High arsenic levels increase activity rather than diversity or abundance of arsenic metabolism genes in paddy soils

Si-Yu Zhang (syzhang@des.ecnu.edu.cn)
East China Normal University https://orcid.org/0000-0001-6701-5790
Xiao Xiao
Nanchang Hangkong University
Song-Can Chen
Helmholtz-Centre for Environmental Research - UFZ: Helmholtz-Zentrum fur Umweltforschung UFZ
Yong-Guan Zhu
Chinese Academy of Sciences
Konstantinos T. Konstantinidis
Georgia Tech: Georgia Institute of Technology

Research Article

Keywords: Paddy soil, Arsenic contamination, Arsenic metabolism genes, Microbial Abundance and diversity, Transcriptional activity

Posted Date: August 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-142737/v2

Version of Record: A version of this preprint was published at Applied and Environmental Microbiology on September 28th, 2021. See the published version at https://doi.org/10.1128/AEM.01383-21.

Abstract

Arsenic (As) metabolism genes are generally present in soils but their diversity, relative abundance, and transcriptional activity in response to different As concentrations remain unclear, limiting our understanding of the microbial activities that control the fate of an important environmental pollutant. To address this issue, we applied metagenomics and metatranscriptomics to paddy soils showing a gradient of As concentrations to investigate As resistance genes (*ars*) including *arsR*, *acr3*, *arsB*, *arsC*, *arsM*, *arsI*, *arsP*, and *arsH* as well as energy-generating As respiratory oxidation (*aioA*) and reduction (*arrA*) genes. Somewhat unexpectedly, the relative DNA abundances and diversity of *ars*, *aioA*, and *arrA* genes were not significantly different between low and high (~10 vs ~100 mg kg^{-1}) As soils. By comparison to available metagenomes from other soils, geographic distance rather than As levels drove the different compositions of microbial communities. Arsenic significantly increased *ars* genes abundance only when its concentration was higher than 410 mg kg^{-1}. In contrast, between low and high As soils, metatranscriptomics revealed a significant increase in transcription of *ars* and *aioA* genes, which are induced by arsenite, the dominant As species in paddy soils, but not *arrA* genes, which are induced by arsenate. These patterns appeared to be community-wide as opposed to taxon-specific. Collectively, our findings advance understanding of how microbes respond to high As levels and the diversity of As metabolism genes in paddy soils and indicated that future studies of As metabolism in soil, or other environments, should include the function (transcriptome) level.

IMPORTANCE

Arsenic (As) is a toxic metalloid pervasively present in the environment. Microorganisms have evolved the capacity to metabolize As, and As metabolism genes are ubiquitously present in the environment even in the absence of high concentrations of As. However, these previous studies were carried out at the DNA level and thus, the activity of the As metabolism genes detected remains essentially speculative. Here, we show that the high As levels in paddy soils increased the transcriptional activity rather than the relative DNA abundance and diversity of As metabolism genes. These findings advance our understanding of how microbes respond to and cope with high As levels and have implications for better monitoring and managing an important toxic metalloid in agricultural soils and possibly other ecosystems.

Introduction

Arsenic (As) is a toxic metalloid, widely distributed in Earth's crust. Relatively high amounts of As are introduced into the environment from natural sources such as volcanic eruptions, weathering of rocks and geothermal activities as well as anthropogenic activities such as mining, pigment production and application of As-based pesticides in agriculture (1, 2). Because of its pervasive presence in the environment, microorganisms have evolved the capacity to metabolize As (3), and this capacity is hypothesized to be an ancient mechanism that emerged at least 2.72 billion years ago (4, 5).

Microbial mediated As biotransformation include As detoxification to mitigate toxicity and As respiration to generate energy. Several mechanisms have evolved to detoxify both inorganic and organic As. The inorganic arsenate [As(V)] compounds in cytoplasm can be reduced by a reductase (ArsC) (6), followed by arsenite [As(III)] efflux outside the cell, which is catalyzed by two evolutionarily unrelated As(III) efflux permeases, i.e. ArsB and Acr3 (7). Additionally, As(III) can be methylated by an As(III) *S*-adenosylmethionine (SAM) methyltransferase (ArsM) to the more toxic organic methylarsenite [MMAs(III)] (8), followed by the organoarsenical detoxification pathways, which have been recently shown to be present in various microbial genomes. The latter detoxification pathways include MMAs(III)'s oxidization to methylarsenate [MMAs(V)], which is catalyzed by the MMAs(III)-specific oxidase ArsH (9), and actively transported outside the cell by the MMAs(III) efflux permease ArsP (10), or MMAs(III)'s demethylation to the less toxic As(III) by the ArsI lyase, which cleaves the carbon-As bond in MMAs(III) (11). These As resistance genes are usually organized in an *ars* operon, which is controlled by one of three known ArsR transcriptional repressors (ArsR1, ArsR2 and ArsR3) regulated selectively by As(III) and one (ArsR4) by MMAs(III), depending on the exact version of the operon (12, 13). The As respiratory pathways include oxidation of As(III) either to detoxify it to less toxic As(V), or couple it to ATP production (14), and dissimilatory reduction that couples As(V) reduction to anaerobic heterotrophic growth (15). As(III) oxidation is catalyzed by As(III) oxidase (Aio) whose large subunit (AioA) has been well characterized in both bacteria and archaea (16). On the other hand, As(V) reduction is catalyzed by respiratory As(V) reductase (Arr), whose large catalytic subunit (ArrA) serves as a reliable marker for the process (17).

Due to the anaerobic conditions typically prevailing in paddy soils, As(III) is commonly the predominant form of As, which is more mobile and bioavailable to microbes for biotransformation (18-20). Indeed, microbial genes involved in As resistance such as *arsC* and *arsM*, and genes involved in As respiration such as *aioA* and *arrA* have been detected to be widely presented in the paddy soils with As contamination (21), and their activity to biotransform As has also been experimentally demonstrated (22-24). Recent studies of soils with low As levels, i.e. typically less than 15 mg kg^{-1} (or ppm), also showed relative high abundance and diversity of microbial As metabolism genes; most notably, *arsB*, *acr3*, *arsH* , *arsR, arsC*, *arsM*, *aioA* and *arrA* genes in paddy soils (25) and *arsC*, *arsM*, *aioA* and *arrA* genes in estuarine sediments across southeastern China (26). These results reveal high prevalence of As metabolism microbes in the environment even in absence of high concentration of As. However, these previous studies were carried out at the DNA level and thus, the expression of the As metabolism genes detected remains speculative. Further, these DNA-based results showing prevalent As metabolism microbes in both high and low As level environments are somewhat surprising because gene functions and species selected by a specific factor tend to be more abundant in ecosystems where the factor (selective pressure) is stronger as this has been shown for various soil factors such as pH, total organic carbon (TOC), total nitrogen (TN), phosphorus (TP) (27) and organic pollutants (28).

Herein, we applied both metagenomics and metatranscriptomics to investigate the resident (at the DNA level) and active (at the RNA level) As-associated microbial genes in paddy soils of different As

concentrations, aiming to address the following questions: (1) Do As metabolism genes in high As levels paddy soils show higher relative abundances and/or transcriptional activities than their counterparts in low As levels paddy soils to cope with the higher As level present? (2) Which are the key As metabolism genes and pathways coping with high As levels? (3) What are other factors than As levels driving microbial communities compositional and gene transcriptional differences in soils with As contamination?

Results

Soil physicochemical characteristics

A total of 18 samples, three biological replicates from the same field for each site, were collected from six paddy soils in Hunan, China with different As concentrations (Fig. S1 and Table S1). Arsenic (As) concentrations in these paddy soils varied significantly (ANOVA, $p < 0.01$) between high (67.3-104.0 mg kg^{-1}) and low (2.5-10.8 mg kg^{-1}) As levels. The high As concentrations in the paddy soils likely resulted from the historic mine activities occurring in the same area. The total concentrations of nitrogen (TN) in high As paddy soils were also significant higher (ANOVA, $p < 0.01$) than those with low As (1.64-1.77 g kg^{-1} vs 1.17-1.24 g kg^{-1}). No significant difference was found in the other soil physiochemical characteristics, including pH, Eh, total concentration of carbon (TC), phosphorus (TP), sulfur (TS), and concentrations of heavy metals including Pb, Cd, Cr, Sb and Fe. Even though the Eh in sample CZ_AsHig, and Pb and Cd concentrations in sample YCP_AsLow were substantially different from the other samples, these differences in soil properties did not reach statistical significance between high and low As paddy soils, presumably due to small sample size. Details of the soil properties were summarized in supplementary Table S2.

Broad microbial community diversity indices

Six shotgun metagenomes and six shotgun metatranscriptomes were obtained after the triplicate DNA and RNA samples from each site were pooled together, respectively. For metagenomes (Table S3), 12-15 Gbp (billion base pairs, 13 Gbp, on average) of shotgun metagenomic data and an average read length of ~115 bp were acquired after trimming for each pooled DNA sample. On average, 6.6 Gbp (ranged from 3.9 to 8.3 Gbp) of shotgun metatranscriptomic data and average read length of ~129 bp were obtained after trimming and removal of rRNA reads for each pooled RNA sample (Table S4). The coverage achieved by sequencing based on the read redundancy value estimated by Nonpareil was around 22-27%, except for sample SKS_AsHig, which had a lower alpha-diversity (measured by Nonpareil) than the others (*Nd* 22.6 vs 23.6-23.8; note that *Nd* is in log scale) and thus, higher coverage (42%). No significant difference ($p > 0.05$) was found in the alpha-diversity between high and low As paddy soils (average of 23.3 vs 23.7). The average genome size of microbial community was 5.7 Mbp and 5.6 Mbp in high and low As paddy soils, respectively. Considering the difference of average genome size of microbial

community between these paddy soils was no more than 2-fold, estimation of As metabolism genes abundance was essentially the same between the RPKM (the number of reads mapping on reference sequences/gene length in kb per one million reads analyzed) and RPKG (RPKM/average genome size) metrics.

Paddy soil community taxonomic composition

16S rRNA gene (16S) fragments extracted from the metagenomes were about 0.03-0.04% of the total reads (Table S3) and around 65-70% of them were classifiable at the phylum level. Bacteria were predominant in these paddy soils, accounting for 61-69%, while Archaea accounted for only 0.5-2.4%; the remaining reads were unclassified against the SILVA database (29). The dominant bacteria comprised of phyla Proteobacteria, Acidobacteria, Chloroflexi, Nitrospirae, Verrucomicrobia and Bacteroidetes, representing 50-59% of the total 16S-carrying reads. Archaea community was mostly represented by the Crenarchaeota phylum at 0.05-0.78% of the total, followed by Euryarchaeota at 0.36-0.77% (Fig. 1A). There were overall no significant relationships between the level of As and community composition, assessed by Bray-Curtis similarity in 16S rRNA gene-based OTUs (NMDS plot; Fig. S2). DESeq2 analysis (Fig. 1B) revealed statistically significant lower abundance of members of Euryarchaeota, i.e. *Methanomicrobia* and *Thermoplasmata* (0.15% vs 0.85% and 0.02% vs 0.07%, respectively; $p < 0.05$) in high vs low As paddy soils. In contrast, the abundance of Chloroflexi, mostly comprised of Ellin6529 (1.44% vs 0.52%), S085 (0.28% vs 0.09%), *Thermomicrobia* (0.06% vs 0.01%), and *Chlorobi*-OPB56 (0.16% vs 0.05%), GAL15 (0.21% vs 0.01%) and OP11-WCHB1-64 (0.07% vs 0.01%) were significantly ($p < 0.05$) higher in high vs low As paddy soils. We also attempted to recover metagenome-assembled genomes (MAGs; see details in supplementary Results and Discussions), which yielded only five MAGs of high quality (completeness >80% and contamination <5%) presumably due to the high complexity of the datasets. These MAGs were assignable to the class of *Deltaproteobacteria*, *Thermomicrobia*, *Clostridia*, *Acidobacteriaceae* and *Solibacteres*, which is consistent with the dominant microbes in these sites as revealed by the read-based findings reported above.

Paddy soil community functional composition

Protein-coding fragment sequences were predicted for about 90-92% and 44-59% of the trimmed reads in metagenomics and metatranscriptomics, respectively (Table S3 and S4). The high As microbial communities clustered separately than their low As counterparts based on functional gene content (summarized as annotation counts from SEED subsystems) albeit they were also quite different from each other and thus, did not cluster tightly together (Fig. 2A). It was not unexpected to observe clustering at the functional but not taxonomic composition of the sampled communities considering that functional gene content has been shown to be more strongly and faster influenced by environmental conditions than phylogenetic beta-diversity (30). DESeq2 analysis revealed no statistically significant changes in the

relative abundance of genes predicted from metagenomics of different As levels. In contrast, genes associated with As resistance were significantly (adjusted $p < 0.05$) more expressed based on metatranscriptomes by a $\log_2$ fold change of 1-2.3 in high vs low As paddy soils (Fig. 2B and 2C). Other gene functions, including phage tail fiber, phage photosynthesis, phage capsid, phage cyanophage and CBSS-316279 were also found to be differentially expressed between high and low As (Fig. 2B and 2C). This could be related to “hitchhiking” effects (31, 32), functions not assessed here, and/or experimental noise, and were not examined further here.

Relative DNA abundance and expression of As resistance and respiratory genes

Overall, the relative abundance of *ars* genes, i.e., the cumulative abundance of *arsR* (*arsR1*, *arsR2*, *arsR3*, *arsR4*), *acr3*, *arsB*, *arsC* (*arsC1*, *arsC2* and *acr2*), and *arsM*, *arsI*, *arsP* and *arsH* genes, were comparable between high and low As paddy soils as revealed by metagenomes ($1.2 \times 10^{-2} \pm 9.7 \times 10^{-3}$ vs $6.1 \times 10^{-3} \pm 5.3 \times 10^{-4}$ RPKM; Fig. 3A). An outlier to this result was sample SKS_AsHig that had three times higher relative abundance of *ars* genes compared to other samples (Fig. S3). To robustly assess transcriptome activity while accounting for DNA abundance, the transcriptional level of *ars* genes were estimated by normalizing metatranscriptomic counts by the relative abundance of same *ars* genes identified in metagenomes. More than 80% of the As metabolism genes detected in metatranscriptomes were identified in the metagenomes, representing the majority of the As metabolism genes in the samples. The normalized transcriptional expression levels of *ars* genes were significantly higher ($p < 0.05$) in high vs low As paddy soils (2074 ± 701 vs 710 ± 337 RPKM; Fig. 3A). Overall, the 13 *ars* genes responsible for As transcriptional repressor (*arsR*), As(III) efflux (*acr3* and *arsB*), As(V) reduction (*arsC* including *arsC1*, *arsC2* and *acr2*), As(III) methylation (*arsM*), MMAs(III) demethylation (*arsI*), MMAs(III) oxidation (*arsH*) and MMAs(III) efflux (*arsP*) all consistently showed higher transcriptional activities in high vs low As paddy soils (Fig. 3B). However, for sample HP_AsLow, the *arsH*, *arsR* and *arsP* showed a slightly higher transcriptional activity than one of the samples from high As levels sites (CZ_AsHig or SKS_AsHig), which could be attributed to high soil heterogeneity or biases during library preparation and/or sequencing.

For *aioA* genes (Fig. 3A), which encode respiratory As(III) oxidase in the periplasm, a higher relative DNA abundance was revealed, although not statistically significant ($p = 0.114$), in high vs low As paddy soils ($1.0 \times 10^{-4} \pm 6.2 \times 10^{-5}$ vs $2.9 \times 10^{-5} \pm 3.2 \times 10^{-7}$ PRKG, respectively). The transcriptional activity of *aioA* genes was strongly significantly ($p < 0.05$) increased in high vs low As paddy soils (117 ± 39 vs 34 ± 31 PRKM, respectively). Further, there was no significantly higher abundance ($p = 0.598$) or transcriptional activity of *arrA* genes ($p = 0.33$), which is responsible for respiratory As(V) reduction, between high and low As paddy soils (Fig. 3A). The relative abundance of *arrA* genes in metagenomes were comparable between high and low As paddy soils ($1.0 \times 10^{-4} \pm 5.8 \times 10^{-5}$ vs $7.4 \times 10^{-5} \pm 5.9 \times 10^{-5}$ RPKG, respectively), while their transcriptional activity was slightly higher, but not significant ($p = 0.33$), in paddy soils with high vs low As levels (39 ± 27 vs 19 ± 16 RPKM, respectively).

Responses to high As level in paddy soils are community-wide rather than taxon-specific

To evaluate which clades expressed higher As genes in high As soils, and thus, assess whether the transcriptomic shifts of As metabolism gene described above were due to systematic community-wide response as opposed to a few taxa, the Ars, AioA and ArrA-carrying reads in metagenomes (DNA reads) and metatranscriptomes (RNA reads) were placed on the their representative phylogenetic clade (typically >95% nucleotide identity within- vs <95% identity between-clades) and the ratio of RNA/DNA reads assigned at the same reference clade (or gene-based OTU) was compared between samples (Fig. 4 and Fig. S4-S12). Overall, in high As soils, more than 60% (ranging between 63% and 98%; Table S6) of the total clades recruited reads from metatranscriptomes for each As metabolism gene, except for *arsB*, suggesting that the high As levels mostly induced a community-wide response. Moreover, around 54-78% of the total clades for Ars, AioA and ArrA proteins showed a higher ratio of RNA/DNA reads in high than low As paddy soils (Table S6), revealing that the As metabolism genes were expressed more strongly in high than low As samples by many distinct clades (taxa). For *arsB*, due to relative lower abundance of this gene compared to the others (e.g. 75-375 times lower compared to *acr3* which is also responsible for As(III) efflux), only 17% of the total clades recruited *arsB*-carrying reads from high As paddy soils and no clade recruited reads from the low As paddy soils.

Environmental factors affecting the composition of microbial communities in metagenomes

To investigate the differences between microbial communities in high vs low As soils, five previously characterized samples with high As (total As concentration ranged from 34 to 821 mg kg^{-1}) affected by a mine field in Hunan, China (33) and eight samples with low As (total As concentration ranged from 2 to 16 mg kg^{-1}) in Hunan, Zhejiang, Jiangxi and Guangdong, China (25) were included in comparisons with the samples determined by our study. Because these metagenomes had more than two-fold variation in sequencing effort applied, we subsampled the metagenomes to the same level (2 Gbp) in order to avoid false positive signal in detecting features as differentially abundant due to differences in coverage alone (34). PCoA-based analysis of Mash distances of whole metagenomes showed that samples were mostly separated based on their geographic distances, e.g., the most distinct metagenomes were those that originated form the most distant sites (Fig. 5A). The Adonis dissimilarity test corroborated that geographic distribution patterns imposed significant ($p < 0.01$) changes in microbial community compositions. Notably, As level did not cause any significant changes ($p > 0.05$) of the whole microbial communities based on metagenomic (DNA) abundances. For the As-associated microbial communities (Fig. S13), PCoA-based analysis of Mash distances of all As gene-carrying reads recovered from metagenomes revealed a weaker effect of geographic distance in shaping diversity pattens compared to the whole microbial communities. In contrast, As level showed a relatively higher effect, especially on the As-associated microbial communities in the mine field, which is characterized by higher As level, ranging

from 34.1 to 821.2 mg kg $^{-1}$. However, when the analysis was restricted to the paddy soils characterized by our study, no clear significant effect of high vs low As level ((~10 vs ~100 mg kg^{-1}) was observed, which was consistent with the NMDS plot of Bray-Curtis metric based on the counts of As genes annotated from metagenomes (Fig. S14).

Based on the samples determined by our study, the concentrations of total phosphorus (TP) and iron (Fe), in contrast to As, were significant factors in driving the variations observed between the microbial communities (Fig. 5B), explaining about 30% of the differences among the sampled total microbial communities. A significant ($p < 0.01$) interaction was also found between these two factors. Interestingly, the TP concentration also significantly correlated with the relative abundance of different *ars* genes in the metagenomic datasets ($R^2 = 0.83$, $p < 0.05$). The As concentration was found to significantly correlate with the variance in *ars* gene transcriptional activities assessed by Bray-Curtis distance between paddy soils (Adonis test, $R^2 = 0.69$, $p < 0.05$) and the transcriptional activities of *aioA* genes (Pearson's correlation = 0.85, $p < 0.05$). In addition to As, the concentration of total nitrogen (TN) ($R^2 = 0.73$, $p < 0.01$) and chromium ($R^2 = 0.72$, $p < 0.05$) also significantly correlated with differences in the transcriptional activities of *ars* genes, and the concentration of antimony (Sb) (Pearson's correlation = 0.83, $p < 0.05$) significantly correlated with the transcriptional activities of *aioA* genes (Table S7). No significant interaction was found between these environmental factors as revealed by PERMANOVA analysis. For the transcriptional activities of *arrA* genes, no environmental factor was found to affect these activities among the different paddy soils examined.

Discussion

Arsenic metabolism genes were pervasively present in the paddy soils either with high (67.3-105 mg kg^{-1}) or low (2.5-10.8 mg kg^{-1}) As levels, and their relative DNA abundance was mostly comparable between these soils (Fig. 3). Microbial capacity of As metabolism has been shown to be an ancient trait developed in response to the pervasive presence of As in early Earth (3). Therefore, it is likely that the presence and relative abundances of As metabolism genes in the paddy soils are possibly a legacy effect of an early life characteristic, at least to some extent (4). Further, the As resistant mechanisms, including the most well studied detoxification As(V) reduction system (ArsC) followed by the efflux of reduced As(III) (ACR3 and ArsB) (1), and the recently identified organoarsenical detoxification systems (ArsM, ArsI, ArsH and ArsP) (10, 35) have been detected in various microbes (36). The respiratory reduction of As(V) or oxidation of As(III), which is thought to have been developed in the Archaean era (16, 37), have also been described in a broad diversity of microbes (38). Consistently, the prevalent As metabolism genes have been previously reported in various wetlands with low As levels, such as paddy soils and estuarine sediments (< 15 mg kg^{-1}) (21, 25, 26), corroborating the wide prevalence As metabolism genes in our paddy soils.

Notably, the As concentration in high As paddy soils, up to ~100 mg kg $^{-1}$ in the paddy soils studied here, did not significantly correlated to a higher relative abundance of As metabolism genes in the resident As-

associated microbial community. Previous study of As metabolism genes in mining impacted fields showed a significant ($p < 0.05$) linear correlation of *arsC* and *aioA* gene abundance with the increasing of As concentration from 34 to 821 mg kg^{-1} (33). We studied the correlation of *ars* genes abundances with As concentrations using our own and the previously determined samples by Luo and colleagues and found, consistent with the previous study, a significant correlation ($p < 0.05$) between As metabolism gene abundance and high As concentration (especially in the 410-812 mg kg^{-1} range; Fig. S15). These findings indicated that the 12 times fold higher As concentration in high (average of 92 mg kg^{-1}) vs low (average of 8 mg kg^{-1}) paddy soils of our study was not strong enough to significantly differentiate As genes abundances, and that higher difference in As level and/or high absolute As concentrations (e.g., > 410 mg kg^{-1}) is probably necessary in order to elicit more clear gene content and abundance differences. The possibilities that some of the As we measured in our high As soils is not bioavailable (and thus, not selecting for more As-metabolism genes), especially considering that the water soluble or exchangeable (bioavailable) As only accounts for about 0.4-24% of the total As in paddy soils (39). Therefore, to further corroborate these findings, however, the selection pressure exercised by different As levels, As species should also be measured in future studies.

Nevertheless, the As presented in the soil has apparently induced the ArsR-mediated up-regulation of the *ars* operon and Aio and Arr, and therefore resulted in the increased activities of As metabolic pathways, even though the difference in As concentration between high vs. low As paddy soils was probably not great enough to induce difference in DNA abundance due to most bacteria being intrinsically tolerant to such As concentrations (40). The higher abundance of *ars* (*arsR*, *acr3*, *arsB*, *arsC*, *arsM*, *arsI*, *arsP* and *arsH*) and *aioA* genes at the expression (but not DNA) level presumably reflected a truly stronger response of As genes to increased levels of As in the paddy soils at the RNA level than DNA level. Among the As metabolism proteins, both the ArsR and AioA proteins are induced by As(III), which is the dominant As species in the paddy soils during the flooded period (18, 19), and most likely contributed to the significantly increased transcriptional activities of *ars* and *aioA* genes in soils with higher As levels (Fig. 3). But the overall transcriptional activity of *arrA* gene, which is responsible for dissimilarity As(V) reduction (38) was not significantly different between high vs low As paddy soils (Fig. 3). Because ArrA is induced by As(V), it is possible that the dominant As species, i.e. As(III) in paddy soils (18, 19) did not significantly upregulate its transcriptional activity directly. Moreover, our study showed the increased transcriptional activity of As metabolism genes was due to several distinct clades (community-wide) in the high As paddy soils (Fig. 4 and Supplementary Fig. S4-S12), and this result again corroborated the pervasive presence of As metabolism mechanisms in microbes of the paddy ecosystem.

It's also noticeable that the relative abundance of *aioA* genes (DNA level) were about 3.5 times higher in high than low As paddy soils, although not statistically significant ($p = 0.114$). In contrast, the relative abundance of *ars* genes at the DNA level was comparable between high and low As levels (Fig. 3A). Considering that the As respiratory As(III) oxidation catalyzed by Aio in microbes can couple As(III) oxidation to ATP production, while the Ars is only responsible for As resistance (14, 36), our finding is consistent with the expectation that energy-production genes (e.g., *aioA*) will be at higher demand at both

the DNA and transcriptional (RNA) levels with increased availability of the substrate, while for the resistance genes (e.g., *ars*) transcriptional response might be adequate with increased substrate and the substrate might not be similarly toxic to all species in order to select for the corresponding genes.

The As-associated microbial populations in high As soils mostly clustered together and separate from their low As counterparts based on the similarities among the expression levels of the *ars*, *aioA* and *arrA* genes (Fig. S14), except for sample SKS_AsHig, which could be attributed to high soil heterogeneity or the non-significant effect of As levels on *arrA* genes. In contrast, the As-associated populations did not show any obvious clustering in terms of high As vs low As soils based on metagenomic (DNA) abundances (Fig. S14). These results are overall consistent with the relative abundance and expression of As metabolism genes, indicating that high As select more for transcriptome response as opposed to gene abundance at DNA level.

In terms of whole microbial communities, geographic distances, rather than As levels, were found to shape microbial community differences among the paddy soil sites studied here (Fig. 5). The geographic patterns of soil microbial communities have been widely studied previously and are mostly attributed to varied physicochemical and carbon sources in soil environments (41-43). It is possible that the As levels in these paddy soils did not possess a strong-enough selective pressure on the whole microbial communities. In contrast to As, the concentrations of TP (0.31-0.46 g kg^{-1}) and Fe (14.8-27.0 mg kg^{-1}), which were about 1.5-1.8 times different between high and low paddy soils, were significant in determining microbial community diversity among our samples (Fig. 5). Previous studies have shown that available organic soil P has an important influence on microbial community composition, shifting both the composition and function of soils microbes (44, 45). Especially for rhizosphere-associated microbes such as the dominant Proteobacteria in paddy soils (46) their community structure can be altered by soil P availability (47, 48). The Fe in paddy soils is likely to impact the microbially mediated redox processes (49), and its presence in the wetland environments has been demonstrated to correlate with microbial biogeographic patterns. It is also important to mention that the effect of Fe and TP on As microbes and genes may be interlinked due to the significant ($p < 0.01$) interaction of P and Fe since P is often the limiting nutrient in many terrestrial ecosystems due to strong sorption on Fe(III) hydroxides (50). A significant positive correlation between P and Fe in porewaters as well as soil matrix has been previously observed in flooded soils with rice growth (51). Moreover, their interaction could also affect As behavior in paddy soils (52) because iron oxyhydroxides could sequester As(V) (53) and reduce As bioavailability to microbes.

Overall, our study revealed that the high As level of ~100 kg mg^{-1} increased As metabolism genes activities rather than their abundance or diversity in paddy soils. These findings advance our understanding of how microbes respond to high As levels and the diversity of As metabolism genes in paddy soils. Even in cases where there may not be significant difference in the As-associated microbial communities at the DNA level, increased activity of the As-metabolic genes would indicate significant functional contributions to As biotransformation in the paddy ecosystem, which would be relevant for managing these ecosystems. Therefore, the potential functional contributions of the ubiquitous As

metabolism genes in the paddy ecosystem, and likely other environments, should be examined not only at the DNA level but also at the transcript and possibly protein levels. The As-associated microbial populations in soils with As could have significant contributions to the biogeochemical cycling of As and the MAGs reported here could facilitate future studies in this area. The work presented here could also serve as a guide for the number of samples to obtain, amount of sequencing to apply, and what bioinformatics analyses to perform for studying the dynamics of the As or other elements metabolism pathways.

Materials And Methods

Sites description and sample collection

Soil samples were collected from six distinct paddy fields in Hunan province, China, including three sites with high As levels (67.3-104.0 mg kg^{-1}) in Cili (CL_AsHig), Chenzhou (CZ_AsHig) and Shuikoushan (SKS_AsHig), and three sites with low As levels (2.5-10.8 mg kg^{-1}) in Hanpu (HP_AsLow), Lianhua (LH_AsLow) and Yuchangping (YCP_AsLow). Eighteen soil samples, three replicates from the same field for each site, were collected from the soil surface (0-20 cm) in July 2016 during the rice-growing period (flooded). Paddy soil samples were placed in sterile plastic bags and transported to the laboratory on ice for DNA and RNA extraction and soil physiochemistry analysis. Soil properties, including pH, Eh, total concentration of carbon (TC), nitrogen (TN), phosphorus (TP), sulfur (TS), arsenic (As), lead (Pb), cadmium (Cd), chromium (Cr), antimony (Sb) and iron (Fe) were determined on a composite sample of the three replicates following standard methods (21). Details of the underlying methods are provided in the supplementary.

Soil DNA and RNA isolation, library preparation and sequencing

Soil DNA and total RNA was extracted from 0.5 g and 2 g of soil using the FastDNA SPIN kit (MP Biomedicals) and E.Z.N.A Soil RNA Midi kit (Omega, Bio-Tek Inc., USA), respectively, according to the manufacturer's protocol. For each site, DNA or total RNA extracted from the triplicates were pooled together to obtain enough high-quality material for each site and to overcome high sample-to-sample heterogeneity, which is characteristic of soil microbial communities. The quality of DNA and total RNA was assessed by NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Inc., Wilmington, DE, USA) and the ratio of absorbance at 260 nm and 280 nm was around 1.8-2.0, indicating good quality for downstream analysis. Total RNA of the paddy soils was subjected to an rRNA removal procedure using the Ribo-Zero rRNA removal bacteria kit (Illumina Inc., San Diego, CA, USA) following the manufacturer's protocol.

Paired-end DNA and cDNA libraries (2 × 150 bp) were constructed using TruSeq^{TM} DNA Sample prep kit (Illumina Inc., San Diego, CA, USA) and TruSeq^{TM} RNA Sample prep kit (Illumina Inc., San Diego, CA, USA), respectively, following the manufacturer's protocol. The six DNA and cDNA libraries were sequenced

respectively on an Illumina HiSeq 3000 platform at Majorbio Bio-pharm Technology Company in Shanghai, China with HiSeq 3000/4000 PE Cluster kit and HiSeq 3000/4000 SBS kit (Illumina Inc., San Diego, CA, USA) according to the manufacturer's protocol.

Taxonomic classification, functional annotation, and population genome binning of soil metagenomes and metatranscriptomes

The raw paired-end reads of metagenomic and metatranscriptomics datasets were trimmed using SolexaQA (54), removing reads with Phred quality score (Q < 20) and minimum fragment read length of 50 bp after trimming for downstream analyses. Parallel-META version 2.4 (55) was used with default settings to recover 16S rRNA gene fragments from metagenomes and to identify and remove residual rRNA sequences after rRNA subtraction from metatranscriptomes. 16S rRNA gene fragments from metagenomes were then processed for Operational Taxonomic Unit (OTU) picking using close OTU picking, defined at the 97% nucleotide sequence identity threshold, and taxonomic identification with the SILVA database (29) as implemented in QIIME 1.9.1 (56) and classified at the phylum and class levels. The relative abundance of each taxon at each level was normalized by the total number of 16S rRNA gene sequence fragments recovered from the corresponding metagenome.

Protein-coding sequences from short-read metagenomes and metatranscriptomes were predicted using FragGeneScan (57) with default settings and searched against Swiss-Prot database (UniProt, downloaded in April 2019) using BLASTp (BLAST + version 2.2.28) (58). Only the best match with amino acid identity $\geq$ 50% and query sequence length coverage $\geq$ 70% by the alignment was retained. Protein-coding sequences were also functionally annotated, using the same threshold, against the SEED database using the subsystem categories (59). Differentially abundant taxa based on 16S rRNA gene sequence fragments and SEED subsystems of metagenomics and metatranscriptomics between high and low As levels were determined with the DESeq2 package using the negative binomial model, adjusted for false discovery rate (60). Heatmap in pheatmap R package (R version 3.5.1) (61) was used to visualize the significantly differentially abundant taxonomic levels and SEED subsystems (adjusted p value < 0.05). Metagenome-assembled genomes (MAGs) were recovered using both MaxBin Version 2.1.1 (62) and MetaBAT Version 2.12.1 (63) as described in more detail in supplementary.

Assessment of the abundance and transcriptional activity of arsenic metabolism genes

Building an As metabolism protein reference database

To identify and quantify As metabolism proteins containing reads, in-house databases of eight As resistance proteins [Ars i.e. ArsR (ArsR1, ArsR2, ArsR3 and ArsR4), ACR3, ArsB, ArsC (ArsC1, ArsC2 and ACR2), ArsM, ArsI, ArsP, ArsH] and As respiratory oxidation (AioA) and reduction (ArrA) proteins were constructed and manually curated as described before (5). Briefly, verified arsenic metabolism protein

sequences were downloaded from UniProt or NCBI's NR database to build HMM profiles, and searched against 786 representative species which are sampled from an original dataset used for reconstruction of tree of life (species tree) (64), to identify homologs of As metabolism proteins. The resulting matching sequences were manually inspected for the presence of conserved domains based on the Pfam models (release 28.0) (65) and forming deep-branching clades in the phylogenetic tree of the (manually) verified reference sequences to further determine if the sequence in question was truly As metabolism protein. The tree was built using maximum likelihood as implemented in RAxML v8.4.1 with the PROTGAMMAAUTO model (66). Sequences forming long-branches and/or lacking one or more of the conserved domains were removed from further analysis.

To build a more comprehensive As metabolism proteins database that better captures the diversity of As metabolism proteins, proteins on the contigs assembled using IDBA with default settings (67) from metagenomes and metatranscriptomes (for more details, see supplementary) were also searched against the curated As metabolism proteins database using BLASTp (BLAST + version 2.2.28) with a cut-off for a match of amino acid identity $\geq$ 40% and reference length coverage $\geq$ 70% by the alignment. A total of 1040 Ars proteins, including 87 ACR3, 281 ArsC, 1 ArsH, 69 ArsI, 173 ArsM, 21 ArsP and 408 ArsR were identified using this approach. However, for the AioA and ArrA proteins that are longer than the Ars proteins (800-900 vs 200-400 amino acids), no match was identified, and it was likely attributable to the limiting assembly, consistent with the relatively low sequencing effort. Therefore, in order to obtain a comprehensive database for AioA and ArrA proteins, the AioA and ArrA references in curated As metabolism database were searched against UniRef90 (TrEMBL, downloaded in October 2018) database for additional homologs. A more stringent cut-off compared to the Ars proteins mentioned above, i.e. identity $\geq$ 40% and reference length coverage $\geq$ 90% by the alignment, was used to filter the AioA and ArrA proteins in UniRef90 database. Identified Ars protein sequences from metagenomes and metatranscriptomes, and AioA and ArrA proteins sequences from UniRef90 database were searched against Pfam database (release 32.0) (65) in Jun 2019 and inspected for the presence of known conserved domains. The resulting matching sequences were aligned using MAFFT version 7.407 (68) and visually inspected for forming deep-branching clades in a maximum likelihood phylogenetic tree as described above to determine if the sequences were truly As metabolism proteins. These phylogenetic trees were also used for placement of As metabolism proteins encoding reads as described below.

Estimation of As metabolism genes abundance and transcriptional activity

Short sequences from metagenomes and metatranscriptomes were searched against the comprehensive Ars, AioA and ArrA references databases using BLASTx search (BLAST + version 2.2.28) and filtered for best matches with amino acid identity $\geq$ 90%, alignment length $\geq$ 25 amino acid and e-value $\leq$ 1E-5. MicrobeCensus was used to assess the average genome size of each sample community (69) and normalize the total base-pairs assign to each As gene by gene length, i.e. RPKM = (the number of reads mapping on reference sequences/gene length in kb per one million reads analyzed), and genome equivalent, i.e., RPKM/average genome size, providing an estimate RPKG [(reads mapped to gene)/(gene

length in kb)/(genome equivalents)] of how many cells from those sampled encode the gene of interest. Relative abundance of As metabolism genes in metatranscriptomics was calculated by the statistic RPKM (reads per kb per millions of reads), i.e. RPKM = (the number of reads mapping on reference sequences)/(gene length in kb)/(the total number of reads after removal of rRNA reads). The As metabolism genes transcriptional activity was evaluated by the relative abundance of As metabolism genes in metatranscriptomes (RPKM) normalized by the relative abundance of the same gene variant or OTU (at > 90% amino acid level) identified in metagenomes (RPKM; we used RPKM to normalized the relative abundance of the same gene identified in metagenomes here instead of RPKG in order to keep the normalization criteria consistent between metagenomes and metatranscriptomes in calculating the gene's transcriptional activity), and visualized in heatmap using the pheatmap package in R 3.5.1 (61). One-way ANOVA analysis, performed with the vegan package (70) in R 3.5.1, was used to compare the difference of As metabolism genes abundances and transcriptional activities in high vs low As levels paddy soils.

Phylogenetic placement of As metabolism gene carrying reads

Arsenic metabolism protein-containing reads identified by BLASTx search were firstly translated to amino-acid sequences by FragGeneScan (57) with default settings. The amino acid fragments were then added to the As metabolism protein references alignments using MAFFT version 7.407 (68) with 'addfragments' option and were placed in the corresponding phylogenetic tree of different As metabolism proteins (described above) respectively, using RAxML v8.4.1 (66) with -f v option. The placement of the target As metabolism protein-containing reads was visualized in iTOL (71) after processing the resulting visualization jplace file generated by RAxML with JPlace.to_iTOL.rb from the Enveomics Collection (72).

Alpha and beta community diversity estimation

The abundance-weighted average coverage of the sampled microbial communities achieved by our sequencing effort was estimated by Nonpareil (73, 74) based on the level of redundancy of the sequence reads of metagenomes. Mash distances (75) were used to assess the sample-to-sample sequence composition similarity at the whole metagenome level, based on shared kmers. Because the coverage of these metagenomics varied among different studies, the datasets were subsample to 2Gbp to make them comparable. Principle coordinates analysis (PCoA) and/or non-metrical multidimensional scaling (NMDS) plot were used to visualize the site-to-site similarity matrices of microbial community compositions based on Mash distances, OTU tables, and annotation counts from SEED subsystems based on Bray-Curtis distance. NMDS plot of annotation counts of *ars* [i.e. *arsR* (*arsR1*, *arsR2*, *arsR3*, *arsR4*), *acr3*, *arsB*, *arsC* (*arsC1*, *arsC2* and *acr2*), *arsM*, *arsI*, *arsP*, *arsH*], *aioA*, *arrA* genes from metagenomes and metatranscriptomes based on Bray-Curtis distances were used to visualize the site-to-site similarity of As-associated microbial community compositions.

The Adonis test was used to determine correlations between physicochemical variables measured at each site and similarity in microbial community composition (Mash distance) or *ars* genes abundances/transcriptional activities (Bray-Curtis distances) and tested for significance based on a permutation test with 999 interactions. The interactions between different physicochemical variables were assessed by PERMANOVA analysis while the correlation of physicochemical variables with *aioA* and *arrA* gene abundances and transcriptional activities in different sites were tested by Pearson correlation. These analyses were performed in R 3.5.1 with the vegan (70) and ggplot2 (76) packages.

Declarations

Data availability

Metagenomics and metatranscriptomics datasets have been deposited in National Center for Biotechnology Information (NCBI)'s Short Read Archive (SRA) database, under the bioproject PRJNA616041. NCBI SRA numbers for all datasets were provided in supplementary Table S1.

Acknowledgements

This research was supported, in part, by the U.S. National Science Foundation (Award No. 1831582 and 1759831) and Major Project of the Ministry of Science and Technology of Jiangxi Province (No. CK201302055). We also appreciate the support from the Brook Byers Institute for Sustainable Systems and the Hightower Chair at the Georgia Institute of Technology.

References

1. Oremland RS, Stolz JF. 2003. The ecology of arsenic. Science 300:939-944.

2. Zhu Y-G, Yoshinaga M, Zhao F-J, Rosen BP. 2014. Earth abides arsenic biotransformations. Annual Review of Earth and Planetary Sciences 42:443-467.

3. Saunders JK, Fuchsman CA, McKay C, Rocap G. 2019. Complete arsenic-based respiratory cycle in the marine microbial communities of pelagic oxygen-deficient zones. Proceedings of the National Academy of Sciences 116:9925-9930.

4. Sforna MC, Philippot P, Somogyi A, Van Zuilen MA, Medjoubi K, Schoepp-Cothenet B, Nitschke W, Visscher PT. 2014. Evidence for arsenic metabolism and cycling by microorganisms 2.7 billion years ago. Nature Geoscience 7:811-815.

5. Chen S-C, Sun G-X, Yan Y, Konstantinidis KT, Zhang S-Y, Deng Y, Li X-M, Cui H-L, Musat F, Popp D. 2020. The Great Oxidation Event expanded the genetic repertoire of arsenic metabolism and cycling. Proceedings of the National Academy of Sciences 117:10414-10421.

6. Mukhopadhyay R, Rosen BP. 2002. Arsenate reductases in prokaryotes and eukaryotes. Environmental Health Perspectives 110:745-748.

7. Rosen BP. 2002. Biochemistry of arsenic detoxification. FEBS Letters 529:86-92.

8. Qin J, Rosen BP, Zhang Y, Wang G, Franke S, Rensing C. 2006. Arsenic detoxification and evolution of trimethylarsine gas by a microbial arsenite S-adenosylmethionine methyltransferase. Proceedings of the National Academy of Sciences 103:2075-2080.

9. Chen J, Bhattacharjee H, Rosen BP. 2015. ArsH is an organoarsenical oxidase that confers resistance to trivalent forms of the herbicide monosodium methylarsenate and the poultry growth promoter roxarsone. Molecular Microbiology 96:1042-1052.

10. Chen J, Madegowda M, Bhattacharjee H, Rosen BP. 2015. ArsP: a methylarsenite efflux permease. Molecular Microbiology 98:625-635.

11. Yoshinaga M, Rosen BP. 2014. AC· As lyase for degradation of environmental organoarsenical herbicides and animal husbandry growth promoters. Proceedings of the National Academy of Sciences 111:7701-7706.

12. Chen J, Nadar VS, Rosen BP. 2017. A novel MAs (III)-selective ArsR transcriptional repressor. Molecular Microbiology 106:469-478.

13. Qin J, Fu H-L, Ye J, Bencze KZ, Stemmler TL, Rawlings DE, Rosen BP. 2007. Convergent evolution of a new arsenic binding site in the ArsR/SmtB family of metalloregulators. Journal of Biological Chemistry 282:34346-34355.

14. Slyemi D, Bonnefoy V. 2012. How prokaryotes deal with arsenic. Environmental Microbiology Reports 4:571-586.

15. Ahmann D, Roberts AL, Krumholz LR, Morel FM. 1994. Microbe grows by reducing arsenic. Nature 371:750-750.

16. Lebrun E, Brugna M, Baymann F, Muller D, Lievremont D, Lett M-C, Nitschke W. 2003. Arsenite oxidase, an ancient bioenergetic enzyme. Molecular Biology and Evolution 20:686-693.

17. Malasarn D, Saltikov C, Campbell K, Santini J, Hering J, Newman D. 2004. arrA is a reliable marker for As (V) respiration. Science 306:455-455.

18. Takahashi Y, Minamikawa R, Hattori KH, Kurishima K, Kihou N, Yuita K. 2004. Arsenic behavior in paddy fields during the cycle of flooded and non-flooded periods. Environmental Science & Technology 38:1038-1044.

19. Yamaguchi N, Nakamura T, Dong D, Takahashi Y, Amachi S, Makino T. 2011. Arsenic release from flooded paddy soils is influenced by speciation, Eh, pH, and iron dissolution. Chemosphere 83:925-932.

20. Meharg AA, Zhao F-J. 2012. Biogeochemistry of arsenic in paddy environments, p 71-101, Arsenic & Rice. Springer.

21. Zhang S-Y, Zhao F-J, Sun G-X, Su J-Q, Yang X-R, Li H, Zhu Y-G. 2015. Diversity and abundance of arsenic biotransformation genes in paddy soils from southern China. Environmental Science & Technology 49:4138-4146.

22. Huang H, Jia Y, Sun G-X, Zhu Y-G. 2012. Arsenic speciation and volatilization from flooded paddy soils amended with different organic matters. Environmental Science & Technology 46:2163-2168.

23. Jia Y, Huang H, Zhong M, Wang F-H, Zhang L-M, Zhu Y-G. 2013. Microbial arsenic methylation in soil and rice rhizosphere. Environmental Science & Technology 47:3141-3148.

24. Jia Y, Huang H, Chen Z, Zhu Y-G. 2014. Arsenic uptake by rice is influenced by microbe-mediated arsenic redox changes in the rhizosphere. Environmental Science & Technology 48:1001-1007.

25. Xiao K-Q, Li L-G, Ma L-P, Zhang S-Y, Bao P, Zhang T, Zhu Y-G. 2016. Metagenomic analysis revealed highly diverse microbial arsenic metabolism genes in paddy soils with low-arsenic contents. Environmental Pollution 211:1-8.

26. Zhang SY, Su JQ, Sun GX, Yang Y, Zhao Y, Ding J, Chen YS, Shen Y, Zhu G, Rensing C. 2017. Land scale biogeography of arsenic biotransformation genes in estuarine wetland. Environmental Microbiology 19:2468-2482.

27. Ragot SA, Huguenin-Elie O, Kertesz MA, Frossard E, Bünemann EK. 2016. Total and active microbial communities and phoD as affected by phosphate depletion and pH in soil. Plant and Soil 408:15-30.

28. Thompson I, Bailey M, Ellis R, Maguire N, Meharg A. 1998. Response of soil microbial communities to single and multiple doses of an organic pollutant. Soil Biology and Biochemistry 31:95-105.

29. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research 35:7188-7196.

30. Kuang J, Huang L, He Z, Chen L, Hua Z, Jia P, Li S, Liu J, Li J, Zhou J. 2016. Predicting taxonomic and functional structure of microbial communities in acid mine drainage. The ISME Journal 10:1527-1539.

31. Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genetics Research 23:23-35.

32. Corel E, Méheust R, Watson AK, McInerney JO, Lopez P, Bapteste E. 2018. Bipartite network analysis of gene sharings in the microbial world. Molecular Biology and Evolution 35:899-913.

33. Luo J, Bai Y, Liang J, Qu J. 2014. Metagenomic approach reveals variation of microbes with arsenic and antimony metabolism genes from highly contaminated soil. PLoS One 9: e108185.

34. Rodriguez-r LM, Konstantinidis KT. 2014. Estimating coverage in metagenomic data sets and why it matters. The ISME Journal 8:2349-2351.

35. Yang H-C, Rosen BP. 2016. New mechanisms of bacterial arsenic resistance. Biomedical Journal 39:5-13.

36. Bhattacharjee H, Rosen BP. 2007. Arsenic metabolism in prokaryotic and eukaryotic microbes, p 371-406, Molecular microbiology of heavy metals. Springer.

37. Duval S, Ducluzeau A-L, Nitschke W, Schoepp-Cothenet B. 2008. Enzyme phylogenies as markers for the oxidation state of the environment: the case of respiratory arsenate reductase and related enzymes. BMC Evolutionary Biology 8:206.

38. Silver S, Phung LT. 2005. Genes and enzymes involved in bacterial oxidation and reduction of inorganic arsenic. Applied and Environmental Microbiology 71:599-608.

39. Bhattacharyya P, Tripathy S, Kim K, Kim S-H. 2008. Arsenic fractions and enzyme activities in arsenic-contaminated soils by groundwater irrigation in West Bengal. Ecotoxicology and Environmental Safety 71:149-156.

40. Maizel D, Blum JS, Ferrero MA, Utturkar SM, Brown SD, Rosen BP, Oremland RS. 2016. Characterization of the extremely arsenic-resistant Brevibacterium linens strain AE038-8 isolated from contaminated groundwater in Tucumán, Argentina. International Biodeterioration & Biodegradation 107:147-153.

41. Ma B, Wang H, Dsouza M, Lou J, He Y, Dai Z, Brookes PC, Xu J, Gilbert JA. 2016. Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in eastern China. The ISME Journal 10:1891-1901.

42. Plassart P, Prévost-Bouré NC, Uroz S, Dequiedt S, Stone D, Creamer R, Griffiths RI, Bailey MJ, Ranjard L, Lemanceau P. 2019. Soil parameters, land use, and geographical distance drive soil bacterial communities along a European transect. Scientific Reports 9:1-17.

43. Cao H, Chen R, Wang L, Jiang L, Yang F, Zheng S, Wang G, Lin X. 2016. Soil pH, total phosphorus, climate and distance are the major factors influencing microbial activity at a regional spatial scale.

Scientific Reports 6:1-10.

44. DeForest JL, Scott LG. 2010. Available organic soil phosphorus has an important influence on microbial community composition. Soil Science Society of America Journal 74:2059-2066.

45. Heuck C, Weig A, Spohn M. 2015. Soil microbial biomass C: N: P stoichiometry and microbial use of organic phosphorus. Soil Biology and Biochemistry 85:119-129.

46. Somenahally AC, Hollister EB, Loeppert RH, Yan W, Gentry TJ. 2011. Microbial communities in rice rhizosphere altered by intermittent and continuous flooding in fields with long-term arsenic application. Soil Biology and Biochemistry 43:1220-1228.

47. Castrillo G, Teixeira PJPL, Paredes SH, Law TF, de Lorenzo L, Feltcher ME, Finkel OM, Breakfield NW, Mieczkowski P, Jones CD. 2017. Root microbiota drive direct integration of phosphate stress and immunity. Nature 543:513-518.

48. Jackson BP, Miller W. 2000. Effectiveness of phosphate and hydroxide for desorption of arsenic and selenium species from iron oxides. Soil Science Society of America Journal 64:1616-1622.

49. Kögel-Knabner I, Amelung W, Cao Z, Fiedler S, Frenzel P, Jahn R, Kalbitz K, Kölbl A, Schloter M. 2010. Biogeochemistry of paddy soils. Geoderma 157:1-14.

50. Borch T, Fendorf S. 2007. Phosphate interactions with iron (hydr) oxides: Mineralization pathways and phosphorus retention upon bioreduction. Developments in Earth and Environmental Sciences 7:321-348.

51. Wang Y, Yuan J-H, Chen H, Zhao X, Wang D, Wang S-Q, Ding S-M. 2019. Small-scale interaction of iron and phosphorus in flooded soils with rice growth. Science of the Total Environment 669:911-919.

52. Ji Y, Luo W, Lu G, Fan C, Tao X, Ye H, Xie Y, Shi Z, Yi X, Dang Z. 2019. Effect of phosphate on amorphous iron mineral generation and arsenic behavior in paddy soils. Science of The Total Environment 657:644-656.

53. Liu W, Zhu Y, Hu Y, Williams P, Gault A, Meharg AA, Charnock J, Smith F. 2006. Arsenic sequestration in iron plaque, its accumulation and speciation in mature rice plants (Oryza sativa L.). Environmental Science & Technology 40:5730-5736.

54. Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11:485.

55. Su X, Pan W, Song B, Xu J, Ning K. 2014. Parallel-META 2.0: enhanced metagenomic data analysis with functional annotation, high performance computing and advanced visualization. PloS One 9: e89323.

56. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7:335.

57. Rho M, Tang H, Ye Y. 2010. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Research 38:e191-e191.

58. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

59. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Research 42:D206-D214.

60. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15:550.

61. Kolde R, Kolde MR. 2015. Package 'pheatmap'. R Package 1.

62. Wu Y-W, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32:605-607.

63. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7:e7359.

64. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K. 2016. A new view of the tree of life. Nature Microbiology 1:1-6.

65. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. 2002. The Pfam protein families database. Nucleic Acids Research 30:276-280.

66. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.

67. Peng Y, Leung HC, Yiu S-M, Chin FY. 2010. IDBA–a practical iterative de Bruijn graph de novo assembler. Research in Computational Molecular Biology. Lisbon: Proceedings of the 14th Annual International Conference: 426-444.

68. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30:772-780.

69. Nayfach S, Pollard KS. 2015. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biology 16:51.

70. Wagner H. 2015. Vegan: community ecology package. R package.

71. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Research 44:W242-W245.

72. Rodriguez-R LM, Konstantinidis KT. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints.

73. Rodriguez-r LM, Konstantinidis KT. 2014. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinformatics 30:629-635.

74. Rodriguez-R LM, Gunturu S, Tiedje JM, Cole JR, Konstantinidis KT. 2018. Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. MSystems 3:e00039-18.

75. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology 17:132.

76. Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer.

Figures

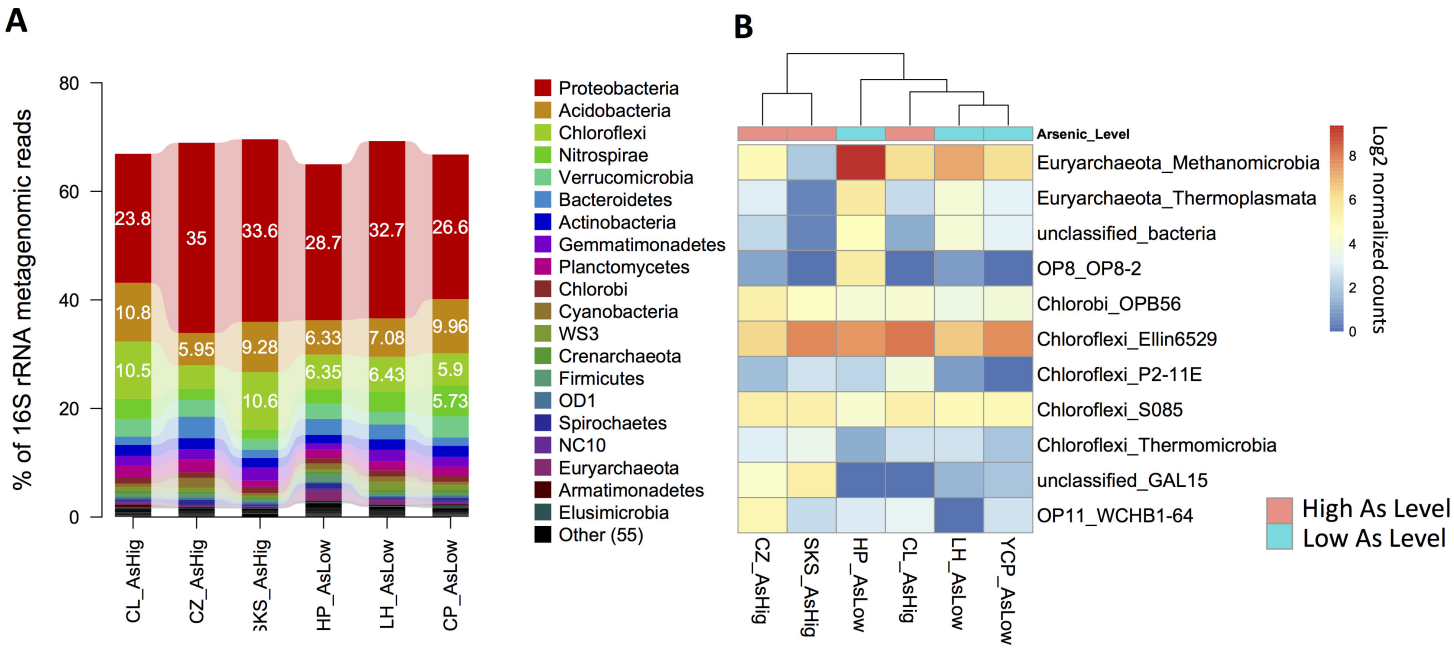

Figure 1

Comparison of microbial community composition in high vs low As sites. (A) Phylum-level community composition and abundance of microbes in paddy soils based on taxonomic classification of identified 16S rRNA gene sequence fragments. The relative abundance of each taxon was normalized by the total number of 16S rRNA gene sequence fragments (both assigned and unassigned) obtained in each corresponding metagenome. Only the top 20 most abundant phyla are shown. (B) Heatmap of microbes at the class level showing statistically significant differences in abundance between high (red) and low (blue) As paddy soils (negative binomial test, adjusted p value < 0.05). Log2 normalized counts were used to calculate the abundances of different microbes at the class level.

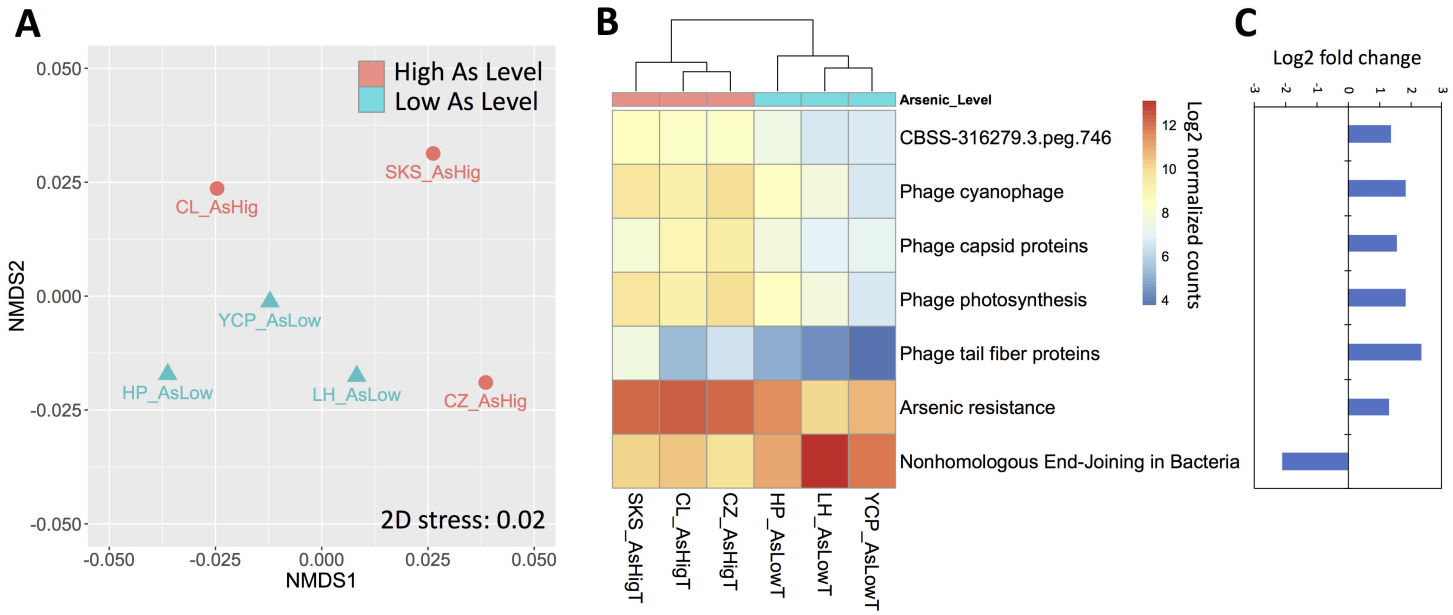

Figure 2

The effect of high As levels on the functional composition of microbial communities. (A) NMDS plot of the Bray-Curtis metric based on the counts of SEED subsystems (level 3) annotated from metagenomes in high (red) and low (blue) As paddy soils. (B) Heatmap of SEED subsystems (level 1) annotated from metatranscriptomes showing statistically significant differences in abundance between high (red) and low (blue) As paddy soils (negative binomial test, adjusted p value < 0.05). Log2 normalized counts were used to calculate the abundances of different SEED subsystems (level 1) annotated from metatranscriptomes. (C) Log2 fold change of SEED subsystems (level 1) annotated from metatranscriptomes that were significantly differentially abundant between high and low As paddy soils (negative binomial test, adjusted p value < 0.05).

Figure 3

Relative abundance and transcription of ars, aioA and arrA genes in high and low As paddy soils. (A) The relative abundance and transcription of ars genes is the cumulative of arsR (arsR1, arsR2, arsR3, arsR4), acr3, arsB, arsC (arsC1, arsC2 and acr2), arsM, arsI, arsP and arsH genes. Relative abundance of As metabolism genes in metagenomes was normalized by the statistic RPKG (reads per kb per genome equivalent, see materials and methods for more details). The As metabolism genes transcriptional activity was evaluated by the relative abundance of As metabolism genes in metatranscriptomics (RPKM reads per kb per millions of reads) normalized by the relative abundance of the same gene identified in metagenomics (RPKM). Horizontal lines represent the median value; * indicates statistically different means at $p < 0.05$ based on one-way ANOVA analysis. (B) Transcriptional activity of the As metabolism genes and the cellular location of their encoded proteins. Transcriptional activities (the same values as in Figure 3A) of ars (acr3, arsB, arsC, arsM, arsI, arsH, arsP and arsR), aioA and arrA genes in high (red) and low (blue) As paddy soils visualized in a heatmap that was produced using the pheatmap package in R 3.5.1. The color ranges from red to blue indicate transcriptional activity from high to low.

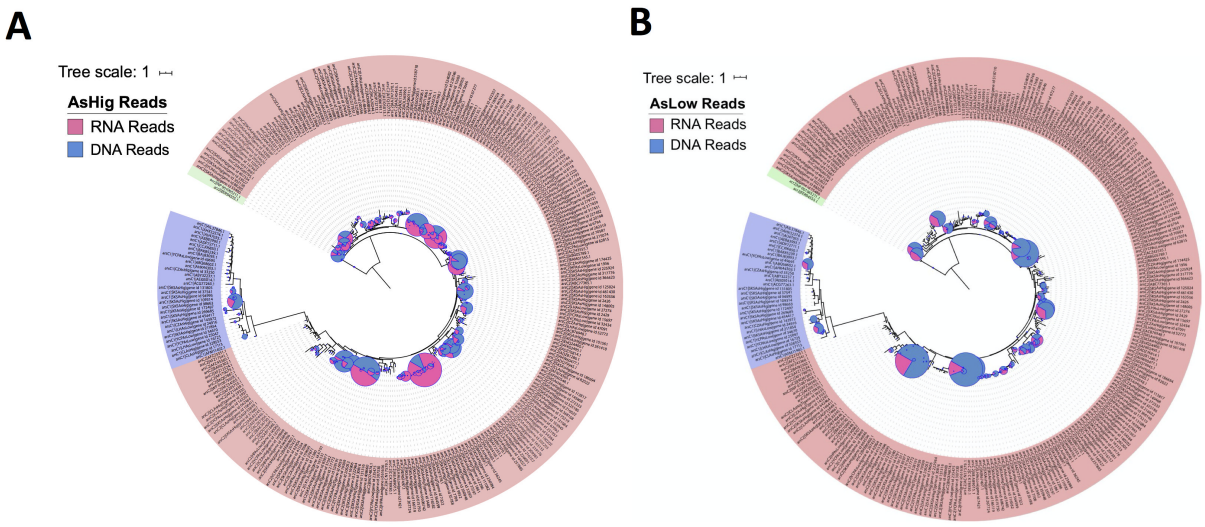

Figure 4

Community-wide shifts as an effect of high As levels. Phylogenetic placement of arsC-carrying reads identified in metagenomes (DNA reads, in blue) and metatranscriptomes (RNA reads, in red) from high (A) and low As paddy soils (B). For this analysis, the three metagenomes and metatranscriptomes from the same As level (i.e. high or low) sites were combined together. The pie size indicates the DNA abundance and the ratio indicate transcriptional activity for each clade (i.e., fraction of RNA vs DNA reads assigned to the class with the same thresholds for a match; see Methods for further details). The clades are colored on the outside based on the ArsC type, i.e., ArsC1 (blue), ArsC2 (red) and ACR2 (green).

Phylogenetic placement of other As metabolism genes carrying reads are provided in the supplementary. Note that several clades in the high As recruited more RNA than DNA reads (red color dominates) whereas the opposite pattern was observed for most clades in the low As datasets.

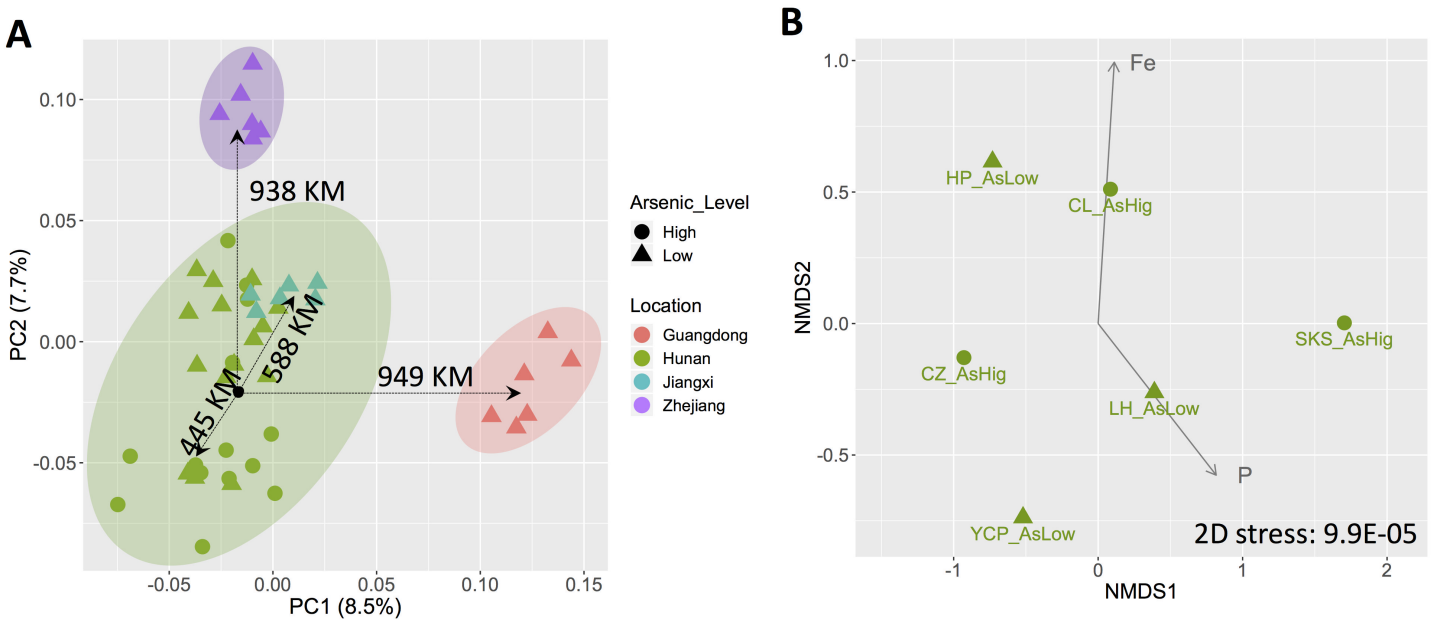

Figure 5

Physicochemical factors affecting whole microbial community composition. (A) PCoA plot of site-to-site similarity matrix based on Mash distances of whole metagenomes from this study and 13 selected metagenomes from previous studies. Symbols in different colors indicate samples from different provinces (green: Hunan; red: Guangdong; blue: Jiangxi; purple: Zhejiang). The shapes of symbols indicate different As levels of these samples (circle: high As levels; triangle: low As levels). Note that samples were clustered according to their geographic distance, not As levels (i.e. CL-CZ: 445 km, CL-JX: 588 km, CL-ZJ: 938 km, CL-GD: 949 km; most distinct metagenomes were clustered further apart). (B) NMDS plot of Mash distances of whole metagenomes determined by this study. The correlation of microbial community differences with environmental factors are shown (arrows). Fe and P concentrations were significantly correlated with the variations observed between the microbial communities based on the Adonis test ($p < 0.05$).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ReSubSIAsGenesPaddySoilAEM.pdf

- RevisedSIAsGenesPaddySoilTables.xlsx