A global survey of arsenic-related genes in soil microbiomes

Taylor K. Dunivin, Susanna Y. Yeh and Ashley Shade

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
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 to 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-associating 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., *acrB* 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.
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.
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.

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
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). 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
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...
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 \textit{arsC} (grx) and \textit{arsM} dominated most samples, their relative proportions varied greatly (Fig. 8a). RefSoil+ data suggests that \textit{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 \textit{arsC} (trx) and \textit{arrA}.

We further examined the arsenic resistance gene abundance at individual sites. We did not include \textit{arr} and \textit{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.)
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 (ArsH) [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.)
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 microorganisms 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 \textit{arsC} (trx), which is hypothesized to be more abundant in high arsenic sites [21, 58]. Additionally, \textit{arrA} encodes a dissipatory 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 \textit{acr3}, \textit{arsB}, \textit{arsC} (trx), \textit{arsD}, \textit{arsM}, and \textit{aioA} were generally endemic, suggesting regional dispersal limitation. Only one \textit{aioA} and three \textit{acr3} sequences were detected in multiple sites. This supports a previous finding that \textit{acr3} and \textit{aioA} from the acid mine drainage in Carnoulès were endemic [64]. Conversely, \textit{arsC} (grx) was cosmopolitan which could suggest genetic migration via HGT or vertical transfer and a limited gene diversification. Both are plausible since \textit{arsC} (grx) was common in ReSoil+ 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, genomics 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: \textit{acr3}, \textit{aioA}, \textit{arsB}, \textit{arsC} (grx), \textit{arsC} (trx), \textit{arsD}, \textit{arsM}, \textit{arrA}, and \textit{arxA}. FunGene databases [27] were constructed for the following arsenic-related genes: \textit{arsB}, \textit{arsC} (grx), \textit{arsC} (trx), \textit{acr3}, \textit{aioA}, \textit{arrA}, and \textit{arxA}. The \textit{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 ReSoil+ database [37] was used to obtain high-quality genomes (chromosomes and plasmids) from soil microorganisms in the Genomes OnLine (GOLD) database [69]. ReSoil+ chromosomes and plasmids were searched with hmmsearch [29] using HMMs from FunGene.cme.msu.edu/). 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 \textit{aioA}, \textit{arrA}, and \textit{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 ReSoil database [42], and the relative abundance of arsenic-related genes within phyla was examined. A 16S rRNA gene maximum likelihood tree of ReSoil+ bacterial strains was constructed with RAxML (v.8.0.6 [71]) based on the Whelan and Goldman (WAG) model with 100 bootstrap replicates ("-m PROTGAM-MAWAG -p 12345 -f a -k -x 12345 -# 100"). Based on accession numbers, gene hits were extracted from ReSoil+ 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: \textit{acr3}, \textit{aioA}, \textit{arrA}, \textit{arsB}, \textit{arsC} (grx), \textit{arsC} (trx), \textit{arsD}, \textit{arsM}, and
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 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% trypsicoy 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 trypsicoy sox broth with 25% glycerol to plates. The plate scrapings were stored at −80 °C until DNA extraction.

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% trypsicoy 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 trypsicoy sox broth with 25% glycerol to plates. The plate scrapings were stored at −80 °C until DNA extraction.

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 1). 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.
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:** Phyllum-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 (gxn) in RefSoil+ organisms. Maximum likelihood tree with 100 bootstrap replications of ArsC (gxn) 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)

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 Russell 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 R25GM151335. 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], Dinsey 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 HMW_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
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
1Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA. 2Environmental and Integrative
Toxicological Sciences Doctoral Program, Michigan State University, East Lansing, MI 48824, USA.
1Department of Plant, Soil, and Microbial Sciences, Michigan State University, East Lansing, MI 48824, USA.
2Institute for Cyber-Enabled Research, Michigan State University, East Lansing, MI 48824, USA.
3Program in Ecology, Evolutionary Biology and Behavior, Michigan State University, East Lansing, MI 48824, USA.
4Plant 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 DHI, Langille MGJ, Belko 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 Poll. 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;87:699–719.
6. Huang H, Jia Y, Sun GY, Zhu Y-G. Arsenic speciation and volatilization from flooded paddy soils amended with different organic matters. Environ Sci Technol. 2012;46:2163–8.
7. Muku H, Ambre Y, Mukti T, Takeshita K, Fukumura 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. Ja 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. Sforza MC, Philippot P, Somogyi A, Van Zulien 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 communities of genes. FEMS Microbiol Rev. 2014;38:90–118.
12. Edwards CF, Hollibaugh JT. Metatranscriptomic analysis of prokaryotic communities of organisms and interactions in the microbiome: communities of organisms and communities of genes. FEMS Microbiol Rev. 2014;38:90–118.
13. Li X, Zhang L, Wang G. Genomic evidence reveals high levels of microbial arsenic metabolism genes in activated sludge and coastal sediments. Appl Microbiol Biotechnol. 2013;97:9775–88.
14. Rosen BP. Biochemistry of arsenic detoxification. FEBS Lett. 2002;529:86–91. https://doi.org/10.1016/S0014-5793(02)03186-1.
15. Li X, Zhang L, Wang G. Genomic evidence reveals high levels of microbial arsenic metabolism genes in activated sludge and coastal sediments. Appl Microbiol Biotechnol. 2013;97:9775–88.
16. Jiang S, Li X, Wang G, Alves L, Wang G. 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.
17. Kuch R, Amadio A, Ordoñez OF, Albarracin VH, Günther W, Farias 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.
18. 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;162:193–204.
19. Rascovan N, Javier M, Martin PV, María EF. Metagenomic study of red biocrusts from Diamante Lake reveals ancient arsenic bioenergetics in halocarchea metagenomic study of red biocrusts from Diamante Lake reveals ancient arsenic bioenergetics in halocarchea. Int Soc Microb Ecol. 2016(10):July1–11.
20. 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.
21. 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. 2015;97:3141–52.
22. Hofmann SW, Hill BP, Hillier LW, Balaji A, Kajarekar T, Bozeman BL, et al. Comparative metagenomics provides insights into the ecosystem functioning of the Shark Bay Stromatolites, Western Australia. Front Microbiol. 2018;9:1359.
23. 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.
24. Wang Q, Fish JA, Gilman M, Sun Y, Brown CT, Tiedje JM, et al. Xnder: employing a novel method for efficient gene-targeted metagenomic assembly. Microbiome. 2015;3:32.
25. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:39.
26. Smrzka SP, Jr TG, Ives AR. Testing for phylogenetic signal in comparative behavioral traits are more labile. Evolution. 2003;57:717–30.
27. Blomberg SP, Jr TG, Ives AR. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution. 2003;57:717–30. https://doi.org/10.1111/j.0014-3820.2003.tb00285.x.
28. Wang Q, Fish JA, Gilman M, Sun Y, Brown CT, Tiedje JM, et al. Xnder: employing a novel method for efficient gene-targeted metagenomic assembly. Microbiome. 2015;3:32.
29. 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.0318340103.
30. Zargar K, Conrad A, Berneck DI, Lowe TM, Stok V, Hoefst S, et al. ArxA: a new clade of arsenite oxidase within the DMSO reductase family of molybdenum oxidoreductases. Environ Microbiol. 2012;14:1635–45.
31. Eddy SR. Hidden Markov models. Curr Opin Struct Biol. 1996;6:361–5.
32. Dunivin TK, Choi J, Howe A, Shade A. RefSoil+: a reference database for arsenic-related soil bacteria. Res Microbiol. 2007;158:128–37. https://doi.org/10.1016/j.resmic.2006.03.002.
33. Ajees AA, Yang J, Rosen BP. The ArsD as(III) metallochaperone. BioMetals. 2012;25:227–39. https://doi.org/10.1007/s10534-011-9594-0.
34. Blomberg SP, Jr TG, Ives AR. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution. 2003;57:717–30. https://doi.org/10.1111/j.0014-3820.2003.tb00285.x.
35. Choi J, Yang Y, Stepanauskas R, Cardenas E, Garoutte A, Williams R, et al. Strategies to improve reference databases for soil microbiomes. ISME J. 2017;11:829–34.
36. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Loeve KJ, et al. A communal catalogue reveals Earth’s multiscale microbial diversity. Nature. 2017;551:457–63. https://doi.org/10.1038/nature24621.
37. 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.2011.05.012.
38. Quinneman M, Heinrich-Salomer A, Muller D, Lievremont D, Jauzein M, Bérit P, et al. Diversity, classification and evolutionary relationships of aquA genes in aerobic arsenite-oxidizing bacteria. Appl Environ Microbiol. 2008;74:4567–73. https://doi.org/10.1128/AEM.02851-07.
39. Insepp 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, Polischuk EA, Rodajo U, Cullen WR. Identification and quantification of arsenic 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 arsenic 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, Chuyu 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. Mira B, Canbraz DL, Dupont RR, McLean JE. New arsenate reductase gene (araA) PCR primers for diversity assessment and quantification in environmental samples. Appl Environ Microbiol. 2017;83:e02725–16.

51. Malassøn M, Engstrøm K, Hallström BM, Wahlberg K, Søndergaard DA, Sall T, et al. ASAMT-mediated tolerance to arsenic evolved by multiple independent horizontal gene transfers from bacteria to eukaryotes. PLoS One. 2017;12:e1–22.

52. Chen S-C, Sun GX, Rosen BP, Zhang S-Y, Zhou B-K, et al. Recurrent horizontal transfer of arsenite methytransferase genes facilitated adaptation of life to arsenic. Sci Rep. 2017;7:741. https://doi.org/10.1038/s41598-017-08312-5.

53. Wang L, Wang J, Jing C. Comparative genomic analysis reveals organization, function and evolution of ars genes in Pantonea spp. Front Microbiol. 2017;8:471.

54. Lee S, Ruiz-Martínez M, Graves LM, Ward TJ, Sileby 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.

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

56. Bahar MM, Megharaj M, Naidu R. Bioremediation of arsenic-contaminated water: recent advances and future prospects. Water Air Soil Pollut. 2013;224:1–22.

57. Fahy A, Giloteaux L, Bertin PN, Le Paslier D, Médigue C, Weissenbach J, et al. The COG database: a tool for genomic-scale analysis of protein families and functions. Nucleic Acids Res. 2014;42:D446–55.

58. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

59. Edgar RC. Muscle: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.

60. 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/msw210.

61. Mukherjee S, Stamatis D, Bertisch J, Ovchinnikova G, Verevskaya O, Ibandy M, et al. Genomes OnLine database (GOLD): v6.6 data updates and feature enhancements. Nucleic Acids Res. 2017;45:D446–55.