Dunivin et al. BMC Biology (2019) 17:45
https://doi.org/10.1186/s12915-019-0661-5

BMC Biology

RESEARCH ARTICLE Open Access

A global survey of arsenic-related genes in soil microbiomes

Taylor K. Dunivin[1,2], Susanna Y. Yeh[3] and Ashley Shade[1,4,5,6*]

Abstract

Background: Environmental resistomes include transferable microbial genes. One important resistome component is resistance to arsenic, a ubiquitous and toxic metalloid that can have negative and chronic consequences for human and animal health. The distribution of arsenic resistance and metabolism genes in the environment is not well understood. However, microbial communities and their resistomes mediate key transformations of arsenic that are expected to impact both biogeochemistry and local toxicity.

Results: We examined the phylogenetic diversity, genomic location (chromosome or plasmid), and biogeography of arsenic resistance and metabolism genes in 922 soil genomes and 38 metagenomes. To do so, we developed a bioinformatic toolkit that includes BLAST databases, hidden Markov models and resources for gene-targeted assembly of nine arsenic resistance and metabolism genes: *acr3*, *aioA*, *arsB*, *arsC* (grx), *arsC* (trx), *arsD*, *arsM*, *arrA*, and *arxA*. Though arsenic-related genes were common, they were not universally detected, contradicting the common conjecture that all organisms have them. From major clades of arsenic-related genes, we inferred their potential for horizontal and vertical transfer. Different types and proportions of genes were detected across soils, suggesting microbial community composition will, in part, determine local arsenic toxicity and biogeochemistry. While arsenic-related genes were globally distributed, particular sequence variants were highly endemic (e.g., *acr3*), suggesting dispersal limitation. The gene encoding arsenic methylase *arsM* was unexpectedly abundant in soil metagenomes (median 48%), suggesting that it plays a prominent role in global arsenic biogeochemistry.

Conclusions: Our analysis advances understanding of arsenic resistance, metabolism, and biogeochemistry, and our approach provides a roadmap for the ecological investigation of environmental resistomes.

Keywords: Arsenic, Functional gene, Bioinformatics, Targeted gene assembly, Horizontal gene transfer, Biogeography, Phylogeny, Phylogenetic diversity, Resistome, Plasmid

Background

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]. The arsenic resistance and metabolism genes exemplify a suite of well-characterized functional genes that have consequences for biogeochemistry. Arsenic is a toxic metalloid that, upon exposure, can have negative effects for all life, including humans, livestock, and microorganisms. The toxicity and mobility of arsenic depends, in part, on its oxidation state: the trivalent arsenite is more mobile and more toxic than the pentavalent arsenate [3]. The toxicity of methylated arsenic species varies with oxidation state and number of methyl groups (monomethyl, dimethyl, trimethyl). Pentavalent methylarsenicals are progressively less toxic than inorganic arsenate, while trivalent methylarsenicals are progressively more toxic than inorganic arsenite with the exception of trimethylarsine which is the least toxic arsenic species [4, 5]. Additionally, volatilization of arsenic can occur through methylation [6], which has varied impacts. Methylated

* Correspondence: shadeash@msu.edu
[1]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA
[4]Program in Ecology, Evolutionary Biology and Behavior, Michigan State University, East Lansing, MI 48824, USA
Full list of author information is available at the end of the article

forms of arsenic can be released to new areas through air [7], captured during bioremediation [8], or accumulate in crops such as rice [9]. Microbial transformations of arsenic can have consequences for arsenic speciation and methylation; therefore, they impact arsenic ecotoxicity and the fate of arsenic in the environment.

Arsenic biogeochemical cycling by microbial communities is both an ancient [10, 11] and a contemporary [3, 12] phenomenon. Changes to the methylation or oxidation state of arsenic alter biogeochemical cycling of arsenic, and microbes have evolved a variety of mechanisms to carry out these functions. Arsenic-related genes are generally separated into two categories: resistance and metabolism [13]. Arsenic resistance, or detoxification, is encoded by the *ars* operon [14]. The *ars* operon protects the cell from arsenic but does not detoxify arsenic itself in the environment. This operon includes arsenite efflux (ArsB, Acr3) which is potentially precluded by cytoplasmic arsenate reduction with either glutaredoxin (ArsC (grx)) or thioredoxin (ArsC (trx)) [14]. Arsenic metabolisms include methylation (ArsM), oxidation (AioAB, ArxAB), and dissimilatory reduction (ArrAB) [13]. While these genetic determinants of arsenic detoxification and metabolism are well-characterized, the full scope of arsenic detoxification and metabolism gene distribution, diversity, and interspecies transfer is unknown [15–17].

Microbial arsenic resistance is reportedly widespread in the environment. Arsenic-resistant organisms have been found in sites with low arsenic concentrations (< 7 ppm) [18, 19], and it has been speculated that nearly all organisms have arsenic resistance genes [20]. While the number of identified microorganisms with arsenic resistance genes continues to grow [13], the number of microorganisms without arsenic resistance genes is unclear. Furthermore, though the complete arsenic biogeochemical cycle has been detected in the environment [10], the relative contributions of genes encoding detoxification and metabolism remain unknown [11]. A global, biogeographic perspective of environmental arsenic-related genes would improve understanding of their ecology. This information would expand foundational knowledge of arsenic detoxification and metabolism, including local and global abundances, gene diversity, dispersal across different environments, and representations over the microbial tree of life.

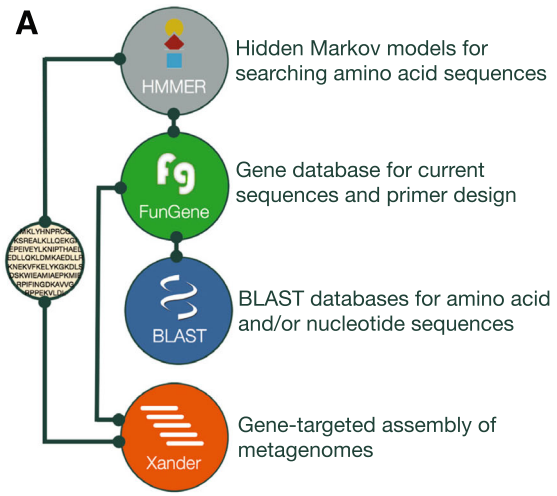
Knowledge gaps concerning the diversity of microbial arsenic-related genes are driven, in part, by numerous inconsistencies in nomenclature and detection methods. Though public microbial metagenome and genome data continue to surge, there are several practical hurdles to achieving a robust, global assessment of microbial arsenic-related genes from this wealth of data. First, tools to detect these genes rely on imperfect annotation [15] and widely vary in nomenclature [21]. Next, the use of different reference databases [12, 22–25] and normalization techniques [25, 26] complicates comparisons between studies. To overcome these hurdles, we developed an open-access toolkit to examine arsenic resistance and metabolism genes in microbial sequence datasets. This toolkit allowed us probe genomic and metagenomic datasets simultaneously to investigate arsenic-related genes in soil microbiomes. We first asked whether arsenic-related genes are universal in soil-associated microorganisms. Next, we tested the hypothesis that genes encoding arsenic detoxification are more abundant than those encoding arsenic metabolism. We also tested the hypothesis that arsenic resistance genes with redundant function (i.e., *acr3* and *arsB*; *arsC* (grx) and *arsC* (trx)) would have complementary environmental abundances. Third, we asked whether estimations of arsenic-related gene abundance are biased by cultivation efforts, as cultivation is often a research emphasis because cultivable, arsenic-resistant microorganisms can be used in bioremediation [17]. Finally, we tested the hypothesis that sequence variants of arsenic-related genes are endemic, not cosmopolitan.

Results

A bioinformatic toolkit for detecting and quantifying arsenic-related genes

We developed a toolkit to improve investigations of microbial arsenic-related genes (Fig. 1a, b) [14, 31–35]. We selected these nine genes because they are markers of arsenic detoxification and metabolism [21, 25] 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 [30], functional gene (FunGene) databases [27], hidden Markov models (HMMs [36]), and gene resources for gene-targeted assembly (Xander [28]) (Fig. 1a). Altogether, this toolkit relies on consistent references and nomenclature and can search both amino acid and nucleotide sequence data.

To demonstrate the utility of our toolkit, we performed an analysis of arsenic-related genes in soil-associated genomes and metagenomes. We used HMMs for marker genes for arsenic detoxification and metabolism to search RefSoil+ genomes, a set of complete chromosomes and plasmids from cultivable soil microorganisms [37]. Additionally, we used a gene-targeted assembler [28] to test 38 public soil metagenomes from Brazil, Canada, Malaysia, Russia, and the USA for arsenic resistance and metabolism



B

Gene	Category	Encoded function	Reference
acr3	Detoxification	arsenite efflux pump	[14]
arsB	Detoxification	arsenite efflux pump	[14]
arsC (grx)	Detoxification	cytoplasmic arsenate reductase	[14]
arsC (trx)	Detoxification	cytoplasmic arsenate reductase	[14]
arsD	Detoxification	transcriptional repressor and metallochaperone	[31]
arsM	Metabolism	arsenite methyltransferase	[32]
aioA	Metabolism	arsenite oxidase (large subunit)	[33]
arrA	Metabolism	dissimilatory arsenate reductase (large subunit)	[34]
arxA	Metabolism	arsenite oxidase (molybdopterin subunit)	[35]

Fig. 1 Arsenic resistance and metabolism gene toolkit schematic. **a** Seed sequences for nine arsenic resistance genes were used to construct an arsenic resistance gene database with existing tools [27–30]. Lines indicate interdependence between modules. **b** Table of arsenic resistance and metabolism genes included in the toolkit. The toolkit is freely available on GitHub: https://github.com/ShadeLab/PAPER_Dunivin_meta_arsenic

genes (Additional file 1). Ultimately, these data serve as a broad baseline of arsenic detoxification and metabolism genes in soil.

Phylogenetic distributions and genomic locations of arsenic-related genes

We asked whether arsenic resistance and metabolism genes were universal in RefSoil+ organisms [37]. Of the 922 RefSoil+ genomes spanning 25 phyla (Fig. 2b; Additional file 2), 14.3% (132 genomes) did not contain any tested arsenic-related genes. 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 detoxification and metabolism genes. Supporting this hypothesis, a Crenarchaeota isolate was previously reported to oxidize arsenic [38]. Nonetheless, these data suggest that arsenic-related genes are widespread, but not universal, even among cultivable soil organisms (Fig. 2).

We next asked whether 16S rRNA gene phylogeny was predictive of arsenic genotypes using a test for phylogenetic signal (Bloomberg's K [39]). 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 (Additional files 3, 4, 5, 6, and 7), chromosomally encoded *acr3*, *arsB*, *arsC* (grx), *arsC* (trx), and *arsM* were similar between phylogenetically related organisms (false discovery rate adjusted $p < 0.01$; Fig. 2a).

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

Arsenite efflux pumps

We examined the phylogenetic diversity of distinct genes encoding arsenite efflux pumps, *acr3* and *arsB*, for soil-associated microorganisms (Fig. 3, Additional files 3 and 4). Gene *acr3* is separated into two clades: *acr3*(1) and *acr3*(2) [40]. Clade *acr3*(1) is typically composed of Proteobacterial sequences while *acr3*(2) is typically composed of Firmicutes and Actinobacterial sequences [21, 40, 41]. Though RefSoil+ genomes were mostly composed of *acr3*(2) sequences from Proteobacteria (Fig. 3a; Additional file 3), we observed greater taxonomic diversity observed than previously reported for this clade [21, 40, 41]. 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%). Plasmid-borne *arsB* sequences were only present in Proteobacteria and Deinococcus-Thermus strains (Fig. 3b; Additional file 4). Sequences from Actinobacteria, Proteobacteria, and Firmicutes were each present in two distinct phylogenetic groups, and previous studies also observed separation of *arsB* sequences based on phylum [40, 41]. 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 (Additional file 4).

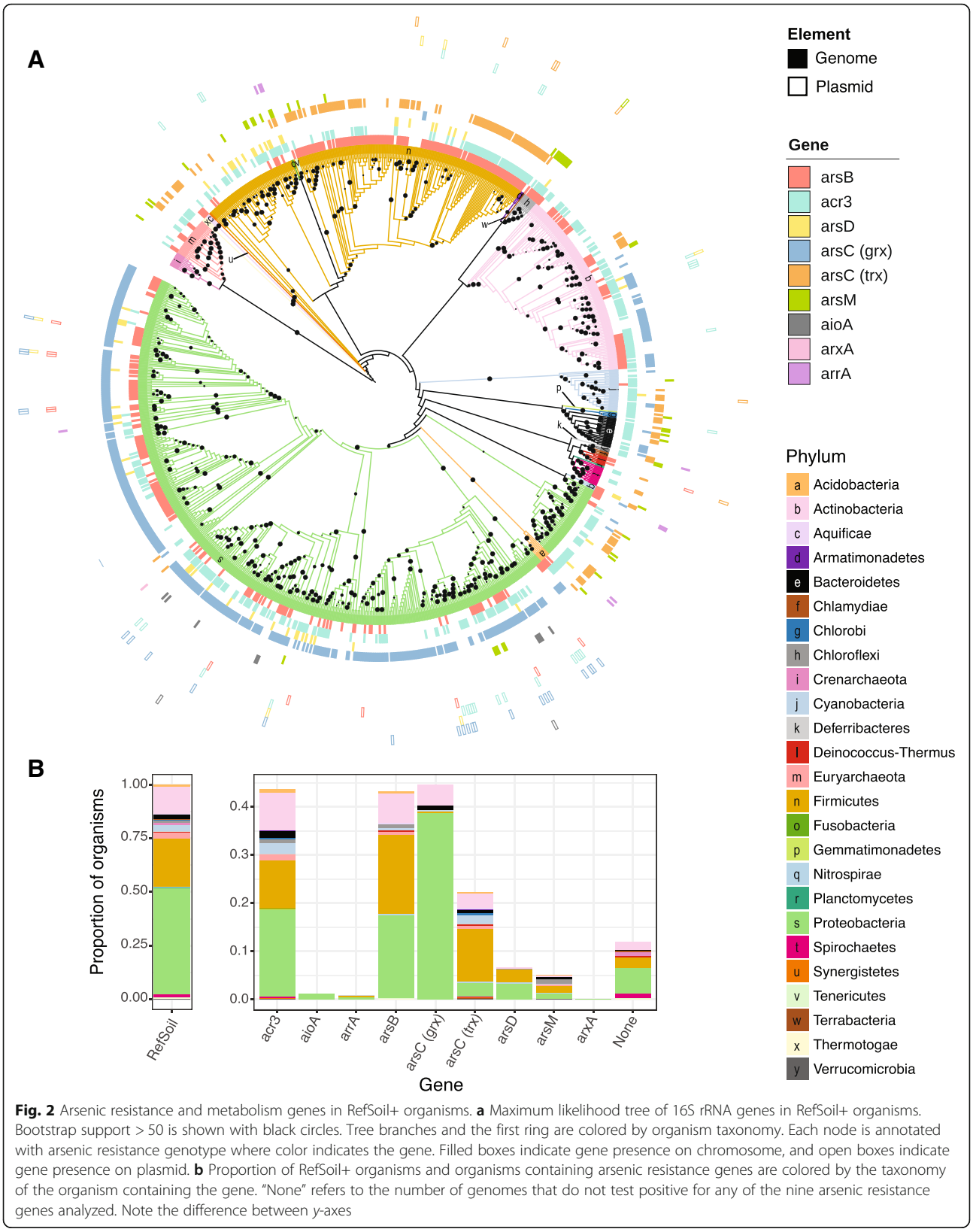


Fig. 2 Arsenic resistance and metabolism genes in RefSoil+ organisms. **a** Maximum likelihood tree of 16S rRNA genes in RefSoil+ organisms. Bootstrap support > 50 is shown with black circles. Tree branches and the first ring are colored by organism taxonomy. Each node is annotated with arsenic resistance genotype where color indicates the 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 are 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

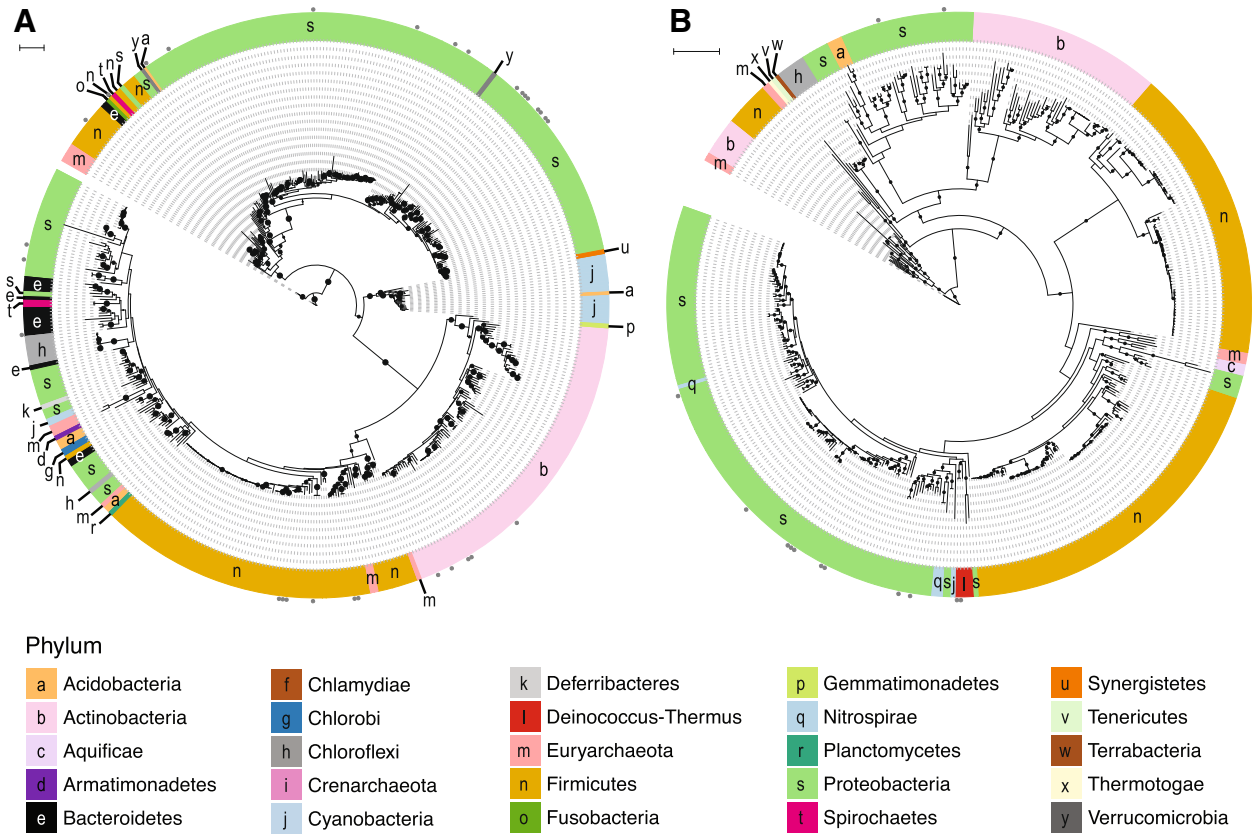


Fig. 3 Phylogeny of arsenite efflux pumps in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of **a** Acr3 and **b** ArsB sequences predicted from RefSoil+ genomes. Tree scale = 1. Leaf tip color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black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

Cytoplasmic arsenate reductases

Cytoplasmic arsenate reductase (ArsC (trx)) was phylogenetically widespread in RefSoil+ microorganisms (Fig. 4a; Additional file 5). While some *arsC* (trx) sequences were plasmid-borne, the majority were chromosomally encoded. Similarly, plasmid-encoded *arsC* (grx) made up 4.6% of RefSoil+ hits (Fig. 4b; Additional file 6). 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

arsM was relatively uncommon in RefSoil+ microorganisms (5.2%) (Fig. 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, and Verrucomicrobia (Fig. 5; Additional file 7). Notably, only one RefSoil+ microorganism, *Rubrobacter radiotolerans* (NZ_CP007516.1), had a plasmid-borne *arsM*.

Arsenic metabolism genes *aioA*, *arrA*, and *arxA* were phylogenetically conserved (Fig. 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 (Fig. 6).

Cultivation bias and environmental distributions of arsenic-related genes

To gain a cultivation-dependent perspective of the abundances of arsenic-related genes in soils, we used inferred environmental abundances of RefSoil microorganisms [42, 43]. The environmental abundance of RefSoil microorganisms, which are cultivable, soil-associated microorganisms, was previously estimated by comparing 16S rRNA gene sequences in RefSoil with those in soil metagenomes [42]. We used this estimated abundance of cultivable microorganisms along with arsenic-related gene

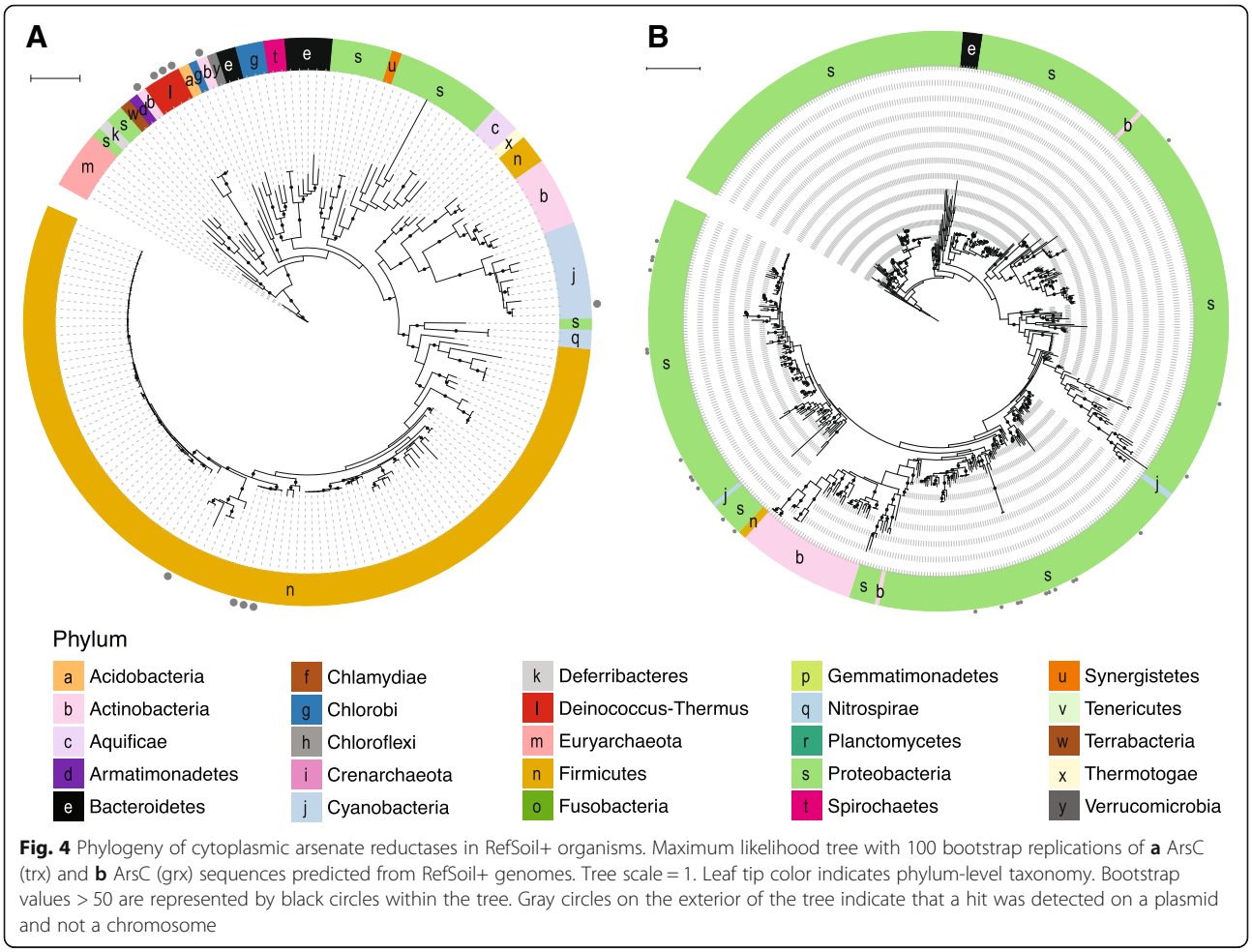


Fig. 4 Phylogeny of cytoplasmic arsenate reductases in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of **a** ArsC (trx) and **b** ArsC (grx) sequences predicted from RefSoil+ genomes. Tree scale = 1. Leaf tip color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the tree. Gray circles on the exterior of the tree indicate that a hit was detected on a plasmid and not a chromosome

information from this study (Fig. 2) to estimate the environmental abundances of arsenic-related genes from the cultivated bacteria. Arsenic metabolism genes (*aioA*, *arrA*, *arsM*, *arxA*) were predicted to be less common in the environment compared with arsenic detoxification genes (*acr3*, *arsB*, *arsC* (grx), *arsC* (trx), and *arsD*) (Fig. 7a; Mann-Whitney *U* test $p < 0.01$). Despite similar distributions of *acr3* and *arsB* in RefSoil+ (Fig. 2b), *acr3* was more abundant in most soil orders (Fig. 7a; Mann-Whitney *U* test $p < 0.05$). For genes encoding cytoplasmic arsenate reductases, *arsC* (grx) was more abundant than *arsC* (trx) (Mann-Whitney *U* test $p < 0.01$).

To gain a cultivation-independent perspective of the abundances of arsenic-related genes, we examined their normalized abundance from soil metagenomes (Fig. 7b). An undetected gene does not confirm absence, so we present a conservative estimate that only includes metagenomes testing positive for a gene. Arsenic detoxification genes (*acr3*, *arsB*, *arsC* (grx), *arsC* (trx), and *arsD*) were more abundant than arsenic metabolism genes (*aioA*, *arrA*, *arsM*, and *arxA*) (Mann-Whitney *U* test $p < 0.01$; Fig. 7b). Genes encoding arsenite efflux pumps differed in their abundance with *acr3* being more abundant than *arsB* (Mann-Whitney *U* test $p < 0.01$). We also observed differences in cytoplasmic arsenate reductases: *arsC* (grx) was more abundant than *arsC* (trx) (Mann-Whitney *U* test $p < 0.01$).

We explored cultivation bias of arsenic-related genes with a case study comparing cultivation-dependent (lawn growth on the standard medium TSA50) and cultivation-independent communities from the same soil. Genes in the *ars* operon (*acr3*, *arsB*, *arsD*, and *arsC* (trx)) were elevated in the cultivation-dependent metagenome (Fig. 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 cultivation-independent methods, these results recapitulate the general discrepancies between RefSoil+ genomes and soil metagenomes (Fig. 7b). This bias has

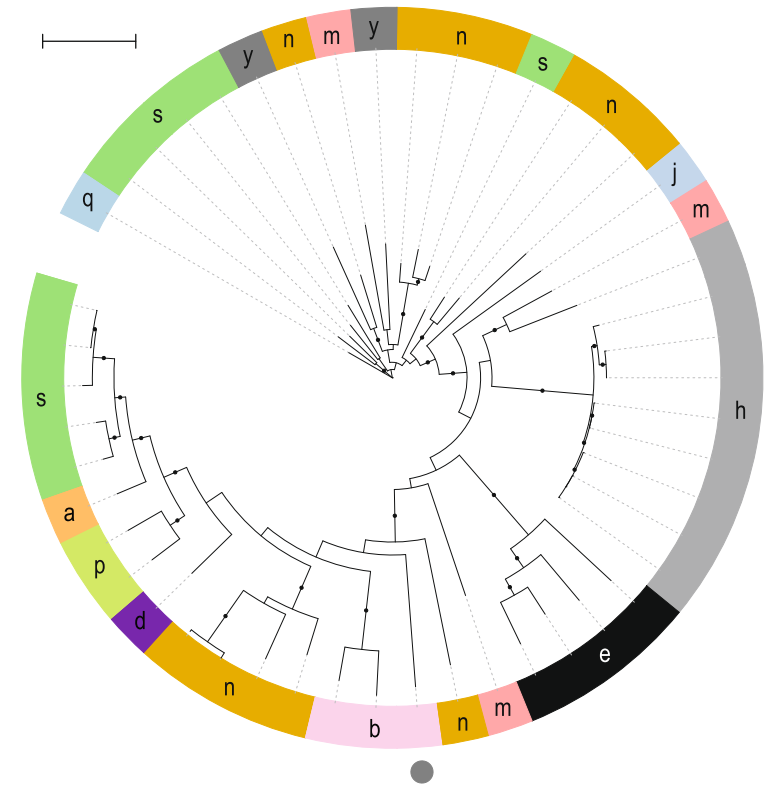


Fig. 5 Phylogeny of ArsM in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of ArsM sequences predicted from RefSoil+ genomes. Tree scale = 1. Leaf tip color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the tree. Gray circles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

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-related gene endemism

Arsenic-related genes are globally distributed, but their biogeography is poorly understood. Broadly, arsenic-related genes had comparable abundance

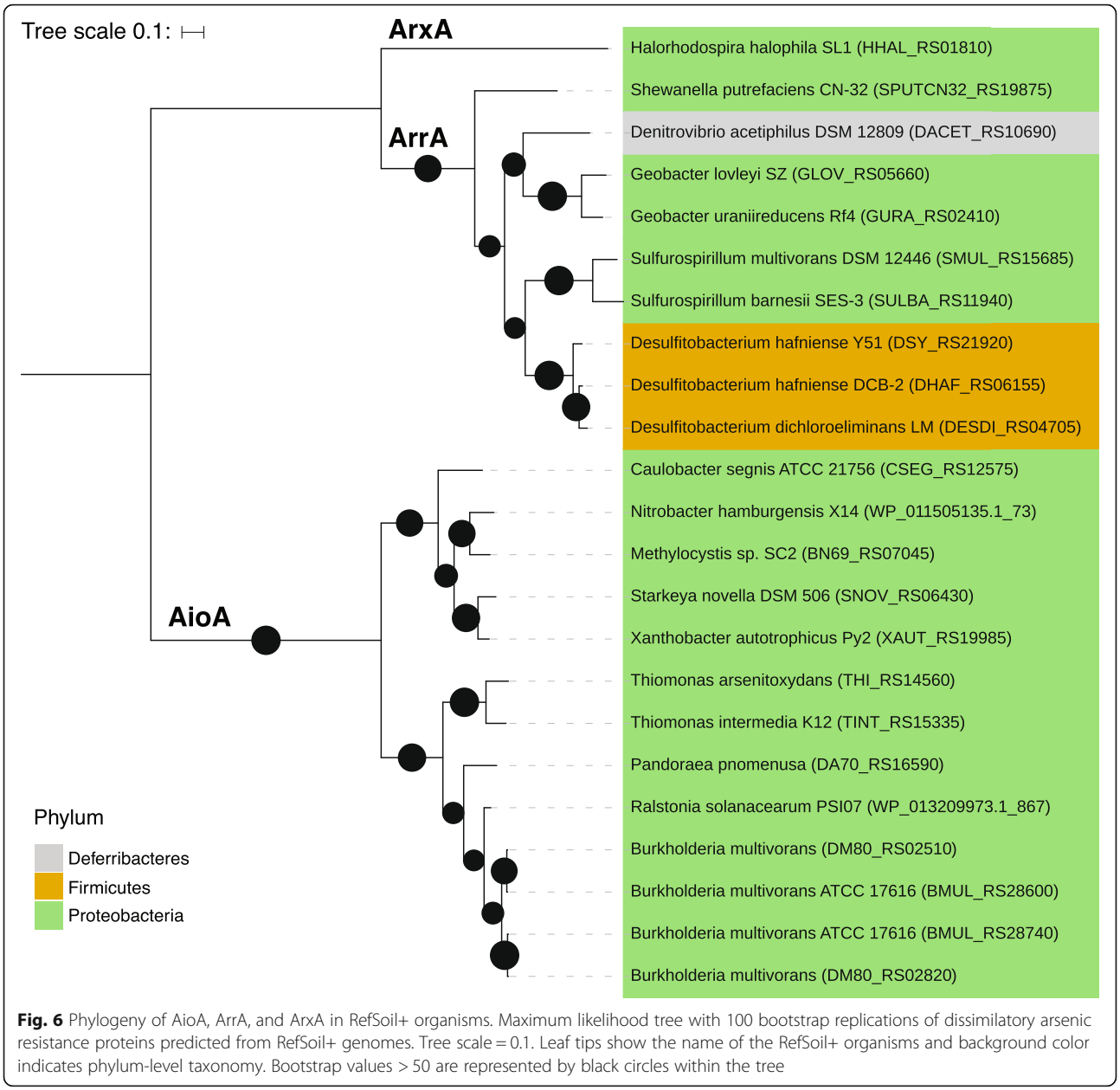


Fig. 6 Phylogeny of AioA, ArrA, and ArxA in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of dissimilatory arsenic resistance proteins predicted from RefSoil+ genomes. Tree scale = 0.1. Leaf tips show the name of the RefSoil+ organisms and background color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the tree

among different soils (Fig. 7a, b). The relative distributions of distinct arsenic detoxification and metabolism 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 USA (Additional file 1). These 16 sites had differences in community membership (Additional file 9) and arsenic-related gene content (Fig. 8a). Geographic location was not predictive of arsenic-related gene content (Mantel's $r = 0.03493$; $p > 0.05$). Soils had different distributions of arsenic-related 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 (Fig. 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 [44]. The mangrove sample had the most even proportions of arsenic-related genes (Fig. 8a). This distribution was driven by a high abundance of *arsC* (trx) and *arrA*.

We further examined the arsenic resistance gene abundance at individual sites. We did not include *arr* and *arx* in this analysis due to limited available data. For each gene, the abundance varied greatly, but replicates

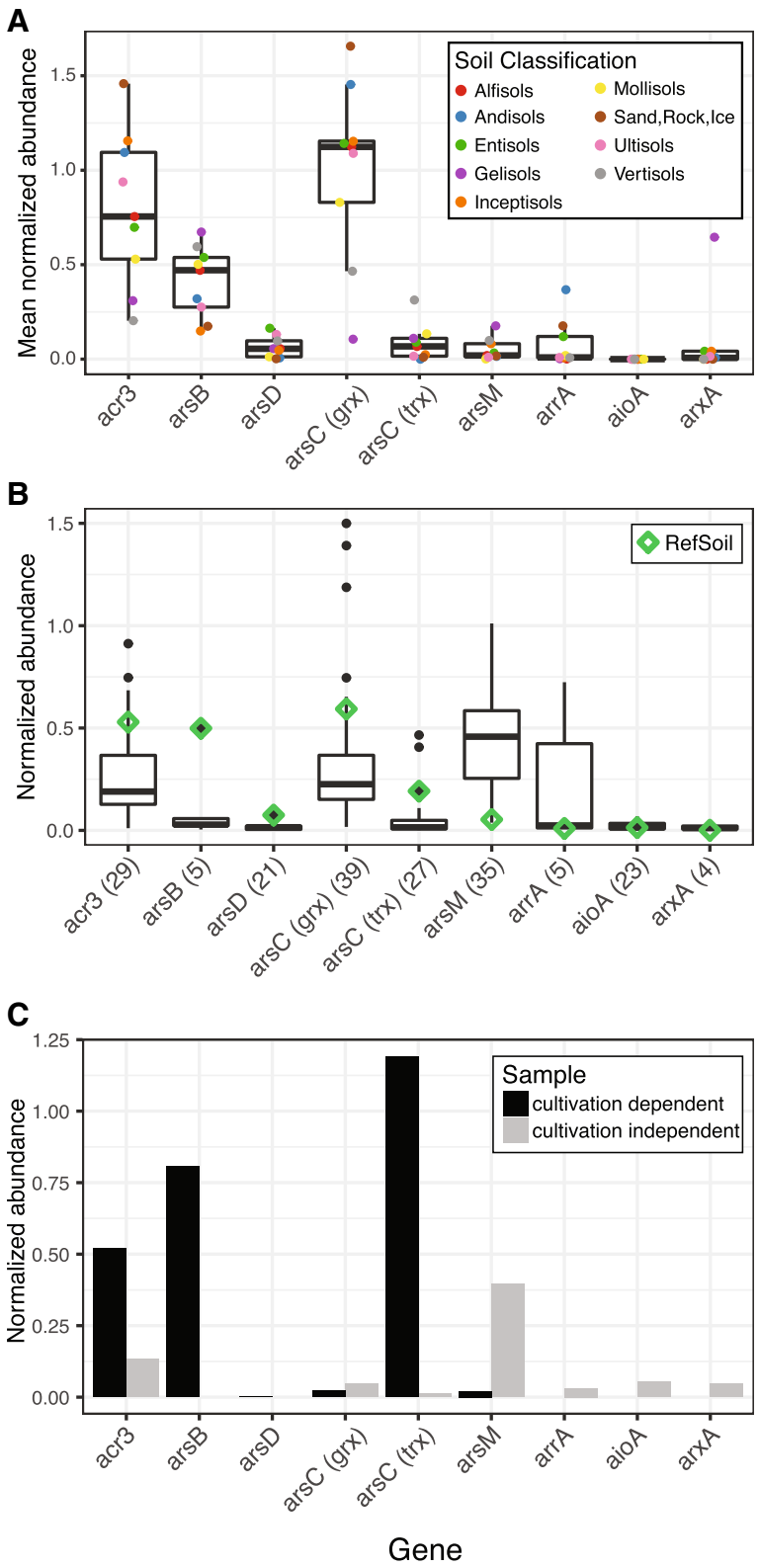


Fig. 7 (See legend on next page.)

(See figure on previous page.)
Fig. 7 Comparison of arsenic resistance and metabolism gene abundance between cultivation-dependent and cultivation-independent methods. **a** Mean normalized abundance of arsenic-related genes based on RefSoil microorganisms abundance estimated from corresponding 16S rRNA gene abundance in Earth Microbiome Project datasets. Points are colored by soil order. **b** Normalized abundance of arsenic resistance genes in RefSoil+ and 38 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 cultivation-independent metagenomes from the same soil sample

within one site had similar abundances (Fig. 8b). The majority of arsenic-related gene sequences (99.3%) were endemic and only found in one to two sites, but 24 sequences were detected in three or more sites (Fig. 8c; Additional file 10). The majority (70.8%) of cosmopolitan sequences belonged to *arsC* (grx). This analysis suggests that arsenic-related genes *acr3*, *arsB*, *arsC* (trx), *arsD*, *arsM*, and *aioA* are generally endemic.

Discussion

A bioinformatic toolkit for detecting and quantifying arsenic-related genes

We developed a toolkit for detecting arsenic-related genes from sequence data that supports a variety of applications (Fig. 1a): arsenic-related genes can be detected in amino acid sequences from completed genomes (HMMs [29], BLAST [30]), nucleotide sequences in draft genomes (BLAST), and metagenomes and metatranscriptomes (Xander [28]). 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. While primers already exist for arsenic-related genes: *aioA* [45, 46], *acr3* [41], *arsB* [41], *arsC* (grx) [47], *arsC* (trx) [48], *arsM* [9], and *arrA* [49–51], these FunGene [27] databases can be used for testing primer breadth, designing new primers, and browsing sequences.

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) [20]), or redox transformations of elements involved in arsenic biogeochemical cycling (e.g., nitrate reductase (NarG) and sulfate reductase (DsrAB) [3, 20]). This toolkit serves as both a resource and an example workflow for developing similar toolkits to examine functional genes, beyond arsenic-related genes, in microbial sequence datasets.

Phylogenetic diversity and distribution of arsenic-related genes

It has been conjectured that nearly all organisms have arsenic resistance genes [20], and though this assumption has propagated in the literature, it had never been explicitly quantified. Our data suggest that arsenic detoxification and metabolism genes are ubiquitous, but not universal in RefSoil+ microorganisms (Fig. 2). It is possible for these 132 organisms to have untested or novel arsenic-related genes; nonetheless, these nine well-characterized genes were not universally detected. Additionally, phylogeny was predictive of the presence of *acr3*, *arsB*, *arsC* (grc), *arsC* (trx), and *arsM*. This correlation suggests that taxonomy is predictive of arsenic genotype despite documented potential for HGT [19, 40, 48, 52, 53]. This result could be explained by ancient rather than contemporary HGT, as seen with *arsM* [53] and *arsC* (grx) [48]. Therefore, we next assessed evidence for HGT by examining the phylogenetic congruence and genomic location (e.g., chromosome or plasmid) of arsenic-related gene sequences.

Horizontal transfer of arsenic-related genes has been well documented [19, 40, 48, 52–55] and is an important consideration for understanding the propagation and taxonomic identity of arsenic-related genes. We examined the phylogenetic diversity of arsenic-related genes in RefSoil+ microorganisms, including plasmids and chromosomes, and compared them with the 16S rRNA gene taxonomy.

Efflux pumps

While known *acr3* sequences separate into two clades [21, 40, 41], plasmid-borne *acr3* sequences were present across clades, 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 (Fig. 3b; Additional file 4). This observation is in agreement with previous reports comparing the diversity of *arsB* to *acr3* [40, 41]. Multiple, phylogenetically distinct copies of *arsB* were present in some RefSoil+ organisms, which 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.

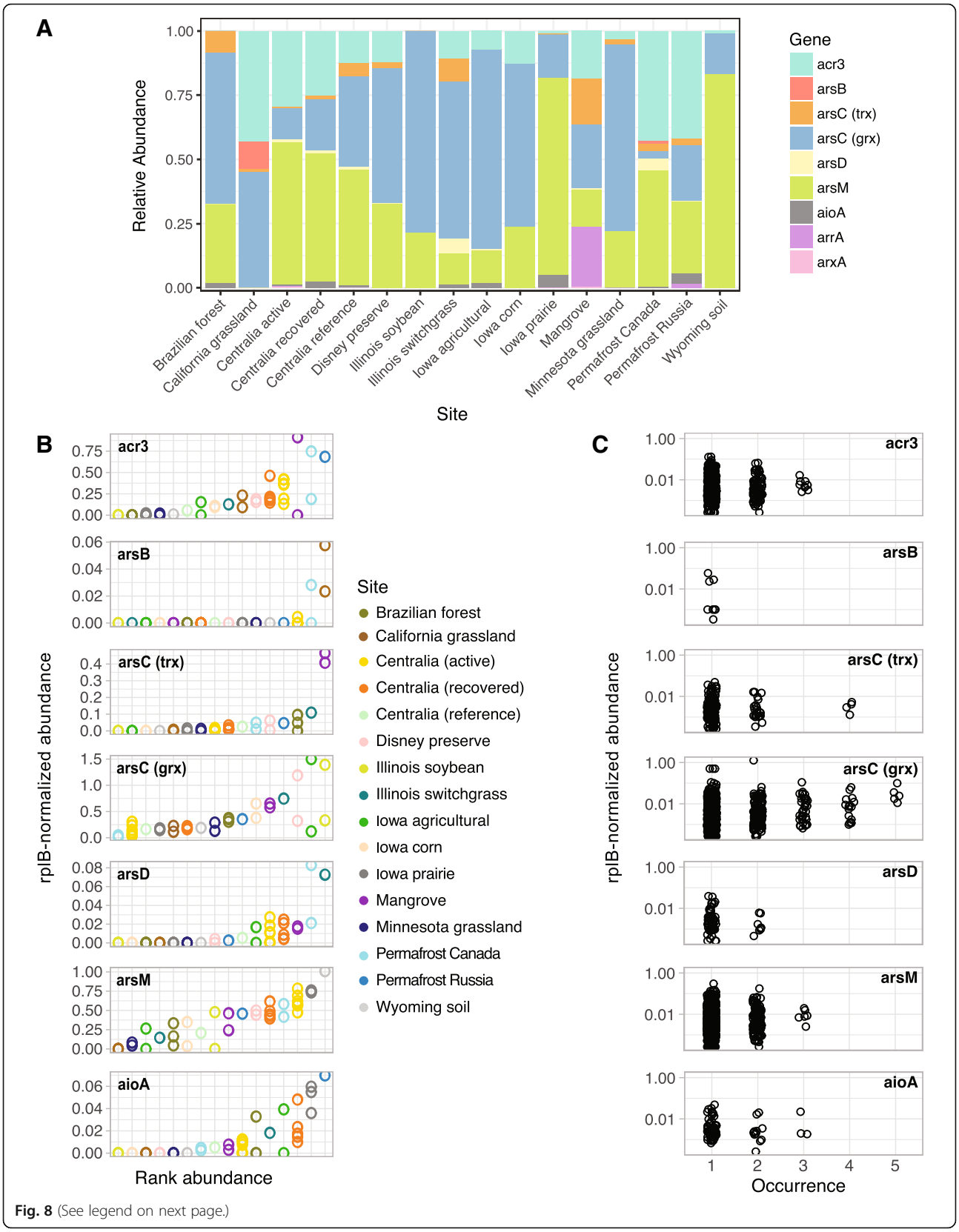


Fig. 8 (See legend on next page.)

(See figure on previous page.)
Fig. 8 Arsenic resistance and metabolism gene biogeography. **a** Relative abundance of arsenic resistance genes in soil metagenomes. **b** Rank *rplB*-normalized abundance of arsenic-related genes in soil metagenomes. Sites are ordered by rank mean abundance. Note the differences in *y*-axes. **c** Abundance-occurrence plots of arsenic-related gene sequences clustered at 90% amino acid identity. Number of samples included are as follows: Brazilian forest $n = 3$, California grassland $n = 2$, Centralia active $n = 7$, Centralia recovered $n = 5$, Centralia reference $n = 1$, Disney preserve $n = 2$, Illinois soybean $n = 2$, Illinois switchgrass $n = 1$, Iowa agricultural $n = 2$, Iowa corn $n = 2$, Iowa prairie $n = 3$, Mangrove $n = 2$, Minnesota grassland $n = 2$, Permafrost Canada $n = 2$, Permafrost Russia $n = 1$, and Wyoming soil $n = 1$

Cytoplasmic arsenate reductases

arsC (trx) was predominantly found on RefSoil+ chromosomes, not plasmids, suggesting vertical transfer of *arsC* (trx) is common. *arsC* (trx) was present in both Bacteria and Archaea, and sequences from the two domains formed two distinct clades. *arsC* (trx) sequences that cluster separately from Bacterial-*arsC* (trx) sequences have been documented in Thermococci, Archaeoglobi, Thermoplasmata, and Halobacteria [56]. Together, this distribution supports an early evolutionary origin for *arsC* (trx). Thus, *arsC* (trx) appears to be an evolutionarily old enzyme that is phylogenetically conserved despite its presence on plasmids and potential for HGT. Plasmid-encoded *arsC* (grx) were also observed in RefSoil+ microorganisms, highlighting a contemporary potential for HGT that has been documented in soil [48]. Thus, both genes encoding cytoplasmic arsenate reductases were more common on chromosomes.

Arsenic metabolisms

The evolutionary history of the gene encoding arsenite S-adenosylmethionine methyltransferase, *arsM*, was recently investigated [52, 53]. Both studies independently determined that *arsM* evolved billions of years ago and was subject to HGT [52, 53]. In this work, *arsM* sequences from Euryarchaeota were dispersed throughout the *arsM* phylogeny, supporting the potential for inter-kingdom transfer events that were recently suggested [52, 53]. Very few RefSoil+ organisms had arsenic metabolism genes *aioA*, *arrA*, or *arxA*, which limits phylogenetic analysis. Nonetheless, they were mostly found in Proteobacteria, which is in agreement with previous work [13].

Cultivation bias and environmental distributions of arsenic-related genes

Cultivation-based assessments of arsenic-related gene content are important since cultivable strains are often favored for bioremediation [57]. We estimated distributions of arsenic-related genes in cultivable microorganisms from soils and found a greater abundance of arsenic detoxification genes *acr3*, *arsB*, and *arsC* (trx) (Fig. 7a). A previous study also reported an abundance of *acr3* over *arsB* in cultivable microoganisms from forest soils and attributed this to the greater phylogenetic distribution of *acr3* compared with *arsB* [41]. Additionally, they found that *arsC* (grx) was more abundant than *arsC* (trx) in cultivated microorganisms from these soils. It has been posited in cultivation-independent studies that *arsC* (trx) is more efficient than *arsC* (grx) and that high local arsenic concentrations result in a relatively greater abundance of *arsC* (trx) [21, 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 assess arsenic-related gene content without cultivation bias, we examined arsenic-related genes in soil metagenomes. As predicted by cultivable organisms, arsenic metabolism 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 result could be due, in part, to the large number of RefSoil+ microorganisms with multiple copies of these genes (Additional file 8). Cultivation-independent genomes (e.g., single-cell-amplified genomes and metagenome-assembled genomes) could provide greater context about the environmental distributions of copy numbers of arsenic-related genes.

Notably, *arsM* was abundant in soil (median 48%), which greatly exceeds cultivation-dependent estimations, and in a case study of cultivation-dependent and cultivation-independent techniques, *arsM* was more abundant in the cultivation-independent sample (Fig. 7c). Due to the early phylogenetic origins of *arsM* and its independent functionality [53], this abundance of *arsM* in soil metagenomes is not unexpected. *arsM* is typically studied in paddy soils [6, 59, 60], but metagenomes in this study suggest it is an important component of the arsenic biogeochemical cycle in a variety of soils.

Arsenic-related gene endemism

We examined the relative abundance of arsenic-related genes in soil metagenomes and observed differences in genetic potential for arsenic transformation that could impact biogeochemical cycling (Fig. 8a). Notably, the mangrove sample had the most even proportions of arsenic-related genes. While the arsenic concentrations in this sample are unknown, mangroves are considered sources and sinks for arsenic [61–63]. This could explain

the greater abundance of *arsC* (trx), which is hypothesized to be more abundant in high arsenic sites [21, 58]. Additionally, *arrA* encodes a dissimilatory arsenate reductase that functions in an anaerobic environment [34], so its greater abundance in sediment is expected. Soil geochemical data was not available for all metagenomes examined in this work, so direct comparisons of arsenic-related gene content and soil geochemistry were not possible. This highlights the importance and utility of depositing geochemical data with DNA sequences. Future work, however, could further examine relationships between arsenic resistance genes and soil geochemical data, including arsenic concentration and redox potential.

We also measured whether arsenic-related gene sequence variants were endemic or cosmopolitan in soil metagenomes (Fig. 8c). We found that genes *acr3*, *arsB*, *arsC* (trx), *arsD*, *arsM*, and *aioA* were generally endemic, suggesting regional dispersal limitation. Only one *aioA* and three *acr3* sequences were detected in multiple sites. This supports a previous finding that *acr3* and *aioA* from the acid mine drainage in Carnoulès were endemic [64]. Conversely, *arsC* (grx) was cosmopolitan which could suggest genetic migration via HGT or vertical transfer and a limited gene diversification. Both are plausible since *arsC* (grx) was common in RefSoil+ plasmids and had low phylogenetic diversity (Fig. 4b; Additional file 6).

Conclusions

We developed a bioinformatic toolkit for detecting arsenic detoxification and metabolism genes in microbial sequence data and applied it to analyze the genomes and metagenomes from soil microorganisms. This toolkit informs hypotheses about the evolutionary histories of these genes (including potential for vertical and horizontal transfers) and how community ecology in situ may influence their prevalence and distribution. This study reports the phylogenetic diversity, genomic locations, and biogeography of arsenic-related 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. Furthermore, the toolkit acts as a framework for similar studies of other functional genes of interest.

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 [21, 25], so we selected nine well-characterized genes: *acr3*, *aioA*, *arsB*, *arsC* (grx), *arsC* (trx), *arsD*, *arsM*, *arrA*, and *arxA*. FunGene databases [27] were constructed for the following arsenic-related genes: *arsB*, *arsC* (grx), *arsC* (trx), *acr3*, *aioA*, *arrA*, and *arxA*. The *arxA* database was constructed with seed sequences from [12]. For all other genes, UniProt [65] was used to obtain full-length, reviewed sequences when possible. NCBI clusters of orthologous groups (COG) [66] for each gene were examined for evidence of function in the literature. All COG and UniProt sequences were aligned using MUSCLE [67]. Aligned sequences were included in a maximum likelihood tree with 50 bootstrap replications made with MEGA (v7.0, [68]). 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 [27]. All arsenic-related gene databases are freely available on FunGene (http://fungene.cme.msu.edu/).

Arsenic-related genes in cultivable soil microorganisms

The RefSoil+ database [37] was used to obtain high-quality genomes (chromosomes and plasmids) from soil microorganisms in the Genomes OnLine (GOLD) database [69]. RefSoil+ chromosomes and plasmids were searched with hmmsearch [29] using HMMs from FunGene with an *e*-value cutoff of 10^{-10}. The top hits were analyzed in R [70]. For each gene, scores and percent alignments were plotted to determine quality cutoffs. Stringent percent alignment scores were included since this search was against completed genome sequences: only hits with scores > 100 and percent alignment > 90% were included. Hits with the lowest scores were manually examined to test for false positives. Due to false positives, hits against *aioA*, *arrA*, and *arxA* were further quality filtered to have scores > 1000. When one open reading frame (ORF) contained multiple hits, the hit with a lower score was removed. Taxonomy was assigned using the RefSoil database [42], and the relative abundance of arsenic-related genes within phyla was examined. A 16S rRNA gene maximum likelihood tree of RefSoil+ bacterial strains was constructed with RAxML (v.8.0.6 [71]) 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, gene hits were extracted from RefSoil+ sequences and used to construct maximum likelihood trees for each gene.

Reference database construction

Reference gene databases of diverse, near full-length sequences were constructed using limited sequences from FunGene databases [27] 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 (Additional file 11). Briefly, minimum amino acid length 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 dropoff point. Sequences were de-replicated before being used in subsequent analysis, and final sequence counts are included in Additional file 11. Reference databases were converted to publicly available BLAST databases using BLAST+ [30]. Reference and BLAST databases are publicly available on GitHub (https://github.com/ShadeLab/PAPER_Dunivin_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, PA (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 min. Remaining soil was stored at − 80 °C until DNA extractions. The soil slurry was allowed to settle for 2 min. One hundred microliters of the slurry was then removed and serial diluted using PBS to a 10^{-2} dilution. One hundred microliters 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 [72] and the MoBio DNEasy PowerSoil Kit (MoBio, Solana Beach, CA, USA) according to 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 × 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 (Additional file 1). Datasets from Centralia, PA, were generated in our research group. All other metagenome datasets 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 datasets was ordered by the number of base pairs (bp). Metagenomic datasets 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 so that comparisons could be made both within and between soil sites. When a project had multiple samples, datasets with the greatest bp were selected. While we prioritized samples with multiple datasets, several replicate samples were omitted early on due to > 30% of data removed during quality filtering, and samples Illinois soil, Russian permafrost, and Wyoming soil have just one sample. This search ultimately yielded 26 datasets from 12 locations and 5 countries (Additional file 2).

Soil metagenome processing and gene targeted assembly

Sequences from MG-RAST datasets 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 [73]. Quality-filtered sequences were used in all downstream analyses. For each dataset, a gene targeted metagenome assembler [28] 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 amino acids) or to improve quality (i.e., maximum length is increased to improve specificity) (Additional file 11). 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

[30]. 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 [74]. Dereplicated, clustered sequences were converted into OTU tables with coverage-adjusted abundance. These tables were subsequently analyzed in R [70]. 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-related genes among datasets, total counts of *rplB* were used to normalize the abundance of each OTU. Relative abundance of arsenic-related genes was also calculated for each sample.

Additional files

Additional file 1: Available metadata and accession numbers for soil metagenomes used in this study. (DOCX 22 kb)

Additional file 2: Phylum-level summary of arsenic-related genes in RefSoil+ chromosomes and plasmids. (DOCX 14 kb)

Additional file 3: Phylogeny of Acr3 in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of Acr3 sequences predicted from RefSoil+ genomes. 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. (PNG 4395 kb)

Additional file 4: Phylogeny of ArsB in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of ArsB sequences predicted from RefSoil+ genomes. 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. (PNG 9385 kb)

Additional file 5: Phylogeny of ArsC (trx) in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of ArsC (trx) sequences predicted from RefSoil+ genomes. 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. (PNG 1911 kb)

Additional file 6: Phylogeny of ArsC (grx) in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of ArsC (grx) sequences predicted from RefSoil+ genomes. 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. (PNG 4752 kb)

Additional file 7: Phylogeny of ArsM in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of ArsM sequences predicted from RefSoil+ genomes. 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. (EPS 6021 kb)

Additional file 8: Histogram of arsenic-related gene copy numbers in RefSoil+ organisms. Total copy number is based on hits from both chromosomes and plasmids from the same organism. (EPS 24 kb)

Additional file 9: Phylum-level community structure of soil metagenomes in this study. (EPS 65 kb)

Additional file 10: Summary of endemic arsenic-related gene sequences. A sequence was considered endemic if it was present in less than three different soil sites. (DOCX 43 kb)

Additional file 11: Summary of reference arsenic resistance and metabolism gene sequences from FunGene databases. (DOCX 51 kb)

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. The funders had no role in the design of the study and collection, analysis, and interpretation of data.

Availability of data and materials

The full arsenic-related gene toolkit (BLAST databases, hidden Markov models, and gene resources for Xander) is publicly available on GitHub (https://github.com/ShadeLab/PAPER_Dunivin_meta_arsenic) [75]. Cultivation-dependent and cultivation-independent Centralia metagenomes from this study are available on NCBI under BioProject PRJNA492298 [76]. All other metagenomes are publicly available, including those from Brazilian forest [77], California grassland [78], Centralia [79], Disney preserve [80], Illinois soybean [81], Illinois switchgrass [82], Iowa agricultural [83], Iowa corn [84], Iowa prairie [85], Mangrove [86], Minnesota grassland [87], Permafrost Canada [88], Permafrost Russia [89], and Wyoming soil [90]. All RefSoil+ and metagenome analyses from this work are also available on GitHub in the HMM_search and gene_targeted_assembly directories respectively.

Authors' contributions

TKD and AS designed the study. TKD and SY contributed to code and performed the analysis. TKD and AS wrote the manuscript. All authors read and approved the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

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

Received: 19 November 2018 Accepted: 2 May 2019
Published online: 30 May 2019

References

1. Nelson MB, Martiny AC, Martiny JBH. Global biogeography of microbial nitrogen-cycling traits in soil. Proc Natl Acad Sci. 2016;113:8033–40.
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. Huang JH. Impact of microorganisms on arsenic biogeochemistry: a review. Water Air Soil Pollut. 2014;225:1848.
4. Páez-Espino D, Tamames J, De Lorenzo V, Cánovas D. Microbial responses to environmental arsenic. BioMetals. 2009;22:117–30.
5. Watanabe T, Hirano S. Metabolism of arsenic and its toxicological relevance. Arch Toxicol. 2013;87:969–79.
6. 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;46:2163–8.
7. Mukai H, Ambe Y, Muku T, Takeshita K, Fukuma T. Seasonal variation of methylarsenic compounds in airborne particulate matter. Nature. 1986;324:239–41.
8. Wang P, Sun G, Jia Y, Meharg AA, Zhu Y. A review on completing arsenic biogeochemical cycle: microbial volatilization of arsines in environment. J Environ Sci (China). 2014;26:371–81. https://doi.org/10.1016/S1001-0742(13)60432-5.
9. Jia Y, Huang H, Zhong M, Wang F, Zhang L, Zhu Y-G. Microbial arsenic methylation in soil and rice rhizosphere. Environ Sci Technol. 2013;47:3141–8. https://doi.org/10.1021/es303649v.
10. 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;7:811–5.
11. Zhu Y-G, Yoshinaga M, Zhao F-J, Rosen BP. Earth abides arsenic biotransformations. Annu Rev Earth Planet Sci. 2014;42:443–67.
12. Edwardson CF, Hollibaugh JT. Metatranscriptomic analysis of prokaryotic communities active in sulfur and arsenic cycling in mono Lake, California, USA. ISME J. 2017;11:2195–208. https://doi.org/10.1038/ismej.2017.80.
13. Andres J, Bertin PN. The microbial genomics of arsenic. FEMS Microbiol Rev. 2016;40:299–322. https://doi.org/10.1093/femsre/fuv050.
14. Rosen BP. Biochemistry of arsenic detoxification. FEBS Lett. 2002;529:86–92. https://doi.org/10.1016/s0014-5793(02)03186-1.
15. 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.
16. 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;16:647–65.
17. 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;9:820.
18. Jackson CR, Dugas SL, Harrison KG. Enumeration and characterization of arsenate-resistant bacteria in arsenic free soils. Soil Biol Biochem. 2005;37:2319–22.
19. 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:e0191893.
20. Zhu YG, Xue XM, Kappler A, Rosen BP, Meharg AA. Linking genes to microbial biogeochemical cycling: lessons from arsenic. Environ Sci Technol. 2017;51:7326–39.
21. 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. https://doi.org/10.1038/s41598-017-00896-0.
22. 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;160:193–204.
23. Rascovan N, Javier M, Martín PV, 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(July):1–11.
24. 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:e0119465.
25. 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;97:9579–88.
26. 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;9:1359.
27. Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM, et al. FunGene: the functional gene pipeline and repository. Front Microbiol. 2013;4:291.
28. 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.
29. Johnson L, Eddy S, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics. 2011:39.
30. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
31. Ajees AA, Yang J, Rosen BP. The ArsD as(III) metallochaperone. BioMetals. 2011;24:391–9.
32. 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;103:2075–80.
33. 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;185:135–41.
34. Saltikov CW, Newman DK. Genetic identification of a respiratory arsenate reductase. Proc Natl Acad Sci. 2003;100:10983–8. https://doi.org/10.1073/pnas.1834303100.
35. 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;14:1635–45.
36. Eddy SR. Hidden Markov models. Curr Opin Struct Biol. 1996;6:361–5.
37. Dunivin TK, Choi J, Howe A, Shade A. RefSoil+: a reference database for genes and traits of soil plasmids. mSystems. 2019;4. https://doi.org/10.1128/mSystems.00349-18.
38. Mikael Sehlin H, Börje Lindström E. Oxidation and reduction of arsenic by Sulfolobus acidocaldarius strain BC. FEMS Microbiol Lett. 1992;93:87–92.
39. Blomberg SP, Jr TG, Ives AR. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution. 2003;57:717–45. https://doi.org/10.1111/j.0014-3820.2003.tb00285.x.
40. 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.
41. Achour AR, Bauda P, Billard P. Diversity of arsenite transporter genes from arsenic-resistant soil bacteria. Res Microbiol. 2007;158:128–37.
42. 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;11:829–34.
43. 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;551:457–63. https://doi.org/10.1038/nature24621.
44. 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;43:1450–5. https://doi.org/10.1016/j.soilbio.2011.03.012.
45. Quemeneur M, Heinrich-Salmeron A, Muller D, Lièvremont D, Jauzein M, Bertin PN, et al. Diversity surveys and evolutionary relationships of aoxB genes in aerobic arsenite-oxidizing bacteria. Appl Environ Microbiol. 2008;74:4567–73. https://doi.org/10.1128/AEM.02851-07.
46. Inskeep WP, Macur RE, Hamamura N, Warelow TP, Ward SA, Santini JM. Detection, diversity and expression of aerobic bacterial arsenite oxidase genes. Environ Microbiol. 2007;9:934–43.

47. Sun Y, Polishchuk EA, Radoja U, Cullen WR. Identification and quantification of arsC genes in environmental samples by using real-time PCR. J Microbiol Methods. 2004;58:335–49.
48. 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;65:147–52. https://doi.org/10.1016/j.ibiod.2010.10.007.
49. Song B, Chyun E, Jaffé PR, Ward BB. Molecular methods to detect and monitor dissimilatory arsenate-respiring bacteria (DARB) in sediments. FEMS Microbiol Ecol. 2009;68:108–17.
50. Mirza BS, Sorensen DL, Dupont RR, McLean JE. New arsenate reductase gene (arrA) PCR primers for diversity assessment and quantification in environmental samples. Appl Environ Microbiol. 2017;83:e02725–16.
51. Malasarn D, Saltikov CW, Campbell KM, Santini JM, Hering JG, Newman DK. arrA is a reliable marker for as(V) respiration. Science (80-), vol. 306; 2004. p. 455.
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. 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. https://doi.org/10.1038/s41598-017-08313-2.
54. Wang L, Wang J, Jing C. Comparative genomic analysis reveals organization, function and evolution of ars genes in Pantoea spp. Front Microbiol. 2017;8:471.
55. 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;79:2471–6.
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. Bahar MM, Megharaj M, Naidu R. Bioremediation of arsenic-contaminated water: recent advances and future prospects. Water Air Soil Pollut. 2013;224:1–20.
58. Escudero LV, 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. https://doi.org/10.1371/journal.pone.0078890.
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;47:7147–54.
60. 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.
61. 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.
62. 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;150:307–22.
63. 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;79:284–92. https://doi.org/10.1016/j.marpolbul.2013.11.024.
64. 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;70:154–67.
65. The Uniprot Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45:D115–9.
66. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000;28:33–6.
67. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.
68. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33:1870–4. https://doi.org/10.1093/molbev/msw054.
69. 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:D446–56.
70. R Core Team. R: a language and environment for statistical computing. 2017.
71. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
72. 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;34:229–35.
73. Dunivin TK, Shade A. Community structure explains antibiotic resistance gene dynamics over a temperature gradient in soil. FEMS Microbiol Ecol. 2018;94:fiy016.
74. 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;73:5261–7.
75. Dunivin TK, Yeh SS, Shade A. Toolkit for detecting arsenic resistance genes: GitHub; 2019. https://github.com/ShadeLab/PAPER_Dunivin_meta_arsenic
76. Shade A. Surface soil microbial communities from active vent of coal mine fire in Centralia Pennsylvania: NCBI; 2018. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA492298
77. Bohannan B. ARMO (mgp3731). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=3731. Accessed July 28, 2017.
78. Allison SD. Loma_Ridge_grassland (mgp1992). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=1992. Accessed July 28, 2017.
79. Shade A. Surface soil microbial communities from Centralia Pennsylvania, which are recovering from an underground coalmine fire. Joint Genome Institute Integrated Microbial Genomes & Microbiomes. https://gold.jgi.doe.gov/analysis_projects?id=Ga0209810
80. Stanish LF. NEON Soil Metagenomes (mgp13948). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=13948. Accessed July 28, 2017.
81. Hartman G. ISA-SMC-2011 (mgp2076). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=2076. Accessed July 28, 2017.
82. Meyer F. Fermi-syntheticlongreads (mgp14596). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=14596. Accessed July 28, 2017.
83. Hofmockel KS. Hofmockel soil aggregate COB KBASE (mgp2592). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=2592. Accessed July 28, 2017.
84. Brown TC. GP corn unassembled (mgp6368). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=6368. Accessed July 28, 2017.
85. Brown TC. GED prairie unassembled (mgp6377). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=6377. Accessed July 28, 2017.
86. Shu-Chien A. Mining of new genes and pathways from soil of mangrove forest (mgp11628). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=11628. Accessed July 28, 2017.
87. Zak D. CedarCreek_minsoil_June2013 (mgp5588). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=5588. Accessed July 28, 2017.
88. Onstott T. Axel Heiberg permafrost: part 4A, unassembled metagenomes (mgp252). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=252. Accessed July 28, 2017.
89. Rivkina E. Permafrost sediments, North-East Siberia, Kolyma lowland (mgp7176). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=mgp7176. Accessed July 28, 2017.
90. Zak D. Ungulate Exclosure 2015 (mgp15600). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=15600. Accessed July 28, 2017.