Title:

A global survey of arsenic related genes in soil microbiomes

Authors

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

Affiliations

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

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

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

[4]Program in Ecology, Evolutionary Biology and Behavior, Michigan State University, East Lansing, MI 48824, USA; Department of Plant, Soil, and Microbial Sciences, Michigan State University, East Lansing, MI 48824, USA; and the Plant Resilience Institute, Michigan State University, East Lansing, MI, 48834, USA

*correspondence: shadeash@msu.edu

Abstract

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. 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. 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

Running title:

Arsenic genomics in soil microbiomes

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 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 (**Figure 1AB**)[14, 27–31]. 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 [32], functional gene (FunGene) databases [33], hidden Markov models (HMMs [34]), and gene resources for gene-targeted assembly (Xander [35]) (**Figure 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 [36]. Additionally, we used a gene-targeted assembler [35] to test 38 public soil metagenomes from Brazil, Canada, Malaysia, Russia, and the United States for arsenic resistance 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 [36]. Of the 922 RefSoil+ genomes spanning 25 phyla (**Figure 2B**;

147 **Additional File 2**), 14.3% (132 genomes) did not contain any tested arsenic related
148 genes. Of the 25 phyla in RefSoil+, two phyla (Chlamydiae and Crenarchaeota) did not
149 have any of these genes. These phyla, however, had few RefSoil+ representatives
150 (three and nine, respectively), so other members of these phyla may have arsenic
151 detoxification and metabolism genes. Supporting this hypothesis, a Crenarchaeota
152 isolate was previously reported to oxidize arsenic [37]. Nonetheless, these data suggest
153 that arsenic related genes are widespread, but not universal, even among cultivable soil
154 organisms (**Figure 2**).
155 We next asked whether 16S rRNA gene phylogeny was predictive of arsenic
156 genotypes using a test for phylogenetic signal (Bloomberg's K [38]). No phylogenetic
157 signal was observed for plasmid-borne sequences or genes encoding arsenic
158 metabolisms (*aioA*, *arrA*, *arxA*); however, relatively few RefSoil+ microorganisms tested
159 positive for these genes. Despite their phylogenetic breadth (**Additional Files 3 – 7**),
160 chromosomally-encoded *acr3*, *arsB*, *arsC* (grx), *arsC* (trx), and *arsM* were similar
161 between phylogenetically related organisms (false discovery rate adjusted $p < 0.01$;
162 **Figure 2A**).
163
164 *Phylogenetic diversity of arsenic related genes: insights into vertical and horizontal*
165 *transfer*
166
167 <u>Arsenite efflux pumps</u>. We examined the phylogenetic diversity of distinct genes
168 encoding arsenite efflux pumps, *acr3* and *arsB*, for soil-associated microorganisms
169 (**Figure 3, Additional Files 3 – 4**). Gene *acr3* is separated into two clades: *acr3*(1) and
170 *acr3*(2) [39]. Clade *acr3*(1) is typically composed of Proteobacterial sequences while
171 *acr3*(2) is typically composed of Firmicutes and Actinobacterial sequences [21, 39, 40].
172 Though RefSoil+ genomes were mostly composed of *acr3*(2) sequences from
173 Proteobacteria (**Figure 3A; Additional File 3**), we observed greater taxonomic diversity
174 observed than previously reported for this clade [21, 39, 40]. Surprisingly, there were
175 deep branches in *acr3*(2) that belonged to Bacteroidetes, Euryarchaeota, Firmicutes,
176 Fusobacteria, and Verrucomicrobia. Similarly, *acr3*(1) contained closely related *acr3*
177 sequences present in a diverse array of phyla (10 out of 25). Both clades had

178 sequences present on plasmids (6.1%). Plasmid-borne *arsB* sequences were only
179 present in Proteobacteria and Deinococcus-Thermus strains (**Figure 3B**; **Additional**
180 **File 4)**. Sequences from Actinobacteria, Proteobacteria, and Firmicutes were each
181 present in two distinct phylogenetic groups, and previous studies also observed
182 separation of *arsB* sequences based on phylum [39, 40]. Interestingly, our genome-
183 centric analysis revealed that microorganisms with multiple copies of *arsB* did not
184 harbor identical copies. For example, seven *Bacillus subtilis* subsp. *subtilis* strains had
185 two copies of *arsB*, with one from each of the two clades (**Additional File 4**).
186
187 Cytoplasmic arsenate reductases. Cytoplasmic arsenate reductase (ArsC (trx)) was
188 phylogenetically widespread in RefSoil+ microorganisms (**Figure 4A**; **Additional File**
189 **5**). While some *arsC* (trx) sequences were plasmid-borne, the majority were
190 chromosomally encoded. Similarly, plasmid encoded *arsC* (grx) made up 4.6% of
191 RefSoil+ hits (**Figure 4B**; **Additional File 6**). Notably, several Proteobacteria strains
192 have multiple copies of *arsC* (grx) with distinct sequences. It is possible that this is the
193 result of an early gene duplication event or HGT of a second *arsC* (grx).
194
195 Arsenic metabolisms. *arsM* was relatively uncommon in RefSoil+ microorganisms
196 (5.2%) (**Figure 2**). In the RefSoil+ database, *arsM* was observed in Euryarchaeota as
197 well as several bacterial phyla Acidobacteria, Actinobacteria, Armatimonadetes,
198 Bacteroidetes, Chloroflexi, Cyanobacteria, Firmicutes, Gemmatimonadetes, Nitrospirae,
199 Proteobacteria, Verrucomicrobia (**Figure 5**; **Additional File 7**). Notably, only one
200 RefSoil+ microorganism, *Rubrobacter radiotolerans* (NZ_CP007516.1), had a plasmid-
201 borne *arsM*.
202 Arsenic metabolism genes *aioA*, *arrA*, and *arxA* were phylogenetically conserved
203 (**Figure 6**). Genes encoding arsenite oxidases *aioA* and *arxA* were restricted to
204 Proteobacteria. *aioA* sequences clustered into two clades based on class-level
205 taxonomy: all Alphaproteobacteria sequences cluster separately from Gamma- and
206 Betaproteobacteria sequences. The gene encoding dissimilatory arsenate reduction
207 *arrA* was also phylogenetically conserved in RefSoil+ strains, with strains from
208 Proteobacteria clustering separate from Firmicutes (**Figure 6**).

209

210 *Cultivation bias and environmental distributions of arsenic related genes*

211 To gain a cultivation-dependent perspective of the abundances of arsenic related
212 genes in soils, we used inferred environmental abundances of RefSoil microorganisms
213 [41, 42]. The environmental abundance of RefSoil microorganisms, which are cultivable,
214 soil-associated microorganisms, was previously estimated by comparing 16S rRNA
215 gene sequences in RefSoil with those in soil metagenomes [41]. We used this
216 estimated abundance of cultivable microorganisms along with arsenic related gene
217 information from this study (**Figure 2**) to estimate the environmental abundances of
218 arsenic related genes from the cultivated bacteria. Arsenic metabolism genes (*aioA*,
219 *arrA*, *arsM*, *arxA*) were predicted to be less common in the environment compared with
220 arsenic detoxification genes (*acr3*, *arsB*, *arsC* (grx), *arsC* (trx), and *arsD*) (**Figure 7A**;
221 Mann Whitney U test $p < 0.01$). Despite similar distributions of *acr3* and *arsB* in
222 RefSoil+ (**Figure 2B**), *acr3* was more abundant in most soil orders (**Figure 7A;** Mann
223 Whitney U test $p < 0.05$). For genes encoding cytoplasmic arsenate reductases, *arsC*
224 (grx) was more abundant than *arsC* (trx) (Mann Whitney U test $p < 0.01$).

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

235 We explored cultivation bias of arsenic related genes with a case study
236 comparing cultivation-dependent (lawn growth on the standard medium TSA50) and -
237 independent communities from the same soil. Genes in the *ars* operon (*acr3*, *arsB*,
238 *arsD*, and *arsC* (trx)) were elevated in the cultivation-dependent metagenome (**Figure**
239 **7C**). Additionally, arsenic metabolism genes were not detected (*aioA*, *arrA*, *arxA*) or in

240 low abundance (*arsM*) in the cultivation-dependent sample; however, all four of these
241 arsenic metabolism genes were detected in the cultivation-independent sample. Though
242 this is a single case study of cultivation-dependent and independent methods, these
243 results recapitulate the general discrepancies between RefSoil+ genomes and soil
244 metagenomes (**Figure 7B**). This bias has important implications for studies focusing on
245 arsenic bioremediation because cultivation-dependent studies could misestimate the
246 potential of microbiomes for arsenic detoxification and metabolism *in situ*.
247
248 *Arsenic related gene endemism*
249 Arsenic related genes are globally distributed, but their biogeography is poorly
250 understood. Broadly, arsenic related genes had comparable abundance among different
251 soils (**Figure 7AB**). The relative distributions of distinct arsenic detoxification and
252 metabolism mechanisms in one site, however, are relevant for predicting the impact of
253 microbial communities on the fate of arsenic. To understand site-specific distributions,
254 we explored soil metagenomes from Brazil, Canada, Malaysia, Russia, and the United
255 States (**Additional File 1**). These 16 sites had differences in community membership
256 (**Additional File 9**) and arsenic related gene content (**Figure 8A**). Geographic location
257 was not predictive of arsenic related gene content (Mantel's $r = 0.03493$; $p > 0.05$). Soils
258 had different distributions of arsenic related genes and therefore differed in their
259 potential impact on the biogeochemical cycling of arsenic. While *arsC* (grx) and *arsM*
260 dominated most samples, their relative proportions varied greatly (**Figure 8A**). RefSoil+
261 data suggests that *arsM* can be found in Verrucomicrobia (100%, $n = 2$), which is of
262 particular importance for soil metagenomes since Verrucomicrobia are often
263 underestimated with cultivation-dependent methods [43]. The mangrove sample had the
264 most even proportions of arsenic related genes (**Figure 8A**). This distribution was
265 driven by a high abundance of *arsC* (trx) and *arrA*.
266 We further examined the arsenic resistance gene abundance at individual sites.
267 We did not include *arr* and *arx* in this analysis due to limited available data. For each
268 gene, the abundance varied greatly, but replicates within one site had similar
269 abundances (**Figure 8B**). The majority of arsenic related gene sequences (99.3%) were
270 endemic and only found in one to two sites, but 24 sequences were detected in three or

271 more sites (**Figure 8C; Additional File 10**). The majority (70.8%) of cosmopolitan
272 sequences belonged to *arsC* (grx). This analysis suggests that arsenic related genes
273 *acr3*, *arsB*, *arsC* (trx), *arsD*, *arsM,* and *aioA* are generally endemic.
274
275 **Discussion**
276 *A bioinformatic toolkit for detecting and quantifying arsenic related genes*
277 We developed a toolkit for detecting arsenic related genes from sequence data
278 that supports a variety of applications (**Figure 1A**): arsenic related genes can be
279 detected in amino acid sequences from completed genomes (HMMs [44], BLAST [32]),
280 nucleotide sequences in draft genomes (BLAST), as well as metagenomes and
281 metatranscriptomes (Xander [35]). Because each tool relies on the same seed
282 sequences, there is consistency and opportunity for comparison between sequence
283 datasets that were generated from different sources. While primers already exist for
284 arsenic related genes: *aioA* [45, 46], *acr3* [47], *arsB* [47], *arsC* (grx) [48], *arsC* (trx) [49],
285 *arsM* [9], and *arrA* [50–52], these FunGene [33] databases can be used for testing
286 primer breadth, designing new primers, and browsing sequences.
287 The toolkit is scalable for additional mechanisms for arsenic resistance and other
288 functional genes of interest (e.g., methylarsenite oxidase (ArsH), C-As lyase (ArsI),
289 trivalent organoarsenical efflux permease (ArsP), organoarsenical efflux permease
290 (ArsJ) [20]), or redox transformations of elements involved in arsenic biogeochemical
291 cycling (e.g., nitrate reductase (NarG) and sulfate reductase (DsrAB) [3, 20]). This
292 toolkit serves as both a resource and an example workflow for developing similar
293 toolkits to examine functional genes, beyond arsenic related genes, in microbial
294 sequence datasets.
295
296 *Phylogenetic diversity and distribution of arsenic related genes*
297 It has been conjectured that nearly all organisms have arsenic resistance genes
298 [20], and though this assumption has propagated in the literature, it had never been
299 explicitly quantified. Our data suggest that arsenic detoxification and metabolism genes
300 are ubiquitous, but not universal in RefSoil+ microorganisms (**Figure 2**). It is possible for
301 these 132 organisms to have untested or novel arsenic related genes; nonetheless,

302 these nine well-characterized genes were not universally detected. Additionally,
303 phylogeny was predictive of the presence of *acr3*, *arsB*, *arsC* (grc), *arsC* (trx), and
304 *arsM.* This correlation suggests that taxonomy is predictive of arsenic genotype despite
305 documented potential for HGT [19, 39, 53–55]. This result could be explained by ancient
306 rather than contemporary HGT, as seen with *arsM* [55] and *arsC* (grx) [53]. Therefore,
307 we next assessed evidence for HGT by examining the phylogenetic congruence and
308 genomic location (e.g. chromosome or plasmid) of arsenic related gene sequences.
309 Horizontal transfer of arsenic related genes has been well documented [19, 39,
310 53–57] and is an important consideration for understanding the propagation and
311 taxonomic identity of arsenic related genes. We examined the phylogenetic diversity of
312 arsenic related genes in RefSoil+ microorganisms, including plasmids and
313 chromosomes, and compared them with the 16S rRNA gene taxonomy.
314
315 Efflux pumps. While known *acr3* sequences separates into two clades [21, 39, 40],
316 plasmid-borne *acr3* sequences were present across clades, suggesting a potential for
317 transfer across unrelated taxa. Therefore, studies assigning taxonomy to *acr3* in the
318 absence of host information should consider the clade precisely and proceed with
319 caution. Despite their functional redundancy as arsenite efflux pumps, *acr3* and *arsB*
320 have very distinctive diversity. As compared with *acr3*, *arsB* was less diverse and more
321 phylogenetically conserved (**Figure 3B**; **Additional File 4**). This observation is in
322 agreement with previous reports comparing the diversity of *arsB* to *acr3* [39, 40].
323 Multiple, phylogenetically distinct copies of *arsB* were present in some RefSoil+
324 organisms, which could be due to an early gene duplication and subsequent
325 diversification or to an early transfer event. Therefore, despite relatively lower sequence
326 variation, this *arsB* phylogeny suggests an interesting evolutionary history that could be
327 investigated further.
328
329 Cytoplasmic arsenate reductases. *arsC* (trx) was predominantly found on RefSoil+
330 chromosomes, not plasmids, suggesting vertical transfer of *arsC* (trx) is common. *arsC*
331 (trx) was present in both Bacteria and Archaea, and sequences from the two domains
332 formed two distinct clades. *arsC* (trx) sequences that cluster separately from Bacterial-

333 *arsC* (trx) sequences have been documented in Thermococci, Archaeoglobi,
334 Thermoplasmata, and Halobacteria [58] Together, this distribution supports an early
335 evolutionary origin for *arsC* (trx). Thus, *arsC* (trx) appears to be an evolutionarily old
336 enzyme that is phylogenetically conserved despite its presence on plasmids and
337 potential for HGT. Plasmid-encoded *arsC* (grx) were also observed in RefSoil+
338 microorganisms, highlighting a contemporary potential for HGT that has been
339 documented in soil [53]. Thus, both genes encoding cytoplasmic arsenate reductases
340 were more common on chromosomes.
341
342 Arsenic metabolisms. The evolutionary history of the gene encoding arsenite S-
343 adenosylmethionine methyltransferase, *arsM*, was recently investigated [54, 55]. Both
344 studies independently determined that *arsM* evolved billions of years ago and was
345 subject to HGT [54, 55]. In this work, *arsM* sequences from Euryarchaeota were
346 dispersed throughout the *arsM* phylogeny, supporting the potential for inter-kingdom
347 transfer events that were recently suggested [54, 55]. Very few RefSoil+ organisms had
348 arsenic metabolism genes *aioA*, *arrA*, or *arxA*, which limits phylogenetic analysis.
349 Nonetheless, they were mostly found in Proteobacteria, which is in agreement with
350 previous work [13].
351
352 *Cultivation bias and environmental distributions of arsenic related genes*
353 Cultivation-based assessments of arsenic related gene content are important
354 since cultivable strains are often favored for bioremediation [59]. We estimated
355 distributions of arsenic related genes in cultivable microorganisms from soils and found
356 a greater abundance of arsenic detoxification genes *acr3*, *arsB*, and *arsC* (trx) (**Figure**
357 **7A**). A previous study also reported an abundance of *acr3* over *arsB* in cultivable
358 microoganisms from forest soils and attributed this to the greater phylogenetic
359 distribution of *acr3* compared with *arsB* [40]. Additionally, they found that *arsC* (grx) was
360 more abundant than *arsC* (trx) in cultivated microorganisms from these soils. It has
361 been posited in cultivation-independent studies that *arsC* (trx) is more efficient than
362 *arsC* (grx) and that high local arsenic concentrations result in a relatively greater
363 abundance of *arsC* (trx) [21, 60]. Our cultivation-dependent abundances suggest that

364 *acr3* and *arsC* (grx), rather than *arsB* and *arsC* (trx), predominantly comprise the
365 arsenic detoxification pathway in soils.
366 To assess arsenic related gene content without cultivation bias, we examined
367 arsenic related genes in soil metagenomes. As predicted by cultivable organisms,
368 arsenic metabolism genes (*aioA*, *arrA*, *arxA*) were generally in low abundance while
369 *acr3* and *arsC* (grx) were in high abundance. Estimates of genes encoding arsenic
370 detoxification (*acr3*, *arsB*, *arsD*, *arsC* (grx), *arsC* (trx)) were considerably lower in these
371 cultivation-independent samples. This result could be due, in part, to the large number
372 of RefSoil+ microorganisms with multiple copies of these genes (**Additional File 8**).
373 Cultivation-independent genomes (e.g., single-cell amplified genomes and
374 metagenome-assembled genomes) could provide greater context about the
375 environmental distributions of copy numbers of arsenic related genes.
376 Notably, *arsM* was abundant in soil (median 48%), which greatly exceeds
377 cultivation-dependent estimations, and in a case-study of cultivation dependent and
378 independent techniques, *arsM* was more abundant in the cultivation independent
379 sample (**Figure 7C**). Due to the early phylogenetic origins of *arsM* and its independent
380 functionality [55], this abundance of *arsM* in soil metagenomes is not unexpected. *arsM*
381 is typically studied in paddy soils [6, 61, 62], but metagenomes in this study suggest it is
382 an important component of the arsenic biogeochemical cycle in a variety of soils.
383
384 *Arsenic related gene endemism*
385 We examined the relative abundance of arsenic related genes in soil
386 metagenomes and observed differences in genetic potential for arsenic transformation
387 that could impact biogeochemical cycling (**Figure 8A**). Notably, the mangrove sample
388 had the most even proportions of arsenic related genes. While the arsenic
389 concentrations in this sample are unknown, mangroves are considered sources and
390 sinks for arsenic [63–65]. This could explain the greater abundance of *arsC* (trx), which
391 is hypothesized to be more abundant in high arsenic sites [21, 60]. Additionally, *arrA*
392 encodes a dissimilatory arsenate reductase that functions in an anaerobic environment
393 [30], so its greater abundance in sediment is expected. Soil geochemical data was not
394 available for all metagenomes examined in this work, so direct comparisons of arsenic

395 related gene content and soil geochemistry were not possible. This highlights the
396 importance and utility of depositing geochemical data with DNA sequences. Future
397 work, however, could further examine relationships between arsenic resistance genes
398 and soil geochemical data, including arsenic concentration and redox potential.
399 We also measured whether arsenic related gene sequence variants were
400 endemic or cosmopolitan in soil metagenomes (**Figure 8C**). We found that genes *acr3*,
401 *arsB*, *arsC* (trx), *arsD*, *arsM*, and *aioA* were generally endemic, suggesting regional
402 dispersal limitation. Only one *aioA* and three *acr3* sequences were detected in multiple
403 sites. This supports a previous finding that *acr3* and *aioA* from the acid mine drainage in
404 Carnoulès were endemic [66]. Conversely, *arsC* (grx) was cosmopolitan which could
405 suggest genetic migration via HGT or vertical transfer and a limited gene diversification.
406 Both are plausible since *arsC* (grx) was common in RefSoil+ plasmids and had low
407 phylogenetic diversity (**Figure 4B**; **Additional File 6**).
408
409 **Conclusions**
410 We developed a bioinformatic toolkit for detecting arsenic detoxification and
411 metabolism genes in microbial sequence data and applied it to analyze the genomes
412 and metagenomes from soil microorganisms. This toolkit informs hypotheses about the
413 evolutionary histories of these genes (including potential for vertical and horizontal
414 transfers) and how community ecology *in situ* may influence their prevalence and
415 distribution. This study reports the phylogenetic diversity, genomic locations, and
416 biogeography of arsenic related genes in soils, integrating information from different
417 'omics datasets and resources to provide a broad synthesis. The toolkit and the
418 synthesis presented here can catalyze future work to understand the ecology and
419 evolution of microbial arsenic biogeochemistry. Furthermore, the toolkit acts as a
420 framework for similar studies of other functional genes of interest.
421
422 **Materials and Methods**
423 *Gene Selection and Functional Gene (FunGene) Database Construction*
424 Marker genes can be used to estimate their potential to influence the arsenic
425 biogeochemical cycle [21, 25], so we selected nine well-characterized genes: *acr3*,

426 *aioA*, *arsB*, *arsC* (grx), *arsC* (trx), *arsD*, *arsM*, *arrA*, and *arxA*. FunGene databases [33]
427 were constructed for the following arsenic related genes: *arsB*, *arsC* (grx), *arsC* (trx),
428 *acr3*, *aioA*, *arrA*, and *arxA*. The *arxA* database was constructed with seed sequences
429 from [12]. For all other genes, UniProt [67] was used to obtain full length, reviewed
430 sequences when possible. NCBI clusters of orthologous groups (COG) [68] for each
431 gene were examined for evidence of function in the literature. All COG and UniProt
432 sequences were aligned using MUSCLE [69]. Aligned sequences were included in a
433 maximum likelihood tree with 50 bootstrap replications made with MEGA (v7.0,[70]).
434 Sequences that did not cluster with known sequences and had no evidence of function
435 were removed. A final FASTA file for each gene was submitted to the Ribosomal
436 Database Project (RDP) to construct a FunGene database [33]. All arsenic related gene
437 databases are freely available on FunGene (http://fungene.cme.msu.edu/).
438
439 *Arsenic related genes in cultivable soil microorganisms*
440 The RefSoil+ database [36] was used to obtain high-quality genomes
441 (chromosomes and plasmids) from soil microorganisms in the Genomes OnLine
442 (GOLD) database [71]. RefSoil+ chromosomes and plasmids were searched with
443 hmmsearch [44] using HMMs from FunGene with an e-value cutoff of 10^{-10}. The top hits
444 were analyzed in R [72]. For each gene, scores and percent alignments were plotted to
445 determine quality cutoffs. Stringent percent alignment scores were included since this
446 search was against completed genome sequences: only hits with scores > 100 and
447 percent alignment > 90% were included. Hits with the lowest scores were manually
448 examined to test for false positives. Due to false positives, hits against *aioA*, *arrA*, and
449 *arxA* were further quality filtered to have scores > 1,000. When one open reading frame
450 (ORF) contained multiple hits, the hit with a lower score was removed. Taxonomy was
451 assigned using the RefSoil database [41], and the relative abundance of arsenic related
452 genes within phyla were examined. A 16S rRNA gene maximum likelihood tree of
453 RefSoil+ bacterial strains was constructed with RAxML (v.8.0.6 [73]) based on the
454 Whelan and Goldman (WAG) model with 100 bootstrap replicates ("-m
455 PROTGAMMAWAG -p 12345 -f a -k -x 12345 -# 100"). Based on accession numbers,

456 gene hits were extracted from RefSoil+ sequences and used to construct maximum
457 likelihood trees for each gene.
458
459 *Reference Database Construction*
460 Reference gene databases of diverse, near full length sequences were
461 constructed using limited sequences from FunGene databases [33] for the following
462 genes: *acr3*, *aioA*, *arrA*, *arsB*, *arsC* (grx), *arsC* (trx), *arsD*, *arsM*, and *arxA*. Seed
463 sequences and hidden Markov models (HMMs) for each gene were downloaded from
464 FunGene, and diverse protein and corresponding nucleotide sequences were selected
465 with gene-specific search parameters (**Additional File 11**). Briefly, minimum amino acid
466 length was set to 70% of the HMM length; minimum HMM coverage was set to 80% as
467 is recommended by Xander software for targeted gene assembly; and a score cutoff
468 was manually selected based on a drop off point. Sequences were de-replicated before
469 being used in subsequent analysis, and final sequence counts are included in
470 **Additional File 11**. Reference databases were converted to publicly available BLAST
471 databases using BLAST+ [74]. Reference and BLAST databases are publicly available
472 on GitHub (https://github.com/ShadeLab/PAPER_Dunivin_meta_arsenic).
473
474 *Sample collection and preparation*
475 A soil surface core (20 cm depth and 5.1 cm diameter) was collected in October
476 2014 from Centralia, PA (GPS coordinates: 40 48.070, 076 20.574). For cultivation-
477 dependent work, a soil slurry was made by vortexing 5 g soil with 25 mL phosphate-
478 buffered saline (PBS) for 1 min. Remaining soil was stored at -80°C until DNA
479 extractions. The soil slurry was allowed to settle for 2 min. 100 µL of the slurry was then
480 removed and serial diluted using PBS to a 10^{-2} dilution. 100 µL of the solution was
481 added to 50% trypticase soy agar (TSA50) with 200 µg/ml cycloheximide to prevent
482 fungal growth. Plates were incubated at 60°C for 72 h. Lawns of growth were extracted
483 by adding 600 µL trypticase soy broth with 25% glycerol to plates. The plate scrapings
484 were stored at -80°C until DNA extraction.
485
486 *DNA extraction and metagenome sequencing*

487 DNA for cultivation-independent analysis was manually extracted from soil using
488 a phenol chloroform extraction [75] and the MoBio DNEasy PowerSoil Kit (MoBio,
489 Solana Beach, CA, USA) according the manufacturer's instructions. DNA extraction for
490 cultivation-dependent analysis was performed in triplicate from 200 µL of plate
491 scrapings using the E.Z.N.A. Bacterial DNA Kit according to the manufacturer's
492 instructions. All DNA was quantified using a Qubit dsDNA BR Assay Kit (Life
493 Technologies, NY, USA) and was submitted for NGS library prep and sequencing at the
494 Michigan State University Genomics Core sequencing facility (East Lansing, MI, USA).
495 Libraries were prepared using the Illumina TruSeq Nano DNA Library Preparation Kit.
496 After QC and quantitation, the libraries were pooled and loaded on one lane of an
497 Illumina HiSeq 2500 Rapid Run flow cell (v1). Sequencing was performed in a 2 x 150
498 bp paired end format using Rapid SBS reagents. Base calling was performed by
499 Illumina Real Time Analysis (RTA) v1.18.61 and output of RTA was demultiplexed and
500 converted to FastQ format with Illumina Bcl2Fastq v1.8.4.
501
502 *Public soil metagenome acquisition*
503 In total, 38 soil metagenomes were obtained for this work (**Additional File 1**).
504 Datasets from Centralia, PA, were generated in our research group. All other
505 metagenome data sets were obtained from MG-RAST ((http://metagenomics.anl.gov/).
506 The MG-RAST database was searched on May 15, 2017, with the following criteria:
507 material = soil, sequence type = shotgun, public = true. The resulting list of
508 metagenome data sets was ordered by number of base pairs (bp). Metagenomic data
509 sets with the most bp were only included if they were sequenced using Illumina to
510 standardize sequencing errors, had an available FASTQ file for internal quality control,
511 and contained < 30% low quality as determined by MG-RAST. Within high quality
512 Illumina samples, priority for inclusion was given to projects with multiple samples so
513 that comparisons could be made both within and between soil sites. When a project had
514 multiple samples, data sets with the greatest bp were selected. While we prioritized
515 samples with multiple datasets, several replicate samples were omitted early on due to
516 > 30% of data removed during quality filtering, and samples Illinois soil, Russian

517 permafrost, and Wyoming soil have just one sample. This search ultimately yielded 26
518 data sets from 12 locations and five countries (**Additional File 2**).
519
520 *Soil metagenome processing and gene targeted assembly*
521 Sequences from MG-RAST data sets as well as Centralia sample Cen13 were
522 quality controlled using the FASTX toolkit (fastq_quality_filter, "-Q33 -q 30 -p 50").
523 Twelve datasets from Centralia, PA, were obtained from the Joint Genome Institute and
524 quality filtered as described previously [76]. Quality filtered sequences were used in all
525 downstream analyses. For each data set, a gene targeted metagenome assembler [35]
526 was used to assemble each gene of interest. For each gene of interest, seed
527 sequences, HMMs, and reference gene databases described above were included. For
528 *rplB,* reference gene database, seed sequences, and HMMs from the Xander package
529 were used. In most instances, default assembly parameters were used except to
530 incorporate differences in protein length (i.e. protein is shorter than default 150 amino
531 acids) or to improve quality (i.e. maximum length is increased to improve specificity)
532 (**Additional File 11**). While the assembler includes chimera removal, additional quality
533 control steps were added. Final assembled sequences (operational taxonomic units,
534 OTUs) were searched against the reference gene database as well as the non-
535 redundant database (nr) from NCBI (August 28, 2017) using BLAST [74]. Genes were
536 re-examined if the top hit had an e-value > 10^{-5} or if top hit descriptors were not the
537 target gene. Genes with low quality results were re-assembled with adjusted
538 parameters.
539
540 *Soil metagenome comparison*
541 To compare assembled sequences between samples, gene-based OTU tables
542 were constructed. Aligned sequences from each sample were dereplicated and
543 clustered at 90 amino acid identity using the RDP Classifier [77]. Dereplicated, clustered
544 sequences were converted into OTU tables with coverage-adjusted abundance. These
545 tables were subsequently analyzed in R [72]. RplB OTUs were used to compare
546 community structure. The six most abundant phyla were extracted to include at least
547 75% of each community; the full community structure is available. To compare the

548 abundance of arsenic related genes among data sets, total counts of *rplB* were used to
549 normalize the abundance of each OTU. Relative abundance of arsenic related genes
550 was also calculated for each sample.
551
552 **Declarations**
553 *Ethics approval and consent to participate*
554 Not applicable.

555 *Consent to publish*

556 Not applicable.
557
558 *Availability of data and materials*
559 The full arsenic related gene toolkit (BLAST databases, hidden Markov models,
560 and gene resources for Xander) is publicly available on GitHub
561 (https://github.com/ShadeLab/PAPER_Dunivin_meta_arsenic) [78]. Cultivation
562 dependent and cultivation independent Centralia metagenomes from this study are
563 available on NCBI under BioProject PRJNA492298 [79]. All other metagenomes are
564 publicly available, including those from Brazilian forest [80], California grassland [81],
565 Centralia [82], Disney preserve [83], Illinois soybean [84], Illinois switchgrass [85], Iowa
566 agricultural [86], Iowa corn [87], Iowa prairie [88], Mangrove [89], Minnesota grassland
567 [90], Permafrost Canada [91], Permafrost Russia [92], and Wyoming soil [93]. All
568 RefSoil+ and metagenome analyses from this work are also available on GitHub in the
569 HMM_search and gene_targeted_assembly directories respectively.

570 *Competing interests*

571 The authors declare no competing interests.

572 *Funding*

573 Metagenome sequencing was supported by the Joint Genome Institute Community
574 Science Project #1834 and by Michigan State University. The work conducted by the
575 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.

Authors' contributions

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

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.

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.

606 4. Páez-Espino D, Tamames J, De Lorenzo V, Cánovas D. Microbial responses to
607 environmental arsenic. BioMetals. 2009;22:117–30.
608 5. Watanabe T, Hirano S. Metabolism of arsenic and its toxicological relevance. Arch
609 Toxicol. 2013;87:969–79.
610 6. Huang H, Jia Y, Sun G-X, Zhu Y-G. Arsenic Speciation and Volatilization from
611 Flooded Paddy Soils Amended with Different Organic Matters. Environ Sci Technol.
612 2012;46:2163–8.
613 7. Mukai H, Ambe Y, Muku T, Takeshita K, Fukuma T. Seasonal variation of
614 methylarsenic compounds in airborne particulate matter. Nature. 1986;324:239–41.
615 8. Wang P, Sun G, Jia Y, Meharg AA, Zhu Y. A review on completing arsenic
616 biogeochemical cycle: Microbial volatilization of arsines in environment. J Environ Sci
617 (China). 2014;26:371–81. doi:10.1016/S1001-0742(13)60432-5.
618 9. Jia Y, Huang H, Zhong M, Wang F, Zhang L, Zhu Y-G. Microbial arsenic methylation
619 in soil and rice rhizosphere. Environ Sci Technol. 2013;47:3141–8.
620 doi:10.1021/es303649v.
621 10. Sforna MC, Philippot P, Somogyi A, Van Zuilen MA, Medjoubi K, Schoepp-Cothenet
622 B, et al. Evidence for arsenic metabolism and cycling by microorganisms 2.7 billion
623 years ago. Nat Geosci. 2014;7:811–5.
624 11. Zhu Y-G, Yoshinaga M, Zhao F-J, Rosen BP. Earth Abides Arsenic
625 Biotransformations. Annu Rev Earth Planet Sci. 2014;42.
626 12. Edwardson CF, Hollibaugh JT. Metatranscriptomic analysis of prokaryotic
627 communities active in sulfur and arsenic cycling in Mono Lake, California, USA. ISME J.
628 2017;11:2195–208. doi:10.1038/ismej.2017.80.
629 13. Andres J, Bertin PN. The microbial genomics of arsenic. FEMS Microbiol Rev.
630 2016;40:299–322. doi:10.1093/femsre/fuv050.
631 14. Rosen BP. Biochemistry of arsenic detoxification. FEBS Lett. 2002;529:86–92.
632 doi:10.1016/s0014-5793(02)03186-1.
633 15. Li X, Zhang L, Wang G. Genomic evidence reveals the extreme diversity and wide
634 distribution of the arsenic-related genes in Burkholderiales. PLoS One. 2014;9:1–11.
635 16. Crognale S, Amalfitano S, Casentini B, Fazi S, Petruccioli M, Rossetti S. Arsenic-
636 related microorganisms in groundwater: a review on distribution, metabolic activities and

637 potential use in arsenic removal processes. Rev Environ Sci Biotechnol. 2017;16:647–
638 65.
639 17. Plewniak F, Crognale S, Rossetti S, Bertin PN, Marco DE, Pelaez AI. A Genomic
640 Outlook on Bioremediation : The Case of Arsenic Removal. Front Microbiol. 2018;9.
641 18. Jackson CR, Dugas SL, Harrison KG. Enumeration and characterization of
642 arsenate-resistant bacteria in arsenic free soils. Soil Biol Biochem. 2005;37:2319–22.
643 19. Dunivin TK, Miller J, Shade A. Taxonomically-linked growth phenotypes during
644 arsenic stress among arsenic resistant bacteria isolated from soils overlying the
645 Centralia coal seam fire. PLoS One. 2018;13.
646 20. Zhu YG, Xue XM, Kappler A, Rosen BP, Meharg AA. Linking Genes to Microbial
647 Biogeochemical Cycling: Lessons from Arsenic. Environ Sci Technol. 2017;51:7326–39.
648 21. Kurth D, Amadio A, Ordoñez OF, Albarracín VH, Gärtner W, Farías ME. Arsenic
649 metabolism in high altitude modern stromatolites revealed by metagenomic analysis.
650 Sci Rep. 2017;7:1024. doi:10.1038/s41598-017-00896-0.
651 22. Chi L, Bian X, Gao B, Tu P, Ru H, Lu K. The effects of an environmentally relevant
652 level of arsenic on the gut microbiome and its functional metagenome. Toxicol Sci.
653 2017;160.
654 23. Rascovan N, Javier M, Martín P V, María EF. Metagenomic study of red biofilms
655 from Diamante Lake reveals ancient arsenic bioenergetics in haloarchaea Metagenomic
656 study of red biofilms from Diamante Lake reveals ancient arsenic bioenergetics in
657 haloarchaea. Int Soc Microb Ecol. 2016;109 July:1–11.
658 24. Costa PS, Reis MP, Avila MP, Leite LR, De Araujo FMG, Salim ACM, et al.
659 Metagenome of a microbial community inhabiting a metal-rich tropical stream sediment.
660 PLoS One. 2015;10.
661 25. Cai L, Yu K, Yang Y, Chen BW, Li XD, Zhang T. Metagenomic exploration reveals
662 high levels of microbial arsenic metabolism genes in activated sludge and coastal
663 sediments. Appl Microbiol Biotechnol. 2013;97:9579–88.
664 26. Babilonia J, Conesa A, Casaburi G, Pereira C, Louyakis AS, Reid RP, et al.
665 Comparative metagenomics provides insight into the ecosystem functioning of the
666 Shark Bay Stromatolites, Western Australia. Front Microbiol. 2018;9.
667 27. Ajees AA, Yang J, Rosen BP. The ArsD as(III) metallochaperone. BioMetals.

668 2011;24:391–9.
669 28. Qin J, Rosen BP, Zhang Y, Wang G, Franke S, Rensing C. Arsenic detoxification
670 and evolution of trimethylarsine gas by a microbial arsenite S-adenosylmethionine
671 methyltransferase. Proc Natl Acad Sci. 2006;103:2075–80.
672 29. Muller D, Lièvremont D, Simeonova DD, Jean-Claude H, Lett M-C. Arsenite oxidase
673 aox genes from a metal-resistant beta-proteobacterium. J Bacteriol. 2003;185:135–41.
674 30. Saltikov CW, Newman DK. Genetic identification of a respiratory arsenate
675 reductase. Proc Natl Acad Sci. 2003;100:10983–8. doi:10.1073/pnas.1834303100.
676 31. Zargar K, Conrad A, Bernick DL, Lowe TM, Stolc V, Hoeft S, et al. ArxA, a new
677 clade of arsenite oxidase within the DMSO reductase family of molybdenum
678 oxidoreductases. Environ Microbiol. 2012;14:1635–45.
679 32. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.
680 BLAST+: Architecture and applications. BMC Bioinformatics. 2009;10:1–9.
681 33. Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM, et al. FunGene: The
682 functional gene pipeline and repository. Front Microbiol. 2013;4.
683 34. Eddy SR. Hidden Markov models. Curr Opin Struct Biol. 1996;6:361–5.
684 35. Wang Q, Fish JA, Gilman M, Sun Y, Brown CT, Tiedje JM, et al. Xander: employing
685 a novel method for efficient gene-targeted metagenomic assembly. Microbiome. 2015;3.
686 36. Dunivin TK, Choi J, Howe A, Shade A. RefSoil+: a Reference Database for Genes
687 and Traits of Soil Plasmids. mSystems. 2019;4. doi:10.1128/mSystems.00349-18.
688 37. Mikael Sehlin H, Börje Lindström E. Oxidation and reduction of arsenic by
689 Sulfolobus acidocaldarius strain BC. FEMS Microbiol Lett. 1992;93:87–92.
690 38. Blomberg SP, Jr TG, Ives AR. Testing for phylogenetic signal in comparative data:
691 behavioral traits are more labile. Evolution. 2003;57:717–45. doi:10.1111/j.0014-
692 3820.2003.tb00285.x.
693 39. Cai L, Liu G, Rensing C, Wang G. Genes involved in arsenic transformation and
694 resistance associated with different levels of arsenic-contaminated soils. BMC Microbiol.
695 2009;9:4.
696 40. Achour AR, Bauda P, Billard P. Diversity of arsenite transporter genes from arsenic-
697 resistant soil bacteria. Res Microbiol. 2007;158:128–37.
698 41. Choi J, Yang F, Stepanauskas R, Cardenas E, Garoutte A, Williams R, et al.

699 Strategies to improve reference databases for soil microbiomes. ISME J. 2017;11:829–
700 34.
701 42. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A
702 communal catalogue reveals Earth's multiscale microbial diversity. Nature.
703 2017;551:457–63. doi:10.1038/nature24621.
704 43. Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, Walters WA, et al.
705 The under-recognized dominance of Verrucomicrobia in soil bacterial communities. Soil
706 Biol Biochem. 2011;43:1450–5. doi:10.1016/j.soilbio.2011.03.012.
707 44. Johnson L, Eddy S, Portugaly E. Hidden Markov Model Speed Heuristic and
708 Iterative HMM Search Procedure. BMC Bioinformatics. 2011;39.
709 45. Quemeneur M, Heinrich-Salmeron A, Muller D, Lièvremont D, Jauzein M, Bertin PN,
710 et al. Diversity Surveys and Evolutionary Relationships of aoxB Genes in Aerobic
711 Arsenite-Oxidizing Bacteria. Appl Environ Microbiol. 2008;74:4567–73.
712 doi:10.1128/AEM.02851-07.
713 46. Inskeep WP, Macur RE, Hamamura N, Warelow TP, Ward SA, Santini JM.
714 Detection, diversity and expression of aerobic bacterial arsenite oxidase genes. Environ
715 Microbiol. 2007;9:934–43.
716 47. Achour AR, Bauda P, Billard P. Diversity of arsenite transporter genes from arsenic-
717 resistant soil bacteria. Res Microbiol. 2007;158:128–37.
718 48. Sun Y, Polishchuk EA, Radoja U, Cullen WR. Identification and quantification of
719 arsC genes in environmental samples by using real-time PCR. J Microbiol Methods.
720 2004;58:335–49.
721 49. Villegas-Torres MF, Bedoya-Reina OC, Salazar C, Vives-Florez MJ, Dussan J.
722 Horizontal arsC gene transfer among microorganisms isolated from arsenic polluted
723 soil. Int Biodeterior Biodegrad. 2011;65:147–52. doi:10.1016/j.ibiod.2010.10.007.
724 50. Song B, Chyun E, Jaffé PR, Ward BB. Molecular methods to detect and monitor
725 dissimilatory arsenate-respiring bacteria (DARB) in sediments. FEMS Microbiol Ecol.
726 2009;68:108–17.
727 51. Mirza BS, Sorensen DL, Dupont RR, McLean JE. New arsenate reductase gene
728 (arrA) PCR primers for diversity assessment and quantification in environmental
729 samples. Appl Environ Microbiol. 2017;83.

730 52. Malasarn D, Saltikov CW, Campbell KM, Santini JM, Hering JG, Newman DK. arrA
731 Is a Reliable Marker for As(V) Respiration. Science (80-). 2004;306:455.
732 53. Villegas-Torres MF, Bedoya-Reina OC, Salazar C, Vives-Florez MJ, Dussan J.
733 Horizontal arsC gene transfer among microorganisms isolated from arsenic polluted
734 soil. Int Biodeterior Biodegrad. 2011;65:147–52.
735 54. Palmgren M, Engström K, Hallström BM, Wahlberg K, Søndergaard DA, Sall T, et
736 al. AS3MT-mediated tolerance to arsenic evolved by multiple independent horizontal
737 gene transfers from bacteria to eukaryotes. PLoS One. 2017;12:1–22.
738 55. Chen S-C, Sun G-X, Rosen BP, Zhang S-Y, Deng Y, Zhu B-K, et al. Recurrent
739 horizontal transfer of arsenite methyltransferase genes facilitated adaptation of life to
740 arsenic. Sci Rep. 2017;7:7741. doi:10.1038/s41598-017-08313-2.
741 56. Wang L, Wang J, Jing C. Comparative genomic analysis reveals organization,
742 function and evolution of ars genes in Pantoea spp. Front Microbiol. 2017;8.
743 57. Lee S, Rakic-Martinez M, Graves LM, Ward TJ, Siletzky RM, Kathariou S. Genetic
744 determinants for cadmium and arsenic resistance among Listeria monocytogenes
745 serotype 4B isolates from sporadic human listeriosis patients. Appl Environ Microbiol.
746 2013;79:2471–6.
747 58. Jackson CR, Dugas SL. Phylogenetic analysis of bacterial and archaeal arsC gene
748 sequences suggests an ancient, common origin for arsenate reductase. BMC Evol Biol.
749 2003;3:18.
750 59. Bahar MM, Megharaj M, Naidu R. Bioremediation of arsenic-contaminated water:
751 Recent advances and future prospects. Water Air Soil Pollut. 2013;224:1–20.
752 60. Escudero L V, Casamayor EO, Chong G, Pedrós-Alió C, Demergasso C.
753 Distribution of microbial arsenic reduction, oxidation and extrusion genes along a wide
754 range of environmental arsenic concentrations. PLoS One. 2013;8:e78890.
755 doi:10.1371/journal.pone.0078890.
756 61. Zhao FJ, Harris E, Yan J, Ma J, Wu L, Liu W, et al. Arsenic methylation in soils and
757 its relationship with microbial arsM abundance and diversity, and As speciation in rice.
758 Environ Sci Technol. 2013;47:7147–54.
759 62. Qiao JT, Li XM, Hu M, Li FB, Young LY, Sun WM, et al. Transcriptional Activity of
760 Arsenic-Reducing Bacteria and Genes Regulated by Lactate and Biochar during

761 Arsenic Transformation in Flooded Paddy Soil. Environ Sci Technol. 2018;52:61–70.
762 63. Li R, Chai M, Qiu GY. Distribution, fraction, and ecological assessment of heavy
763 metals in sediment-plant system in mangrove forest, South China Sea. PLoS One.
764 2016;11:1–15.
765 64. Chatterjee M, Massolo S, Sarkar SK, Bhattacharya AK, Bhattacharya BD, Satpathy
766 KK, et al. An assessment of trace element contamination in intertidal sediment cores of
767 Sunderban mangrove wetland, India for evaluating sediment quality guidelines. Environ
768 Monit Assess. 2009;150:307–22.
769 65. Chaudhuri P, Nath B, Birch G. Accumulation of trace metals in grey mangrove
770 Avicennia marina fine nutritive roots: The role of rhizosphere processes. Mar Pollut Bull.
771 2014;79:284–92. doi:10.1016/j.marpolbul.2013.11.024.
772 66. Fahy A, Giloteaux L, Bertin PN, Le Paslier D, Médigue C, Weissenbach J, et al. 16S
773 rRNA and As-Related Functional Diversity: Contrasting Fingerprints in Arsenic-Rich
774 Sediments from an Acid Mine Drainage. Microb Ecol. 2015;70:154–67.
775 67. The Uniprot Consortium T. UniProt: the universal protein knowledgebase. Nucleic
776 Acids Res. 2017;45.
777 68. Tatusov RL, Galperin MY, Natale DA, Koonin E V. The COG database: a tool for
778 genome-scale analysis of protein functions and evolution. Nucleic Acids Res.
779 2000;28:33–6.
780 69. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and
781 space complexity. BMC Bioinformatics. 2004;5:113.
782 70. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis
783 version 7.0 for bigger datasets. Mol Biol Evol. 2016;33:1870–4.
784 doi:10.1093/molbev/msw054.
785 71. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Verezemska O, Isbandi M, et
786 al. Genomes OnLine Database (GOLD) v.6: Data updates and feature enhancements.
787 Nucleic Acids Res. 2017;45:D446–56.
788 72. R Core Team. R: A Language and Environment for Statistical Computing. 2017.
789 73. Stamatakis A. RAxML Version 8: A tool for phylogenetic analysis and post-analysis
790 of large phylogenies. Bioinformatics. 2014;30.
791 74. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.

792 BLAST+: architecture and applications. BMC Bioinformatics. 2008;10.
793 75. Cho J-C, Lee D-H, Cho Y-C, Cho J-C, Kim S-J. Direct Extraction of DNA from Soil
794 for Amplification of 16S rRNA Gene Sequences by Polymerase Chain Reaction. J
795 Microbiol. 2006;34:229–35.
796 76. Dunivin TK, Shade A. Community structure explains antibiotic resistance gene
797 dynamics over a temperature gradient in soil. FEMS Microbiol Ecol. 2018;94.
798 77. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid
799 assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ
800 Microbiol. 2007;73.
801 78. Dunivin TK, Yeh SS, Shade A. Toolkit for detecting arsenic resistance genes.
802 GitHub. 2019. https://github.com/ShadeLab/PAPER_Dunivin_meta_arsenic.
803 79. Shade A. Surface soil microbial communities from active vent of coal mine fire in
804 Centralia Pennsylvania. NCBI. 2018.
805 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA492298.
806 80. Bohannan B. ARMO (mgp3731). MG-RAST. http://www.mg-
807 rast.org/linkin.cgi?project=3731.
808 81. Allison SD. Loma_Ridge_grassland (mgp1992). MG-RAST. http://www.mg-
809 rast.org/linkin.cgi?project=1992.
810 82. Shade A. Surface soil microbial communities from Centralia Pennsylvania, which
811 are recovering from an underground coalmine fire. Joint Genome Institute Integrated
812 Microbial Genomes & Microbiomes.
813 83. Stanish LF. NEON Soil Metagenomes (mgp13948). MG-RAST. http://www.mg-
814 rast.org/linkin.cgi?project=13948.
815 84. Hartman G. ISA-SMC-2011 (mgp2076). MG-RAST. http://www.mg-
816 rast.org/linkin.cgi?project=2076.
817 85. Meyer F. Fermi-syntheticlongreads (mgp14596). MG-RAST. http://www.mg-
818 rast.org/linkin.cgi?project=14596.
819 86. Hofmockel KS. Hofmockel Soil Aggregate COB KBASE (mgp2592). MG-RAST.
820 http://www.mg-rast.org/linkin.cgi?project=2592.
821 87. Brown TC. GP corn unassembled (mgp6368). MG-RAST. http://www.mg-
822 rast.org/linkin.cgi?project=6368.

88. Brown TC. GED prairie unassembled (mgp6377). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=6377.
89. 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.
90. Zak D. CedarCreek_minsoil_June2013 (mgp5588). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=5588.
91. Onstott T. Axel Heiberg Permafrost: Part 4A, Unassembled Metagenomes (mgp252). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=252.
92. Rivkina E. Permafrost sediments, North-East Siberia, Kolyma lowland (mgp7176). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=mgp7176.
93. Zak D. Ungulate Exclosure 2015 (mgp15600). MG-RAST. http://www.mg-rast.org/linkin.cgi?project=15600.

837 **Figure Legends**

838

839 **Figure 1. Arsenic resistance and metabolism gene toolkit schematic. A)** Seed
840 sequences for nine arsenic resistance genes were used to construct an arsenic
841 resistance gene database with existing tools [33, 35, 44, 74]. Lines indicate
842 interdependence between modules. **B)** Table of arsenic resistance and metabolism
843 genes included in the toolkit. The toolkit is freely available on GitHub:
844 github.com/ShadeLab/meta_arsenic

845

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

856

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

863

864 **Figure 4. Phylogeny of cytoplasmic arsenate reductases in RefSoil+ organisms.**
865 Maximum likelihood tree with 100 bootstrap replications of **A)** ArsC (trx) and **B)**
866 ArsC (grx) sequences predicted from RefSoil+ genomes. Tree scale = 1. Leaf tip
867 color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by

868 black circles within the tree. Grey circles on the exterior of the tree indicate that a
869 hit was detected on a plasmid and not a chromosome.
870
871 **Figure 5. Phylogeny of ArsM in RefSoil+ organisms.** Maximum likelihood tree with
872 100 bootstrap replications of ArsM sequences predicted from RefSoil+ genomes.
873 Tree scale = 1. Leaf tip color indicates phylum-level taxonomy. Bootstrap values >
874 50 are represented by black circles within the tree. Grey circles on the exterior of
875 the tree indicate that a hit was detected on a plasmid and not a chromosome.
876
877 **Figure 6. Phylogeny of AioA, ArrA, and ArxA in RefSoil+ organisms.** Maximum
878 likelihood tree with 100 bootstrap replications of dissimilatory arsenic resistance
879 proteins predicted from RefSoil+ genomes. Tree scale = 0.1. Leaf tips show the
880 name of the RefSoil+ organisms and background color indicates phylum-level
881 taxonomy. Bootstrap values > 50 are represented by black circles within the tree.
882
883 **Figure 7. Comparison of arsenic resistance and metabolism gene abundance**
884 **between cultivation dependent and cultivation independent methods. A)**
885 Mean normalized abundance of arsenic related genes based on RefSoil
886 microorganisms abundance estimated from corresponding 16S rRNA gene
887 abundance in Earth Microbiome Project datasets. Points are colored by soil order.
888 **B)** Normalized abundance of arsenic resistance genes in RefSoil+ and 38
889 metagenomes. Metagenome abundance was normalized to *rplB*, and RefSoil+
890 normalized abundance was calculated using the number of RefSoil+ genomes.
891 Only metagenomes with an arsenic resistance gene detected are shown, and the
892 total number of datasets (including RefSoil+) is shown in parentheses. **C)** *rplB*-
893 normalized abundance of arsenic resistance genes in cultivation dependent and
894 independent metagenomes from the same soil sample.
895
896 **Figure 8. Arsenic resistance and metabolism gene biogeography. A)** Relative
897 abundance of arsenic resistance genes in soil metagenomes. **B)** Rank *rplB*-
898 normalized abundance of arsenic related genes in soil metagenomes. Sites are

899 ordered by rank mean abundance. Note the differences in y axes. **C)**
900 Abundance-occurrence plots of arsenic related gene sequences clustered at
901 90% amino acid identity. Number of samples included are as follows Brazilian
902 forest n = 3, California grassland n = 2, Centralia active n = 7, Centralia
903 recovered n = 5, Centralia reference n = 1, Disney preserve n = 2, Illinois
904 soybean n = 2, Illinois switchgrass n = 1, Iowa agricultural n = 2, Iowa corn n = 2,
905 Iowa prairie n = 3, Mangrove n = 2, Minnesota grassland n = 2, Permafrost
906 Canada n = 2, Permafrost Russia n = 1, and Wyoming soil n = 1.
907

908 **Additional files legends**

909 **Additional File 1**. Available metadata and accession numbers for soil metagenomes
910 used in this study.

911 **Document Name**: Additional_File_1.docx
912 **Document Format**: Word
913 **Document Size**: 23 KB
914

915

916 **Additional File 2.** Phylum-level summary of arsenic related genes in RefSoil+
917 chromosomes and plasmids.

918 **Document Name**: Additional_File_2.docx
919 **Document Format**: Excel
920 **Document Size**: 15 KB
921

922

923 **Additional File 3. Phylogeny of Acr3 in RefSoil+ organisms.** Maximum likelihood tree
924 with 100 bootstrap replications of Acr3 sequences predicted from RefSoil+ genomes. Leaf
925 tips show the name of the RefSoil+ organisms and background color indicates phylum-
926 level taxonomy. Bootstrap values > 50 are represented by black circles within the tree.

927 **Document Name**: Additional_File_3.png
928 **Document Format**: PNG
929 **Document Size**: 4.5 MB
930

931

932 **Additional File 4. Phylogeny of ArsB in RefSoil+ organisms.** Maximum likelihood tree
933 with 100 bootstrap replications of ArsB sequences predicted from RefSoil+ genomes.
934 Leaf tips show the name of the RefSoil+ organisms and background color indicates
935 phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the
936 tree.

937 **Document Name**: Additional_File_4.png
938 **Document Format**: PNG
939 **Document Size**: 9.6 MB
940

941

942 **Additional File 5. Phylogeny of ArsC (trx) in RefSoil+ organisms.** Maximum likelihood

943 tree with 100 bootstrap replications of ArsC (trx) sequences predicted from RefSoil+
944 genomes. Leaf tips show the name of the RefSoil+ organisms and background color
945 indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black circles
946 within the tree.

947 **Document Name**: Additional_File_5.png
948 **Document Format**: PNG
949 **Document Size**: 2 MB
950

951

952 **Additional File 6. Phylogeny of ArsC (grx) in RefSoil+ organisms.** Maximum
953 likelihood tree with 100 bootstrap replications of ArsC (grx) sequences predicted from
954 RefSoil+ genomes. Leaf tips show the name of the RefSoil+ organisms and background
955 color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black
956 circles within the tree.

957 **Document Name**: Additional_File_6.png
958 **Document Format**: PNG
959 **Document Size**: 4.9 MB
960

961

962 **Additional File 7. Phylogeny of ArsM in RefSoil+ organisms.** Maximum likelihood tree
963 with 100 bootstrap replications of ArsM sequences predicted from RefSoil+ genomes.
964 Leaf tips show the name of the RefSoil+ organisms and background color indicates
965 phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the
966 tree.

967 **Document Name**: Additional_File_7.eps
968 **Document Format**: EPS
969 **Document Size**: 6.2 MB
970

971

972 **Additional File 8**. **Histogram of arsenic related gene copy numbers in RefSoil+**
973 **organisms.** Total copy number is based on hits from both chromosomes and plasmids
974 from the same organism.

975 **Document Name**: Additional_File_8.eps
976 **Document Format**: EPS
977 **Document Size**: 25 KB

Additional File 9. Phylum-level community structure of soil metagenomes in this study.

Document Name: Additional_File_9.eps
Document Format: EPS
Document Size: 67 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.

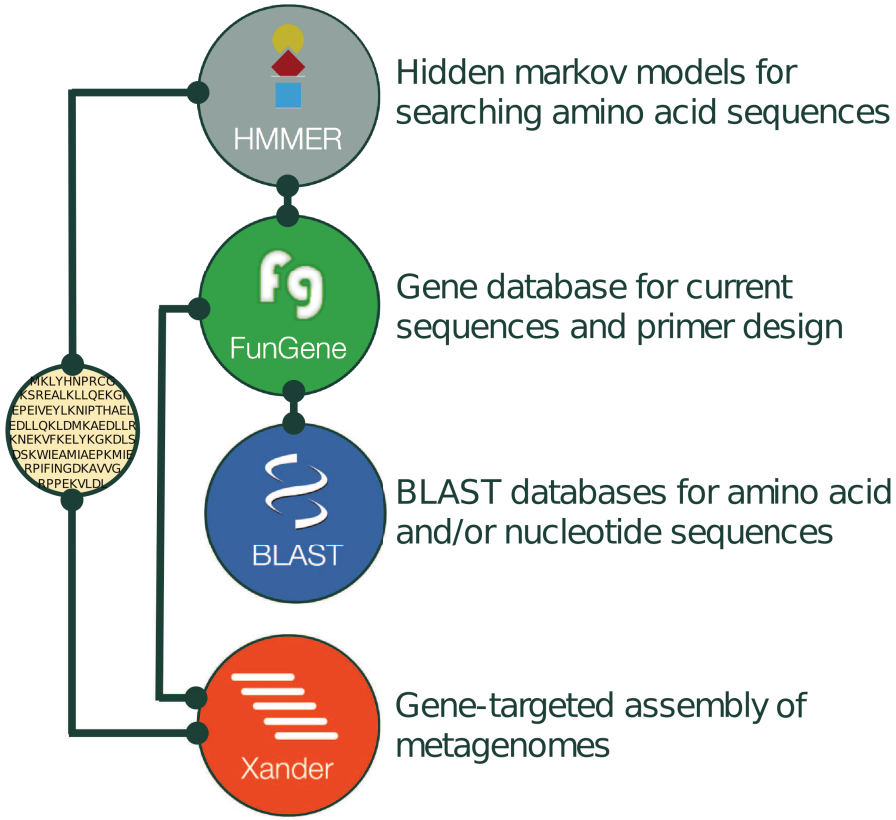
Document Name: Additional_File_10.docx
Document Format: Word
Document Size: 45 KB

Additional File 11. Summary of reference arsenic resistance and metabolism gene sequences from FunGene databases.

Document Name: Additional_File_11.docx
Document Format: Word
Document Size: 53 KB

Figure 1

A



B

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

Figure 2

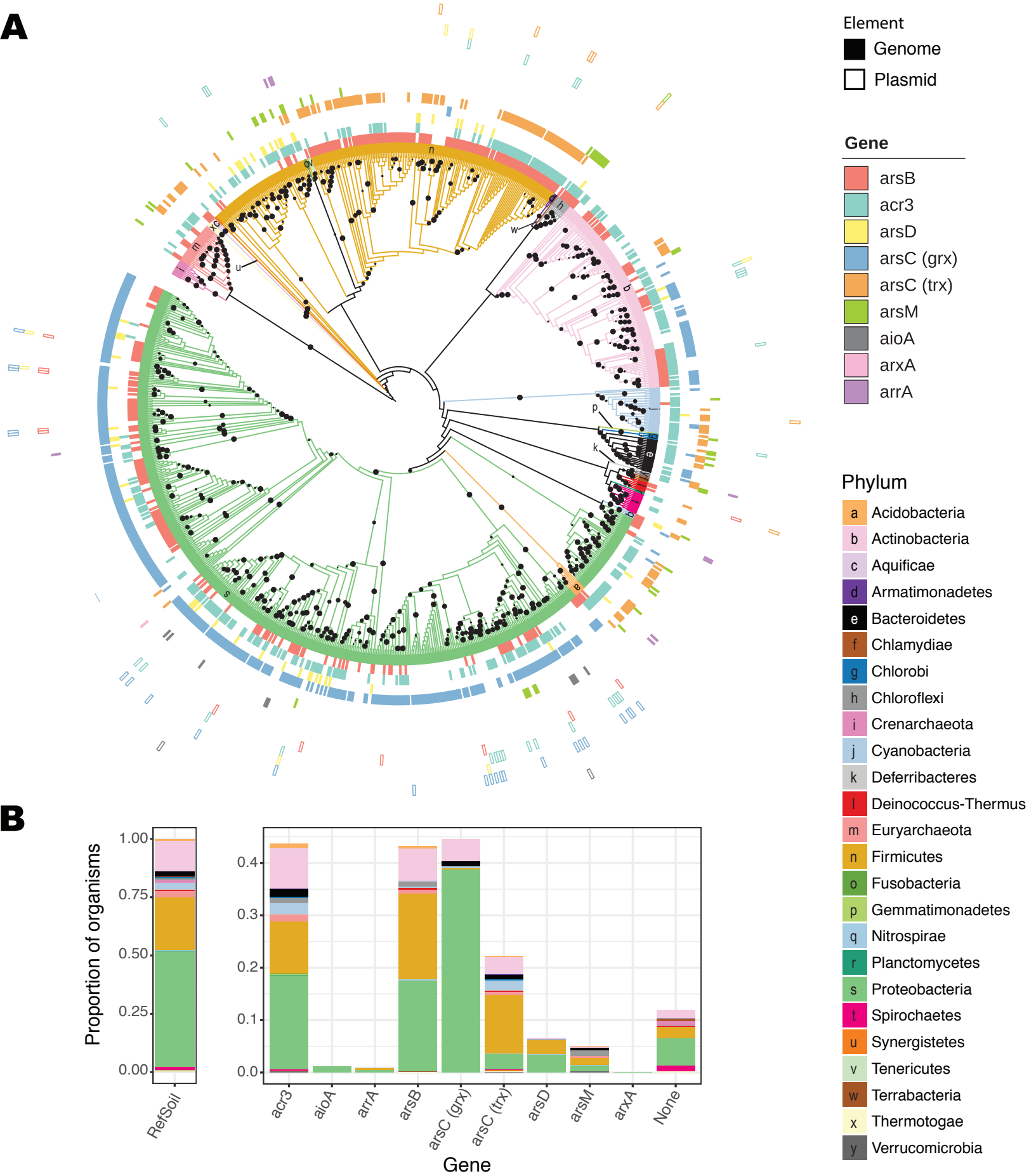

A
Element
Genome
Plasmid
Gene
arsB
acr3
arsD
arsC (grx)
arsC (trx)
arsM
aioA
arxA
arrA
Phylum
a Acidobacteria
b Actinobacteria
c Aquificae
d Armatimonadetes
e Bacteroidetes
f Chlamydiae
g Chlorobi
h Chloroflexi
i Crenarchaeota
j Cyanobacteria
k Deferribacteres
l Deinococcus-Thermus
m Euryarchaeota
n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

Figure 3

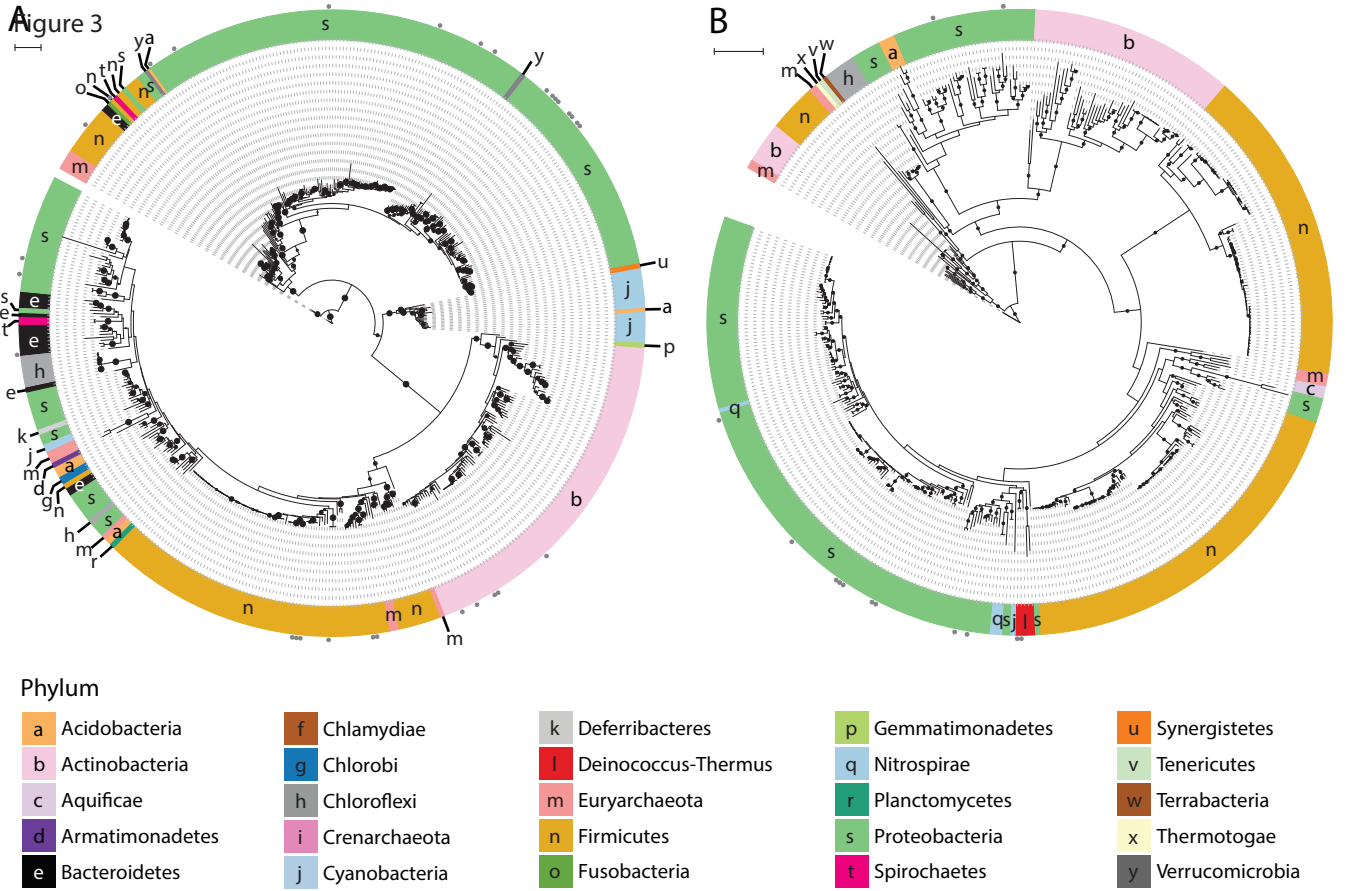

A
B
Phylum
a Acidobacteria
b Actinobacteria
c Aquificae
d Armatimonadetes
e Bacteroidetes
f Chlamydiae
g Chlorobi
h Chloroflexi
i Crenarchaeota
j Cyanobacteria
k Deferribacteres
l Deinococcus-Thermus
m Euryarchaeota
n Firmicutes
o Fusobacteria
p Gemmatimonadetes
q Nitrospirae
r Planctomycetes
s Proteobacteria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

Figure 4

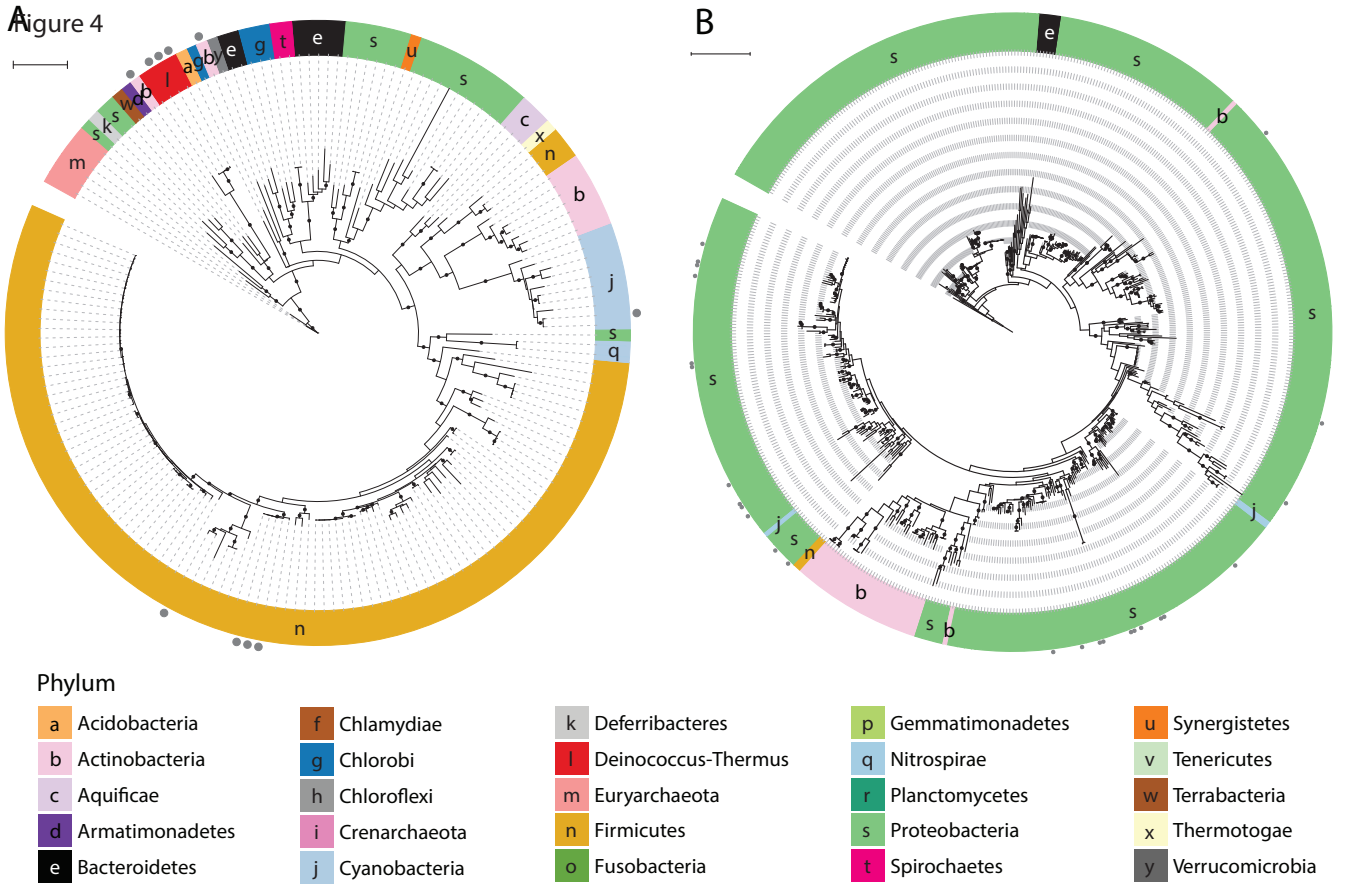

A
B
Phylum
a Acidobacteria
b Actinobacteria
c Aquificae
d Armatimonadetes
e Bacteroidetes
f Chlamydiae
g Chlorobi
h Chloroflexi
i Crenarchaeota
j Cyanobacteria
k Deferribacteres
l Deinococcus-Thermus
m Euryarchaeota
n Firmicutes
o Fusobacteria
p Gemmatimonadetes
q Nitrospirae
r Planctomycetes
s Proteobacteria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

Figure 5

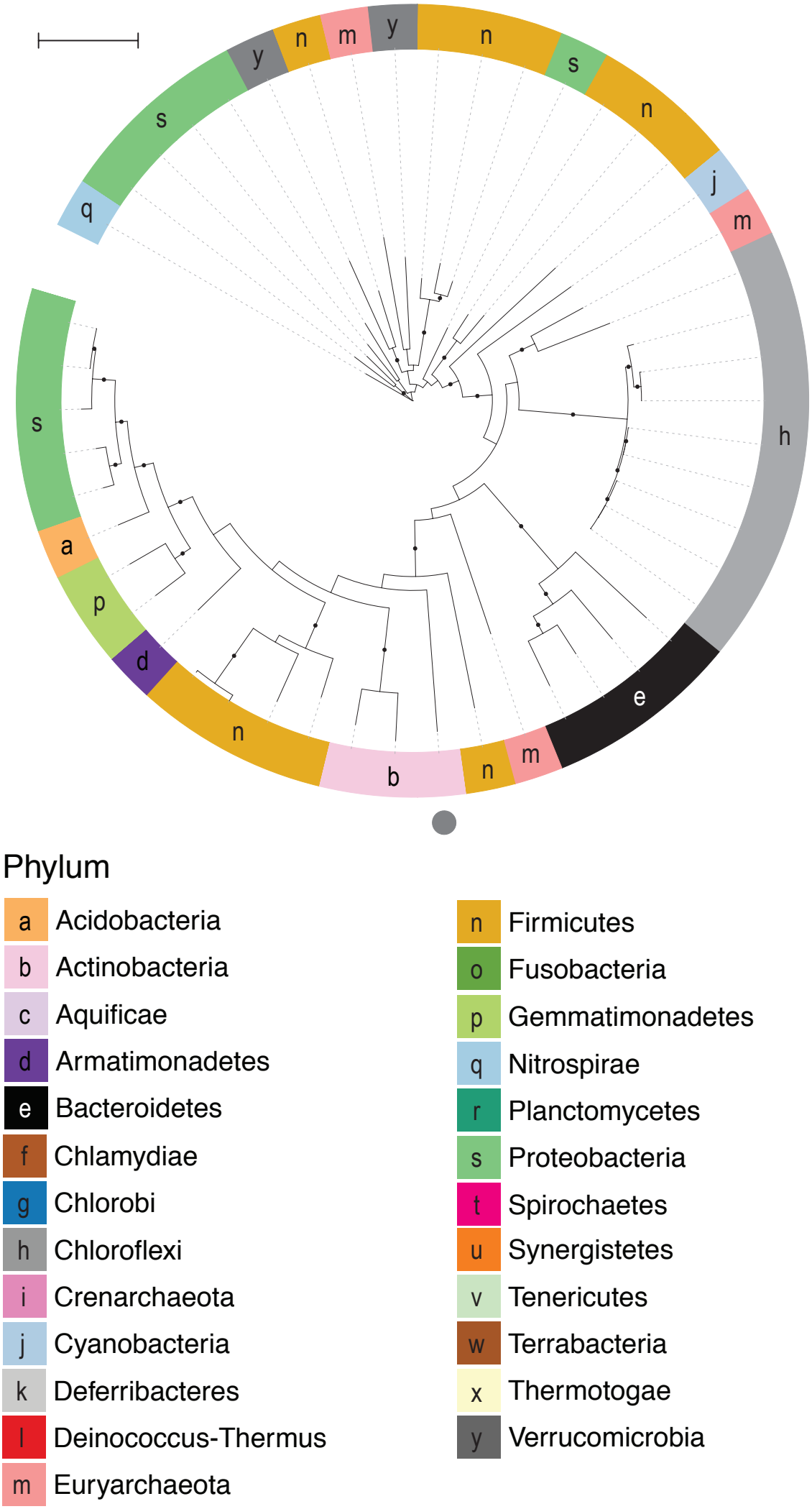

Phylum
a Acidobacteria
b Actinobacteria
c Aquificae
d Armatimonadetes
e Bacteroidetes
f Chlamydiae
g Chlorobi
h Chloroflexi
i Crenarchaeota
j Cyanobacteria
k Deferribacteres
l Deinococcus-Thermus
m Euryarchaeota
n Firmicutes
o Fusobacteria
p Gemmatimonadetes
q Nitrospirae
r Planctomycetes
s Proteobacteria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

Figure 6

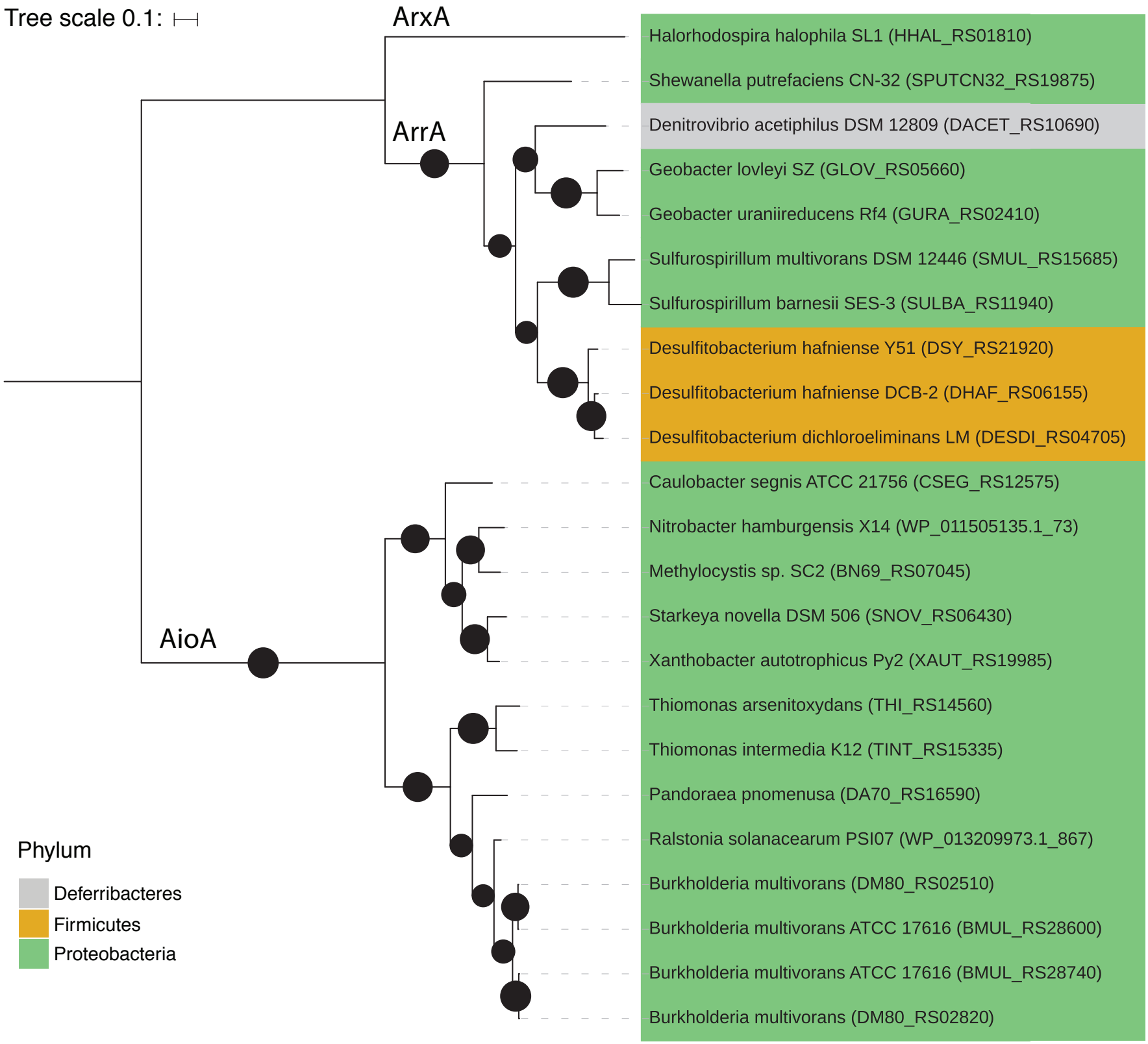

Tree scale 0.1:
ArxA
ArrA
AioA
Halorhodospira halophila SL1 (HHAL_RS01810)
Shewanella putrefaciens CN-32 (SPUTCN32_RS19875)
Denitrovibrio acetiphilus DSM 12809 (DACET_RS10690)
Geobacter lovleyi SZ (GLOV_RS05660)
Geobacter uraniireducens Rf4 (GURA_RS02410)
Sulfurospirillum multivorans DSM 12446 (SMUL_RS15685)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

Figure 7

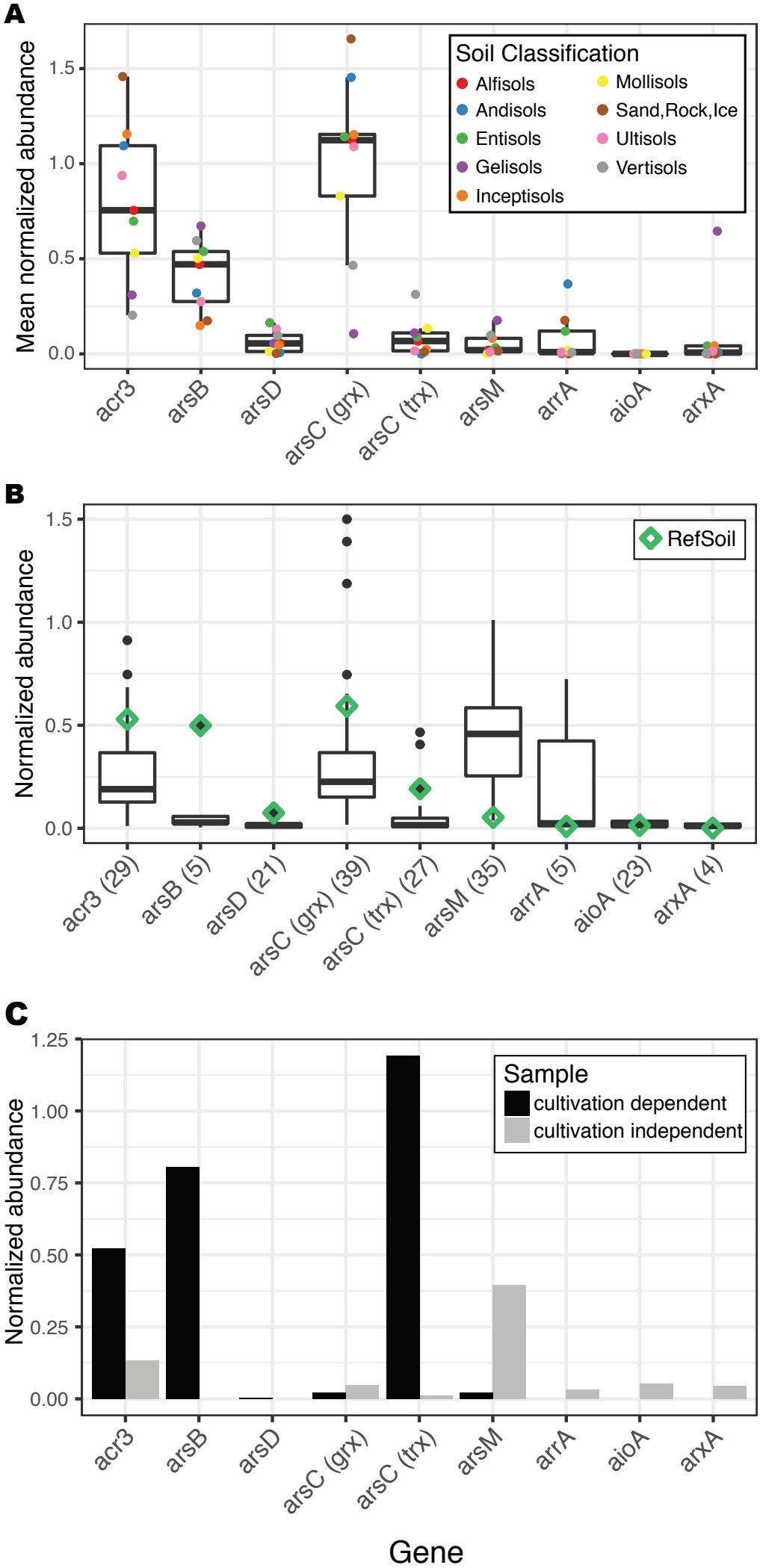
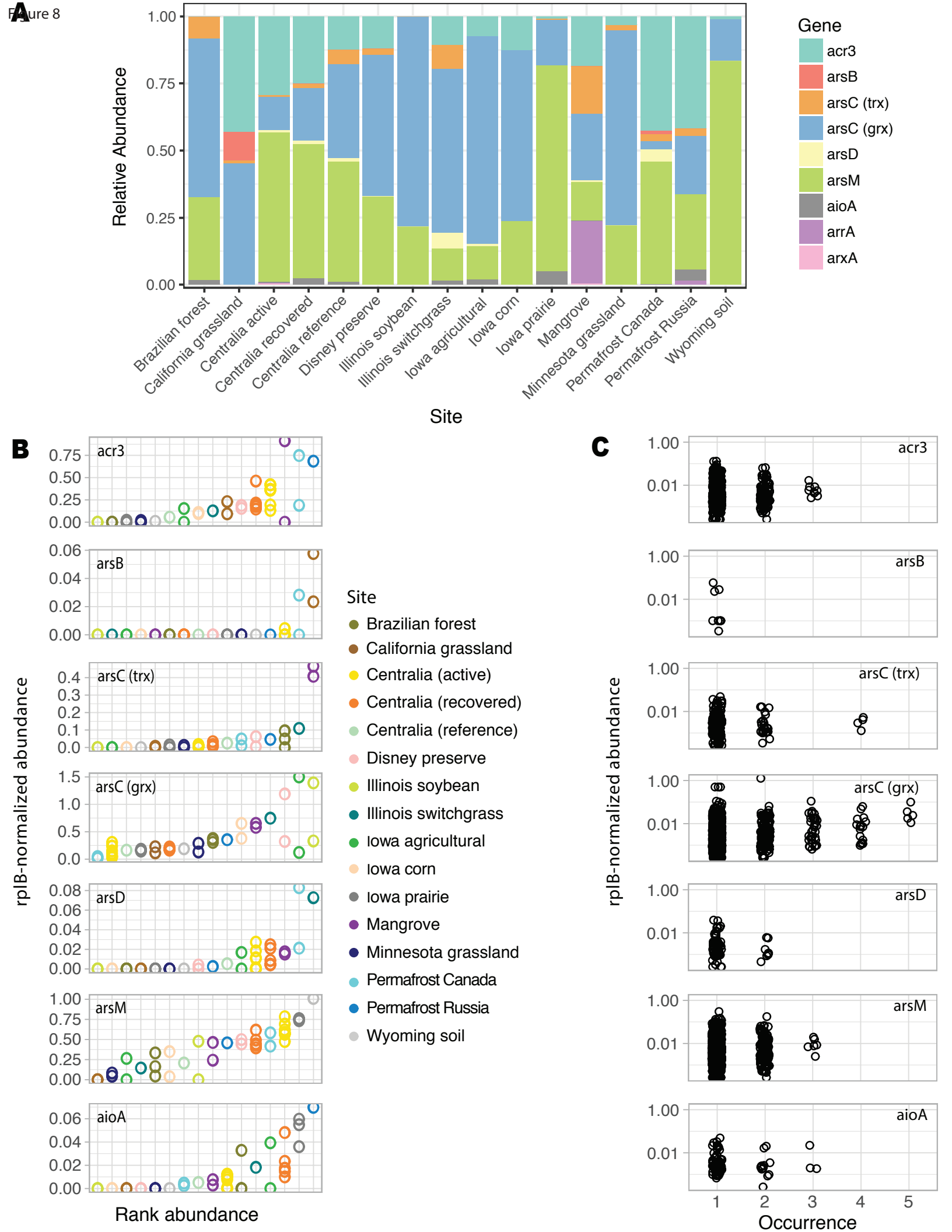

A
Mean normalized abundance
1.5
1.0
0.5
0.0
Soil Classification
Alfisols
Andisols
Entisols
Gelisols
Inceptisols
Mollisols
Sand,Rock,Ice
Ultisols
Vertisols
acr3
arsB
arsD
arsC (grx)
arsC (trx)
arsM
arrA
aioA
arxA
B
Normalized abundance
1.5
1.0
0.5
0.0
RefSoil
acr3 (29)
arsB (5)
arsD (21)
arsC (grx) (39)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

Figure 8


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