative distributions of distinct arsenic resistance mechanisms in one site, however, are relevant for predicting the impact of microbial communities on the fate of arsenic. To understand site-specific distributions, we explored soil metagenomes from Brazil, Canada, Malaysia, Russia, and the United States (**Supplementary Table S2**). These 16 sites had differences in community membership (**Figure S7**) and arsenic resistance gene content (**Figure 8A**). Soils had different distributions of arsenic resistance genes and therefore differed in their potential impact on the biogeochemical cycling of arsenic. While *arsC* (grx) and *arsM* dominated most samples, their relative proportions varied greatly (**Figure 8A**). RefSoil+ data suggests that *arsM* can be found in Verrucomicrobia (100%, n = 2), which is of particular importance for soil metagenomes since Verrucomicrobia are often underestimated with cultivation-dependent methods [62]. The mangrove sample had the most even proportions of arsenic resistance genes (**Figure 8A**). This is driven by a high abundance of *arsC* (trx) and *arrA*. While the arsenic concentrations in this sample are unknown, mangroves are considered sources and sinks for arsenic [63–65]. This could explain the greater abundance of *arsC* (trx), which is hypothesized to be more abundant in high arsenic sites [15, 58]. Additionally, *arrA* encodes a dissimilatory arsenate reductase that functions in an anaerobic environment [41], so its greater abundance in sediment is expected.

We further examined the arsenic resistance gene abundance at individual sites. For each gene, the abundance varied greatly, but replicates within on site had similar abundances (**Figure 8B**). The majority of arsenic resistance gene sequences (99.3%)

were endemic and only found in one to two sites, but 24 sequences were detected in three or more sites (**Figure 8C; Table S4**). The majority (70.8%) of cosmopolitan sequences belonged to *arsC* (grx). This could suggest genetic migration via HGT or vertical transfer and a limited of diversification of the *arsC* (grx). Both are plausible since *arsC* (grx) is common in RefSoil+ plasmids and has low phylogenetic diversity (**Figure 4B**; **Figure S4**). Only one *aioA* and three *acr3* sequences were detected in multiple sites. This supports previous findings that most variants of *acr3* and *aioA* are endemic [66]. Additionally, all assembled *arrA* sequences in this study were site-specific. Previous work has shown *arrA* clones were station-specific in acid mine drainage [67], so we expected endemism is common for *arrA*. Thus, arsenic resistance genes, except *arsC* (grx), are generally endemic, suggesting regional dispersal limitation.

Conclusions

We developed a bioinformatic toolkit for detecting arsenic resistance genes in microbial sequence data and applied it to analyze the genomes and metagenomes from soil microorganisms. It informs hypotheses about the evolutionary histories of these genes (including potentials for vertical and horizontal transfers) and how community ecology *in situ* may influence their prevalence and distributions. This study reports the phylogenetic diversity, genomic locations, and biogeography of arsenic resistance genes in soils, integrating information from different ‘omics datasets and resources to provide a broad synthesis. The toolkit and the synthesis presented here can catalyze future work to understand the ecology and evolution of microbial arsenic biogeochemistry.

Acknowledgements

This work was supported by Michigan State University with computing resources provided by the Michigan State Institute for Cyber-Enabled Research. We thank the Ribosomal Database Project for their advice on reference gene database construction.

Funding

Metagenome sequencing was supported by the Joint Genome Institute Community Science Project #1834 and by Michigan State University. The work conducted by the U.S. Department of Energy (DOE) Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231. TKD was supported by the Ronald and Sharon Rogowski Fellowship for Food Safety and Toxicology and the Russel B. DuVall Graduate Fellowship from the Department of Microbiology and Molecular Genetics. SYY was supported through the Advanced Computational Research Experience program funded by the National Science Foundation under Grant No. 1560168. AS acknowledges support from the National Science Foundation DEB# 1655425 and #1749544, from the USDA National Institute of Food and Agriculture and Michigan State University AgBioResearch. AS and TKD acknowledge support from the National Institutes of Health R25GM115335.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Nelson MB, Martiny AC, Martiny JBH. Global biogeography of microbial nitrogen-cycling traits in soil. *Proc Natl Acad Sci* 2016.
2. Boon E, Meehan CJ, Whidden C, Wong DHJ, Langille MGI, Beiko RG. Interactions in the microbiome: Communities of organisms and communities of genes. *FEMS Microbiol Rev* 2014; **38**: 90–118.
3. Sforna MC, Philippot P, Somogyi A, Van Zuilen MA, Medjoubi K, Schoepp-Cothenet B, et al. Evidence for arsenic metabolism and cycling by microorganisms 2.7 billion years ago. *Nat Geosci* 2014; **7**: 811–815.
4. Zhu Y-G, Yoshinaga M, Zhao F-J, Rosen BP. Earth Abides Arsenic

Biotransformations. *Annu Rev Earth Planet Sci* 2014; **42**.

5. Edwardson CF, Hollibaugh JT. Metatranscriptomic analysis of prokaryotic communities active in sulfur and arsenic cycling in Mono Lake, California, USA. *ISME J* 2017.
6. Huang JH. Impact of microorganisms on arsenic biogeochemistry: A review. *Water Air Soil Pollut* . 2014.
7. Andres J, Bertin PN. The microbial genomics of arsenic. *FEMS Microbiol Rev* 2016; **40**: 299–322.
8. Rosen BP. Biochemistry of arsenic detoxification. *FEBS Lett* 2002; **529**: 86–92.
9. Li X, Zhang L, Wang G. Genomic evidence reveals the extreme diversity and wide distribution of the arsenic-related genes in Burkholderiales. *PLoS One* 2014; **9**: 1–11.
10. Crognale S, Amalfitano S, Casentini B, Fazi S, Petruccioli M, Rossetti S. Arsenic-related microorganisms in groundwater: a review on distribution, metabolic activities and potential use in arsenic removal processes. *Rev Environ Sci Biotechnol* 2017; **16**: 647–665.
11. Plewniak F, Crognale S, Rossetti S, Bertin PN, Marco DE, Pelaez AI. A Genomic Outlook on Bioremediation□: The Case of Arsenic Removal. *Front Microbiol* 2018; **9**: 1–8.
12. Jackson CR, Dugas SL, Harrison KG. Enumeration and characterization of arsenate-resistant bacteria in arsenic free soils. *Soil Biol Biochem* 2005; **37**: 2319–2322.
13. Dunivin TK, Miller J, Shade A. Taxonomically-linked growth phenotypes during arsenic stress among arsenic resistant bacteria isolated from soils overlying the Centralia coal seam fire. *PLoS One* 2018; **13**.
14. Zhu YG, Xue XM, Kappler A, Rosen BP, Meharg AA. Linking Genes to Microbial Biogeochemical Cycling: Lessons from Arsenic. *Environ Sci Technol* 2017.
15. Kurth D, Amadio A, Ordoñez OF, Albarracín VH, Gärtner W, Farías ME. Arsenic metabolism in high altitude modern stromatolites revealed by metagenomic analysis. *Sci Rep* 2017; **7**: 1024.
16. Chi L, Bian X, Gao B, Tu P, Ru H, Lu K. The effects of an environmentally

relevant level of arsenic on the gut microbiome and its functional metagenome. *Toxicol Sci* 2017; 1–12.

17. Rascovan N, Javier M, Martín P V, María EF. Metagenomic study of red biofilms from Diamante Lake reveals ancient arsenic bioenergetics in haloarchaea Metagenomic study of red biofilms from Diamante Lake reveals ancient arsenic bioenergetics in haloarchaea. *Int Soc Microb Ecol* 2016; **109**: 1–11.
18. Costa PS, Reis MP, Avila MP, Leite LR, De Araujo FMG, Salim ACM, et al. Metagenome of a microbial community inhabiting a metal-rich tropical stream sediment. *PLoS One* 2015; **10**: 1–21.
19. Cai L, Yu K, Yang Y, Chen BW, Li XD, Zhang T. Metagenomic exploration reveals high levels of microbial arsenic metabolism genes in activated sludge and coastal sediments. *Appl Microbiol Biotechnol* 2013; **97**: 9579–9588.
20. Babilonia J, Conesa A, Casaburi G, Pereira C, Louyakis AS, Reid RP, et al. Comparative metagenomics provides insight into the ecosystem functioning of the Shark Bay Stromatolites, Western Australia. *Front Microbiol* 2018; **9**: 1–16.
21. Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM, et al. FunGene: The functional gene pipeline and repository. *Front Microbiol* 2013.
22. The Uniprot Consortium T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017; **45**.
23. Tatusov RL, Galperin MY, Natale DA, Koonin E V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000; **28**: 33–36.
24. Edgar R. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004; **5**.
25. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol Biol Evol* 2016; msw054.
26. Dunivin TK, Choi J, Howe AC, Shade A. RefSoil+: A reference for antimicrobial resistance genes on soil plasmids. *bioRxiv* 2018.
27. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Verezemska O, Isbandi M, et al. Genomes OnLine Database (GOLD) v.6: Data updates and feature enhancements. *Nucleic Acids Res* 2017; **45**: D446–D456.

28. Johnson L, Eddy S, Portugaly E. Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure. *BMC Bioinformatics* 2011; **39**.
29. R Development Core Team. R: A language and environment for statistical computing. 2008.
30. Choi J, Yang F, Stepanauskas R, Cardenas E, Garoutte A, Williams R, et al. Strategies to improve reference databases for soil microbiomes. *ISME J* 2017.
31. Stamatakis A. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014; **30**.
32. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics* 2008; **10**.
33. Cho J-C, Lee D-H, Cho Y-C, Cho J-C, Kim S-J. Direct Extraction of DNA from Soil for Amplification of 16S rRNA Gene Sequences by Polymerase Chain Reaction. *J Microbiol* . 2006. , 229–235
34. Dunivin TK, Shade A. Community structure explains antibiotic resistance gene dynamics over a temperature gradient in soil. *FEMS Microbiol Ecol* 2018; **94**.
35. Wang Q, Fish JA, Gilman M, Sun Y, Brown CT, Tiedje JM, et al. Xander□: employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome* 2015; **3**: 32.
36. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 2007; **73**.
37. R Core Team. R: A Language and Environment for Statistical Computing. 2017. Vienna, Austria.
38. Ajees AA, Yang J, Rosen BP. The ArsD as(III) metallochaperone. *BioMetals*. 2011.
39. Qin J, Rosen BP, Zhang Y, Wang G, Franke S, Rensing C. Arsenic detoxification and evolution of trimethylarsine gas by a microbial arsenite S-adenosylmethionine methyltransferase. *Proc Natl Acad Sci* 2006.
40. Muller D, Lièvremont D, Simeonova DD, Jean-Claude H, Lett M-C. Arsenite oxidase aox genes from a metal-resistant beta-proteobacterium. *J Bacteriol* 2003; **185**: 135–141.

41. Saltikov CW, Newman DK. Genetic identification of a respiratory arsenate reductase. *Proc Natl Acad Sci U S A* 2003; **100**: 10983–8.
42. Zargar K, Conrad A, Bernick DL, Lowe TM, Stolc V, Hoeft S, et al. ArxA, a new clade of arsenite oxidase within the DMSO reductase family of molybdenum oxidoreductases. *Environ Microbiol* 2012.
43. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. *BMC Bioinformatics* 2009; **10**: 1–9.
44. Eddy SR. Hidden Markov models. *Curr Opin Struct Biol* . 1996.
45. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res* 2016; **44**: D67–D72.
46. Mikael Sehlin H, Börje Lindström E. Oxidation and reduction of arsenic by Sulfolobus acidocaldarius strain BC. *FEMS Microbiol Lett* 1992; **93**: 87–92.
47. Blomberg SP, Jr TG, Ives AR. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 2003; **57**: 717–745.
48. Mukhopadhyay R, Rosen BP, Phung LT, Silver S. Microbial arsenic: From geocycles to genes and enzymes. *FEMS Microbiol Rev* 2002; **26**: 311–325.
49. Chen S-C, Sun G-X, Rosen BP, Zhang S-Y, Deng Y, Zhu B-K, et al. Recurrent horizontal transfer of arsenite methyltransferase genes facilitated adaptation of life to arsenic. *Sci Rep* 2017; **7**: 7741.
50. Villegas-Torres MF, Bedoya-Reina OC, Salazar C, Vives-Florez MJ, Dussan J. Horizontal arsC gene transfer among microorganisms isolated from arsenic polluted soil. *Int Biodeterior Biodegrad* 2011; **65**: 147–152.
51. Cai L, Liu G, Rensing C, Wang G. Genes involved in arsenic transformation and resistance associated with different levels of arsenic-contaminated soils. *BMC Microbiol* 2009; **9**: 4.
52. Palmgren M, Engström K, Hallström BM, Wahlberg K, Søndergaard DA, Sall T, et al. AS3MT-mediated tolerance to arsenic evolved by multiple independent horizontal gene transfers from bacteria to eukaryotes. *PLoS One* 2017; **12**: 1–22.
53. Wang L, Wang J, Jing C. Comparative genomic analysis reveals organization, function and evolution of ars genes in Pantoea spp. *Front Microbiol* 2017; **8**: 1–12.
54. Lee S, Rakic-Martinez M, Graves LM, Ward TJ, Siletzky RM, Kathariou S.

Genetic determinants for cadmium and arsenic resistance among Listeria monocytogenes serotype 4B isolates from sporadic human listeriosis patients. *Appl Environ Microbiol* 2013; **79**: 2471–2476.

55. Achour AR, Bauda P, Billard P. Diversity of arsenite transporter genes from arsenic-resistant soil bacteria. *Res Microbiol* 2007; **158**: 128–137.

56. Jackson CR, Dugas SL. Phylogenetic analysis of bacterial and archaeal arsC gene sequences suggests an ancient, common origin for arsenate reductase. *BMC Evol Biol* 2003; **3**: 18.

57. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 2017.

58. Escudero L V, Casamayor EO, Chong G, Pedrós-Alió C, Demergasso C. Distribution of microbial arsenic reduction, oxidation and extrusion genes along a wide range of environmental arsenic concentrations. *PLoS One* 2013; **8**: e78890.

59. Zhao FJ, Harris E, Yan J, Ma J, Wu L, Liu W, et al. Arsenic methylation in soils and its relationship with microbial arsM abundance and diversity, and As speciation in rice. *Environ Sci Technol* 2013; **47**: 7147–7154.

60. Huang H, Jia Y, Sun G-X, Zhu Y-G. Arsenic Speciation and Volatilization from Flooded Paddy Soils Amended with Different Organic Matters. *Environ Sci Technol* 2012; **46**: 2163–2168.

61. Qiao JT, Li XM, Hu M, Li FB, Young LY, Sun WM, et al. Transcriptional Activity of Arsenic-Reducing Bacteria and Genes Regulated by Lactate and Biochar during Arsenic Transformation in Flooded Paddy Soil. *Environ Sci Technol* 2018; **52**: 61–70.

62. Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, Walters WA, et al. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biol Biochem* 2011; **43**: 1450–1455.

63. Li R, Chai M, Qiu GY. Distribution, fraction, and ecological assessment of heavy metals in sediment-plant system in mangrove forest, South China Sea. *PLoS One* 2016; **11**: 1–15.

64. Chatterjee M, Massolo S, Sarkar SK, Bhattacharya AK, Bhattacharya BD, Satpathy KK, et al. An assessment of trace element contamination in intertidal

sediment cores of Sunderban mangrove wetland, India for evaluating sediment quality guidelines. *Environ Monit Assess* 2009; **150**: 307–322.

65. Chaudhuri P, Nath B, Birch G. Accumulation of trace metals in grey mangrove Avicennia marina fine nutritive roots: The role of rhizosphere processes. *Mar Pollut Bull* 2014; **79**: 284–292.

66. Fahy A, Giloteaux L, Bertin PN, Le Paslier D, Médigue C, Weissenbach J, et al. 16S rRNA and As-Related Functional Diversity: Contrasting Fingerprints in Arsenic-Rich Sediments from an Acid Mine Drainage. *Microb Ecol* 2015; **70**: 154–167.

67. Desoeuvre A, Casiot C, Héry M. Diversity and Distribution of Arsenic-Related Genes Along a Pollution Gradient in a River Affected by Acid Mine Drainage. *Microb Ecol* 2015; 1–14.

Figure Legends

Figure 1. Arsenic resistance gene toolkit schematic. A) Seed sequences for nine arsenic resistance genes were used to construct an arsenic resistance gene database with existing tools [21, 28, 32, 35]. Lines indicate interdependence between modules. **B)** Table of arsenic resistnac genes included in the toolkit. The toolkit is freely available on GitHub: github.com/ShadeLab/meta_arsenic

Figure 2. Arsenic resistance genes in RefSoil+ organisms. A) Maximum likelihood tree of 16S rRNA genes in RefSoil+ organisms. Bootstrap support > 50 is shown with black circles. Tree branches and first ring are colored by organism taxonomy. Each node is annotated with arsenic resistance genotype where color indicates gene. Filled boxes indicate gene presence on chromosome, and open boxes indicate gene presence on plasmid. **B)** Proportion of RefSoil+ organisms and organisms containing arsenic resistance genes colored by the taxonomy of the organism containing the gene. “None” refers to the number of genomes that do not test positive for any of the nine arsenic resistance genes analyzed. Note the difference between y-axes.

Figure 3. Phylogeny of arsenite efflux pumps in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of **A)** Acr3 and **B)** ArsB sequences from RefSoil+ organisms. Tree scale = 1. Leaf tip color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the tree. Grey circles on the exterior of the tree indicate that a hit was detected on a plasmid and not a chromosome.

Figure 4. Phylogeny of cytoplasmic arsenate reductases in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of **A)** ArsC (trx) and **B)** ArsC (grx) sequences from RefSoil+ organisms. Tree scale = 1. Leaf tip color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the tree. Grey circles on the exterior of the tree indicate that a hit

was detected on a plasmid and not a chromosome.

Figure 5. Phylogeny of ArsM in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of ArsM sequences from RefSoil+ organisms. Tree scale = 1. Leaf tip color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the tree. Grey circles on the exterior of the tree indicate that a hit was detected on a plasmid and not a chromosome.

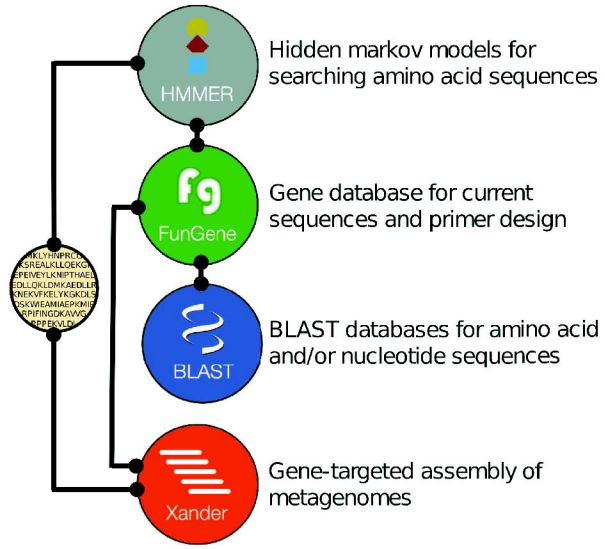
Figure 6. Phylogeny of dissimilatory arsenic resistance proteins in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of dissimilatory arsenic resistance genes from RefSoil+ ogranisms. Tree scale = 0.1. Leaf tips show the name of the RefSoil+ organisms and background color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the tree. Grey circles on the exterior of the tree indicate that a hit was detected on a plasmid and not a chromosome.

Figure 7. Comparison of arsenic resistance gene abundance between cultivation dependent and cultivation independent methods. A) Mean normalized abundance of RefSoil microorganisms based on 16S rRNA gene abundance from Earth Microbiome Project datasets. Points are colored by soil order. **B)** Normalized abundance of arsenic resistance genes in RefSoil+ and metagenomes. Metagenome abundance was normalized to *rplB*, and RefSoil+ normalized abundance was calculated using the number of RefSoil+ genomes. Only metagenomes with an arsenic resistance gene detected are shown, and the total number of datasets (including RefSoil+) is shown in parentheses. **C)** *rplB*-normalized abundance of arsenic resistance genes in cultivation dependent and independent metagenomes from the same soil sample.

Figure 8. Arsenic resistance gene biogeography. A) Relative abundance of arsenic resistance genes in soil metagenomes. **B)** Rank *rplB*-normalized abundance of arsenic resistance genes in soil metagenomes. Sites are ordered by rank mean

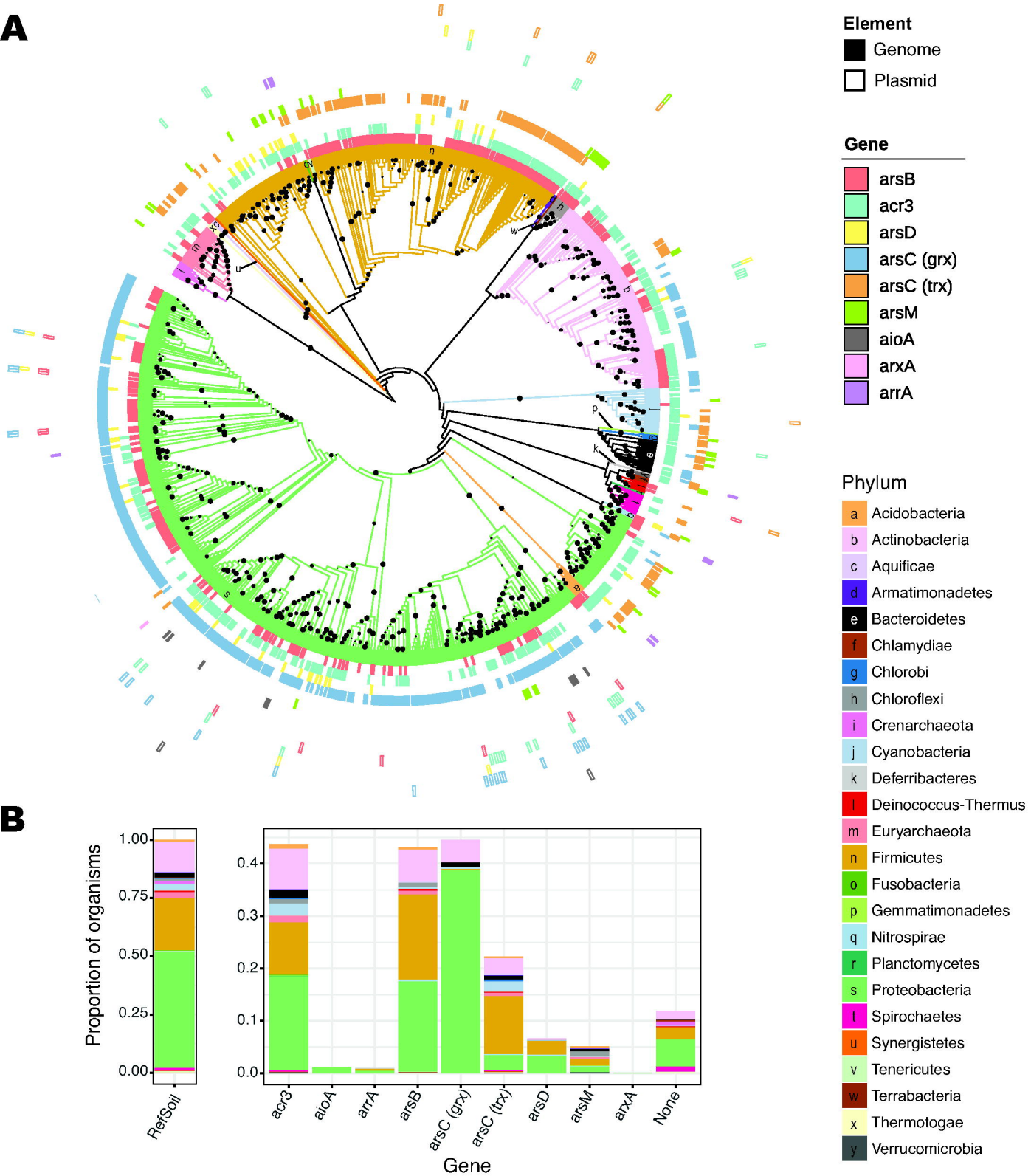
774 abundance. Note the differences in y axes. **C)** Abundance-occurrence plots of
775 arsenic resistance gene sequences clustered at 90% aa identity.

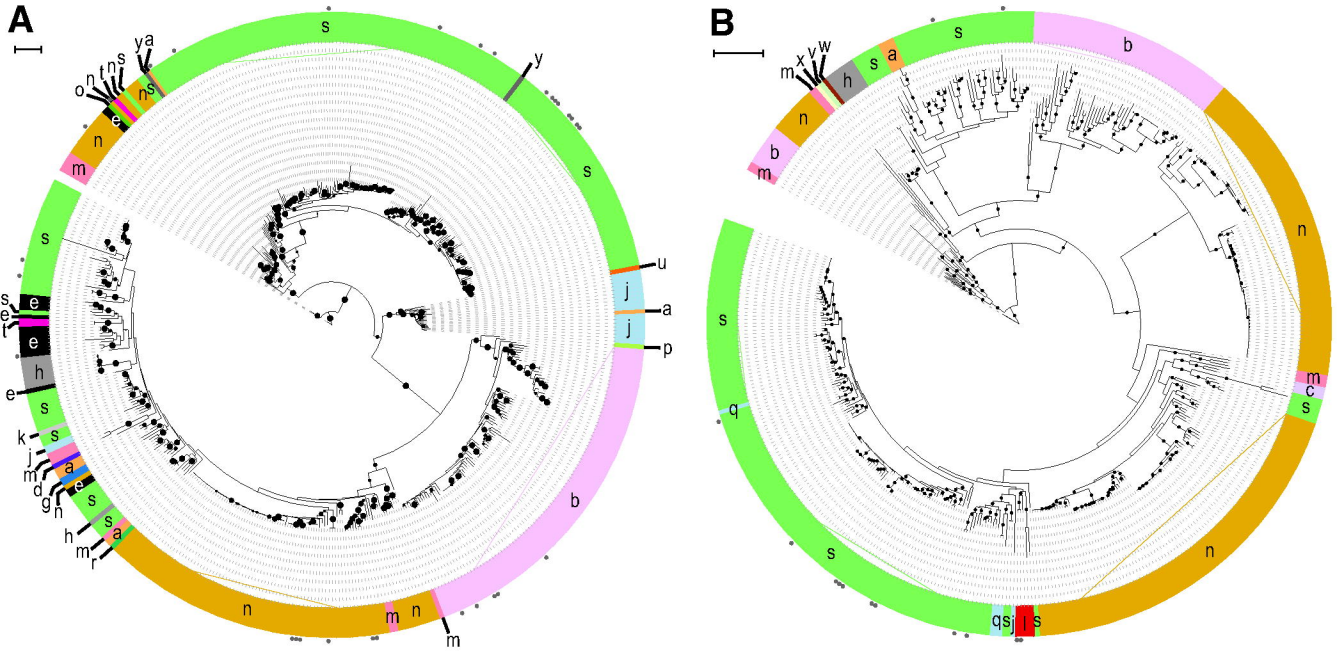
A



B

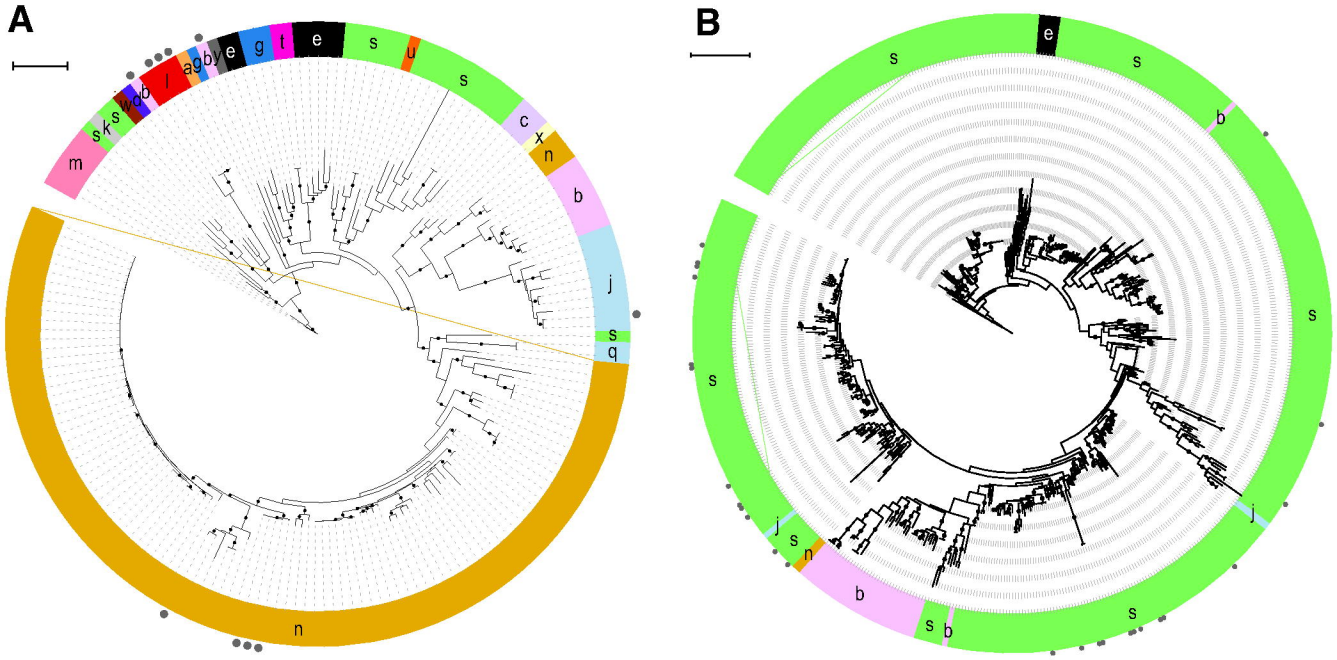
Gene	Category	Encoded function	Reference
acr3	Detoxification	arsenite permease	[8]
arsB	Detoxification	arsenite permease	[8]
arsC (grx)	Detoxification	cytoplasmic arsenate reductase	[8]
arsC (trx)	Detoxification	cytoplasmic arsenate reductase	[8]
arsD	Detoxification	transcriptional repressor and metallochaperone	[37]
arsM	Metabolism	arsenite methyltransferase	[38]
aioA	Metabolism	arsenite oxidase (large subunit)	[39]
arrA	Metabolism	dissimilatory arsenate reductase (large subunit)	[40]
arxA	Metabolism	arsenite oxidase (molybdopterin subunit)	[41]

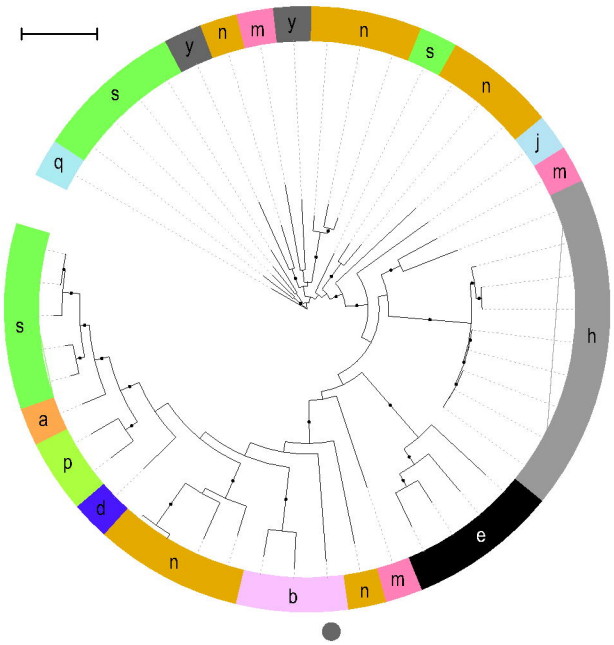

A
Element
Genome
Plasmid
Gene
arsB
acr3
arsD
arsC (grx)
arsC (trx)
arsM
aioA
arxA
arrA
Phylum
a Acidobacteria
b Actinobacteria
c Aquificae
d Armatimonadetes
e Bacteroidetes
f Chlamydiae
g Chlorobi
h Chloroflexi
i Crenarchaeota
j Cyanobacteria
k Deferribacteres
l Deinococcus-Thermus
m Euryarchaeota
n Firmicutes
o Fusobacteria
p Gemmatimonadetes
q Nitrospirae
r Planctomycetes
s Proteobacteria
t Spirochaetes
u Synergistetes
v Tenericutes
w Terrabacteria
x Thermotogae
y Verrucomicrobia
B
Proportion of organisms
1.00
0.75
0.50
0.25
0.00
0.4
0.3
0.2
0.1
0.0
RefSoil
acr3
aioA
arrA
arsB
arsC (grx)
arsC (trx)
arsD
arsM
arxA
None
Gene



Phylum

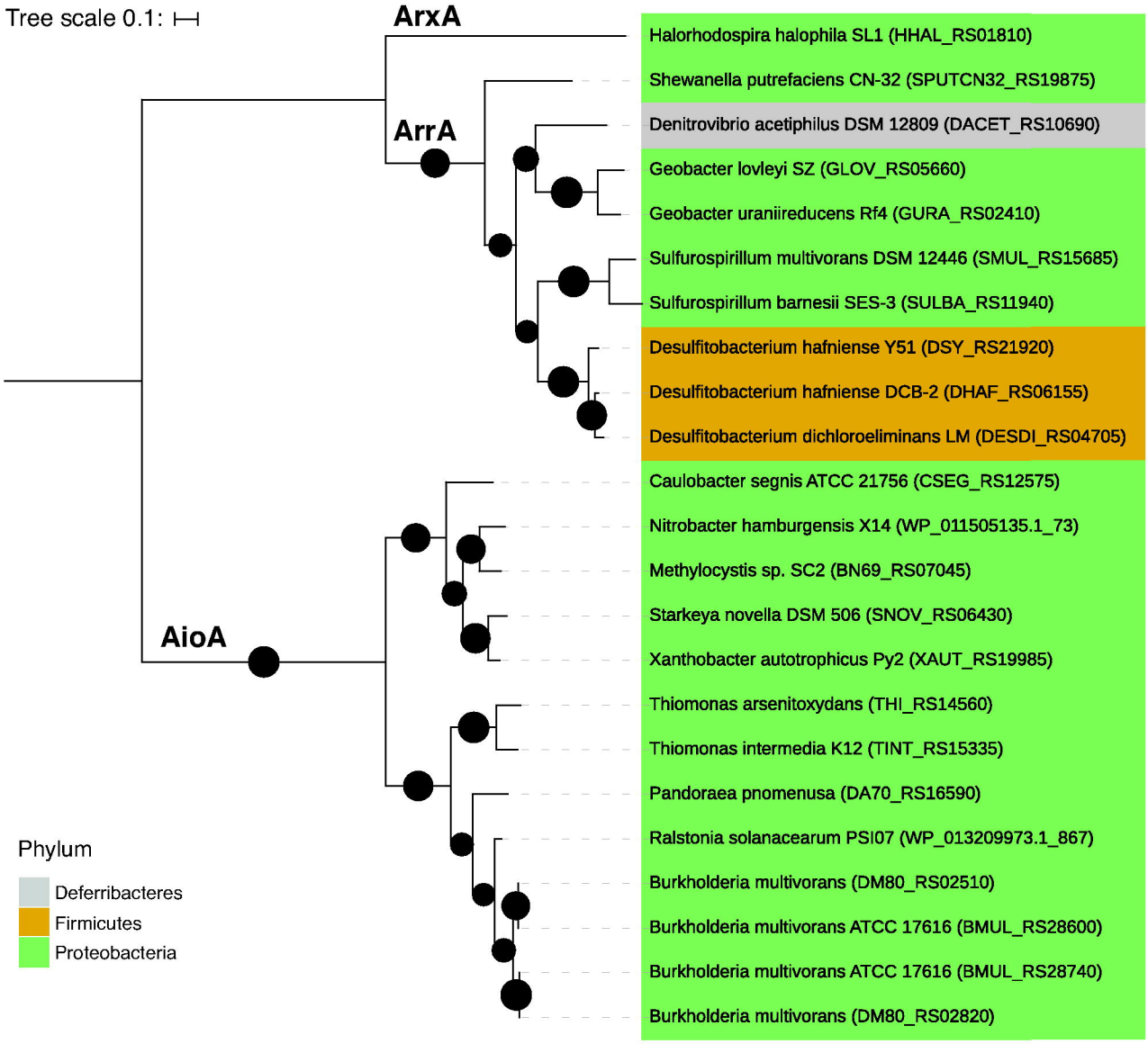
a Acidobacteria	f Chlamydiae	k Deferribacteres	p Gemmatimonadetes	u Synergistetes
b Actinobacteria	g Chlorobi	l Deinococcus-Thermus	q Nitrospirae	v Tenericutes
c Aquificae	h Chloroflexi	m Euryarchaeota	r Planctomycetes	w Terrabacteria
d Armatimonadetes	i Crenarchaeota	n Firmicutes	s Proteobacteria	x Thermotogae
e Bacteroidetes	j Cyanobacteria	o Fusobacteria	t Spirochaetes	y Verrucomicrobia

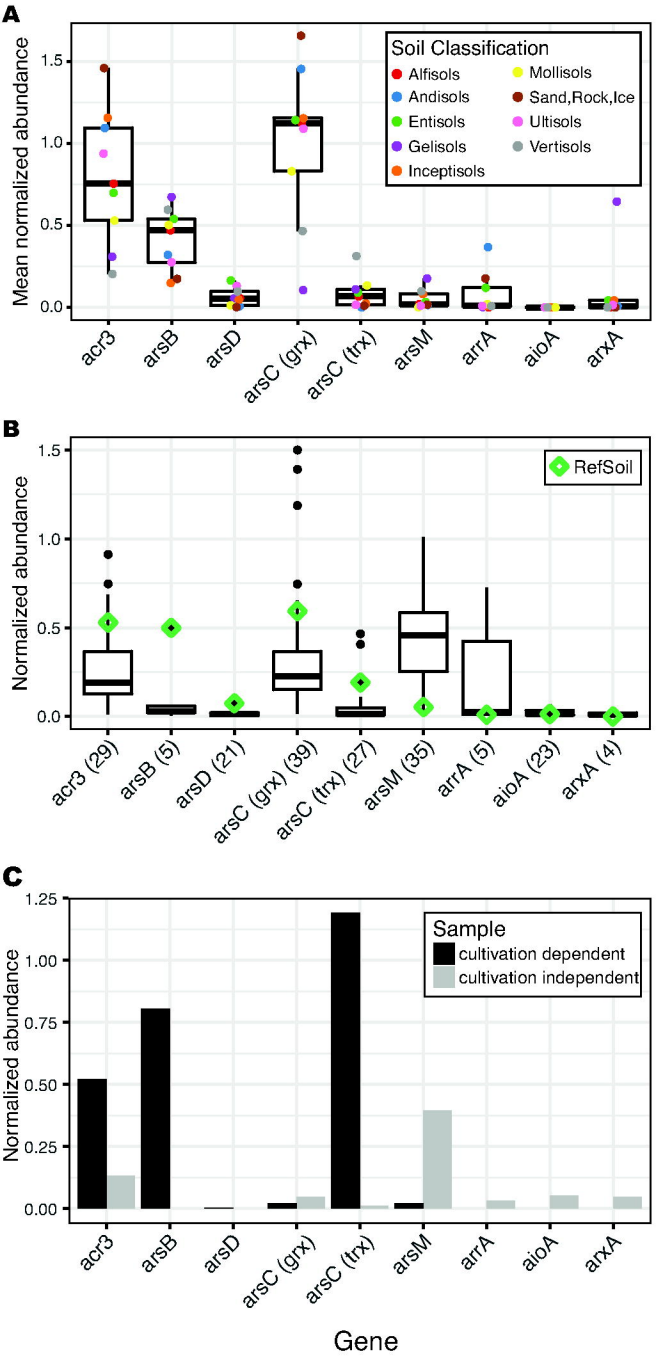

A
B
Phylum
a Acidobacteria
b Actinobacteria
c Aquificae
d Armatimonadetes
e Bacteroidetes
f Chlamydiae
g Chlorobi
h Chloroflexi
i Crenarchaeota
j Cyanobacteria
k Deferribacteres
l Deinococcus-Thermus
m Euryarchaeota
n Firmicutes
o Fusobacteria
p Gemmatimonadetes
q Nitrospirae
r Planctomycetes
s Proteobacteria
t Spirochaetes
u Synergistetes
v Tenericutes
w Terrabacteria
x Thermotogae
y Verrucomicrobia

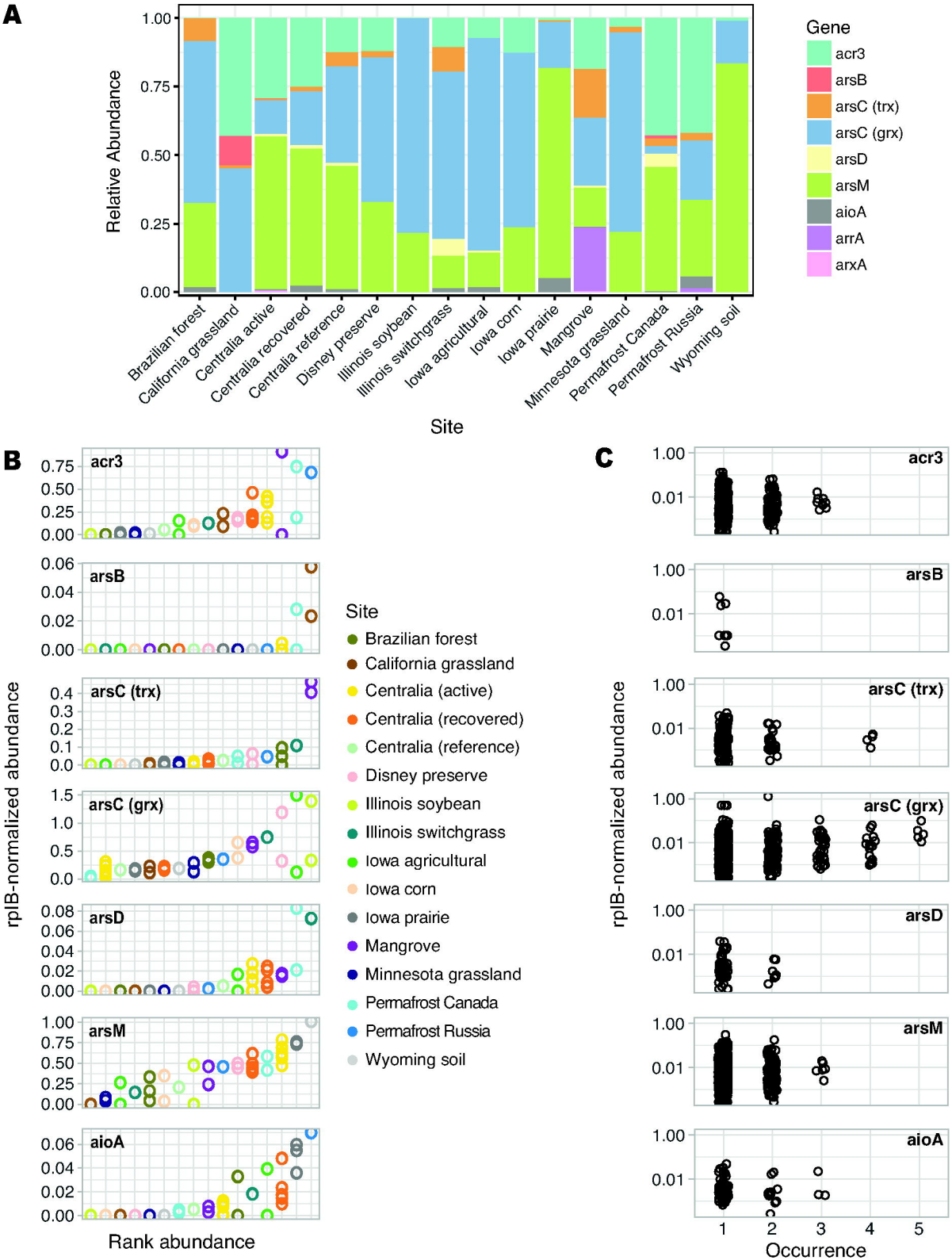


Phylum

a	Acidobacteria	n	Firmicutes
b	Actinobacteria	o	Fusobacteria
c	Aquificae	p	Gemmatimonadetes
d	Armatimonadetes	q	Nitrospirae
e	Bacteroidetes	r	Planctomycetes
f	Chlamydiae	s	Proteobacteria
g	Chlorobi	t	Spirochaetes
h	Chloroflexi	u	Synergistetes
i	Crenarchaeota	v	Tenericutes
j	Cyanobacteria	w	Terrabacteria
k	Deferribacteres	x	Thermotogae
l	Deinococcus-Thermus	y	Verrucomicrobia
m	Euryarchaeota		


Tree scale 0.1:
ArxA
ArrA
AioA
Halorhodospira halophila SL1 (HHAL_RS01810)
Shewanella putrefaciens CN-32 (SPUTCN32_RS19875)
Denitrovibrio acetiphilus DSM 12809 (DACET_RS10690)
Geobacter lovleyi SZ (GLOV_RS05660)
Geobacter uraniireducens Rf4 (GURA_RS02410)
Sulfurospirillum multivorans DSM 12446 (SMUL_RS15685)
Sulfurospirillum barnesii SES-3 (SULBA_RS11940)
Desulfitobacterium hafniense Y51 (DSY_RS21920)
Desulfitobacterium hafniense DCB-2 (DHAF_RS06155)
Desulfitobacterium dichloroeliminans LM (DESDI_RS04705)
Caulobacter segnis ATCC 21756 (CSEG_RS12575)
Nitrobacter hamburgensis X14 (WP_011505135.1_73)
Methylocystis sp. SC2 (BN69_RS07045)
Starkeya novella DSM 506 (SNOV_RS06430)
Xanthobacter autotrophicus Py2 (XAUT_RS19985)
Thiomonas arsenitoxydans (THI_RS14560)
Thiomonas intermedia K12 (TINT_RS15335)
Pandoraea pnomenusa (DA70_RS16590)
Ralstonia solanacearum PSI07 (WP_013209973.1_867)
Burkholderia multivorans (DM80_RS02510)
Burkholderia multivorans ATCC 17616 (BMUL_RS28600)
Burkholderia multivorans ATCC 17616 (BMUL_RS28740)
Burkholderia multivorans (DM80_RS02820)
Phylum
Deferribacteres
Firmicutes
Proteobacteria


A
Mean normalized abundance
1.5
1.0
0.5
0.0
Soil Classification
Alfisols
Andisols
Entisols
Gelisols
Inceptisols
Mollisols
Sand,Rock,Ice
Ultisols
Vertisols
acr3
arsB
arsD
arsC (grx)
arsC (trx)
arsM
arrA
aioA
arxA
B
Normalized abundance
1.5
1.0
0.5
0.0
RefSoil
acr3 (29)
arsB (5)
arsD (21)
arsC (grx) (39)
arsC (trx) (27)
arsM (35)
arrA (5)
aioA (23)
arxA (4)
C
Normalized abundance
1.25
1.00
0.75
0.50
0.25
0.00
Sample
cultivation dependent
cultivation independent
acr3
arsB
arsD
arsC (grx)
arsC (trx)
arsM
arrA
aioA
arxA
Gene


A
Relative Abundance
1.00
0.75
0.50
0.25
0.00
Brazilian forest
California grassland
Centralia active
Centralia recovered
Centralia reference
Disney preserve
Illinois soybean
Illinois switchgrass
Iowa agricultural
Iowa corn
Iowa prairie
Mangrove
Minnesota grassland
Permafrost Canada
Permafrost Russia
Wyoming soil
Site
Gene
acr3
arsB
arsC (trx)
arsC (grx)
arsD
arsM
aioA
arrA
arxA
B
acr3
arsB
arsC (trx)
arsC (grx)
arsD
arsM
aioA
rplB-normalized abundance
Rank abundance
Site
Brazilian forest
California grassland
Centralia (active)
Centralia (recovered)
Centralia (reference)
Disney preserve
Illinois soybean
Illinois switchgrass
Iowa agricultural
Iowa corn
Iowa prairie
Mangrove
Minnesota grassland
Permafrost Canada
Permafrost Russia
Wyoming soil
C
acr3
arsB
arsC (trx)
arsC (grx)
arsD
arsM
aioA
rplB-normalized abundance
1.00
0.01
1
2
3
4
5
Occurrence