The Great Oxidation Event expanded the genetic repertoire of arsenic metabolism and cycling

Song-Can Chen1,a,b,c,d, Guo-Xin Sun1,a,b, Yu Yan2, Konstantinos T. Konstantinidis3,e,f, Si-Yu Zhang2, Ye Deng1, Xiao-Min Li1,a, Hui-Ling Cuia2,3, Florin Musatb, Denny Popp2, Barry P. Rosen2,5, and Yong-Guan Zhua1,2

aState Key Lab of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, 100085 Beijing, China; bDepartment of Isotope Biogeochemistry, Helmholtz Centre for Environmental Research – UFZ, 04318 Leipzig, Germany; cCollege of Resources and Environment, University of Chinese Academy of Sciences, 100049 Beijing, China; dDepartment of Environmental Science and Engineering, Huqiao University, 361021 Xiamen, China; eSchool of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332; fSchool of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332; gDepartment of Environmental Microbiology, Helmholtz Centre for Environmental Research – UFZ, 04318 Leipzig, Germany; hDepartment of Cellular Biology and Pharmacology, Herbert Wertheim College of Medicine, Florida International University, Miami, FL 33199; and iKey Lab of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, 361021 Xiamen, China

Edited by Donald E. Canfield, Institute of Biology and Nordic Center for Earth Evolution (NordCEE), University of Southern Denmark, Odense M., Denmark, and approved March 20, 2020 (received for review January 25, 2020)

The rise of oxygen on the early Earth about 2.4 billion years ago reorganized the redox cycle of harmful metal(loids), including that of arsenic, which doubtlessly imposed substantial barriers to the physiology and diversification of life. Evaluating the adaptive biological responses to these environmental challenges is inherently difficult because of the paucity of fossil records. Here we applied molecular clock analyses to 13 gene families participating in principal pathways of arsenic resistance and cycling, to explore the nature of early arsenic biogeoecycles and decipher feedbacks associated with planetary oxygenation. Our results reveal the advent of nascent arsenic resistance systems under the anoxic environment predating the Great Oxidation Event (GOE), with the primary function of detoxifying reduced arsenic compounds that were abundant in Archean environments. To cope with the increased toxicity of oxidized arsenic species that occurred as oxygen built up in Earth’s atmosphere, we found that parts of preexisting detoxification systems for trivalent arsenicals were merged with newly emerged pathways that originated via convergent evolution. Further expansion of arsenic resistance systems was made feasible by incorporation of oxygen-dependent enzymatic pathways into the detoxification network. These genetic innovations, together with adaptive responses to other redox-sensitive metals, provided organisms with novel mechanisms for adaptation to changes in global biogeoecycles that emerged as a consequence of the GOE.

One of life’s earliest challenges was coping with the toxicity of harmful metal(loids) (1). Understanding the nature and timing of the onset of protective mechanisms is essential for the study of early evolution of Earth and life, yet limited information is available. Arsenic is the most ubiquitous toxic metalloid in nature, with two biologically relevant oxidation states: trivalent arsenite and pentavalent arsenate. Arsenite is generally more toxic than arsenate, and perturbs the physiology of prokaryotes at micromolar levels (2, 3). Relatively high amounts (>20 μM) of dissolved arsenic are nowadays frequently found in oceanic hydrothermal vents or hot springs, environments that may have conditions analogous to similar niches of primordial Earth. For this reason, resistance pathways for transport and biotransformation of arsenic are believed to have emerged early in the evolution of life on Earth (4–6). Environmentally, the rise of atmospheric oxygen during the Great Oxidation Event (GOE) ~2.4 billion years ago (Bya) is thought to have fundamentally changed arsenic chemistry in the Earth’s surface and oceans (2, 7). Prior to the GOE, reduced arsenic species (i.e., arsenite) would have predominated over oxidized arsenics (i.e., arsenate) because the atmosphere and oceans were anoxic and reducing (4, 6, 8). Continental weathering of arsenic at this time is negligible under an atmosphere with very low oxygen levels (<<0.001% compared with present atmospheric level) (9). The rise of atmospheric oxygen (~1% of present atmospheric levels) during the GOE between 2.4 and 2.3 Bya most likely led to intense oxidative weathering of arsenic-bearing minerals that liberated continental arsenic, predominantly as arsenate, for delivery to oceans from rivers (3, 10). These processes would have resulted in the widespread appearance of oxidized arsenic species in the environment. We hypothesized that these dramatic shifts in the redox state of arsenicals and their bioavailability imposed a strong selective pressure on ancient microorganisms toward acquisition of novel enzymatic systems conferring arsenic resistance. Current microbial fossil records lack the power to resolve the timing and causes of the origin of these tolerance and detoxification mechanisms.

Molecular and genetic studies have identified many arsenic resistance (ars) genes in extant organisms (SI Appendix, Table S1). These include efflux permeases, redox enzymes, methyltransferases, and transcriptional repressors. Arsenite efflux is catalyzed by two evolutionarily unrelated groups of arsenic | detoxification | evolution | oxygen | biogeochemistry

Author contributions: B.P.R. and Y.-G.Z. designed research; S.-C.C., G.-X.S., X.-M.L., and H.-L.C. performed research; S.-C.C., Y.Y., K.T.X., S.-Y.Z., Y.D., and D.P. analyzed data; and S.-C.C., G.-X.S., Y.Y., F.M., B.P.R., and Y.-G.Z. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: The sequence data used in this study were provided as supplementary datasets S1-S3.

1 S.-C.C. and G.-X.S. contributed equally to this work.

2To whom correspondence may be addressed. Email: bosen@fudan.edu.cn or ygzhu@rcees.ac.cn.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2001063117/-/DCSupplemental.

First published April 29, 2020.
efflux permeases: ArsB and Acr3 (11). Arsenate detoxification is catalyzed by reductases (ArsC), with homology to the glutaredoxin family (ArsC1), to low-molecular-weight phosphatases (ArsC2), or by members of the CDC25 family of dual-specific phosphatases (Acr2), respectively (12). These enzymes reduce intracellular arsenate to arsenite, the substrate of the two arsenic efflux permeases. Additionally, arsenite can be methylated by ArsM, an arsenite S-adenosylmethionine (SAM) methyltransferase, to the more toxic species methylarsenite and dimethylarsenite. In air, these are oxidized nonenzymatically to the largely nontoxic pentavalent species. However, methylarsenite can be also detoxified by active extrusion from cells catalyzed by the methylarsenite-specific efflux permease ArsP (13), oxidation to methylarsenate by the methylarsenite-specific oxidase ArsH (14, 15), or demethylation to less toxic arsenite by the ArsR C-As lylase that cleaves the carbon–arsenic bond in methylarsenite (16). Arsenic resistance genes are usually organized in operons, which are nearly always under control of an ArsR transcriptional repressor. Four different ArsRs, in which each an arsenic binding site is located at a different place in the protein structure, have been described, with three (ArsR1, ArsR2, and ArsR3) regulated selectively by arsenite (17) and one (ArsR4) by methylarsenite (18).

Here, we estimate the geological birth date of 13 arsenic resistance genes in relation to the GOE, using molecular clock analyses. The detailed evolutionary histories for each gene family were reconstructed by comparing their gene phylogenies with the phylogeny of organisms (the tree of life) under an explicit model of macroevolution events including gene birth, transfer, duplication, and loss. The occurrence of each arsenic detoxification gene was examined with respect to the taxonomy and physiology of the host microorganisms to provide independent evidence for our molecular dating analysis.

Results
Phylogenetic Distribution of Arsenic Detoxification Genes. Protein sequences of the 13 arsenic resistance genes were acquired from genomes of 645 bacteria, 88 archaea, and 53 eukaryotes, representative of phylogenetic diversity across the three domains of life (19). The presence/absence of arsenic resistance genes in each of the sampled taxa were collapsed at phylum level and plotted against a reference tree reconstructed from a concatenated alignment of 16 ribosomal proteins (Fig. 1). The distinct phyletic patterns divide the 13 genes into three sets (A–C). Genes in set A, including arsM, acr3, arsC2, arsP, and arsR1, are widely distributed among major lineages of bacteria, archaea, and/or eukaryotes, whereas set B comprises seven genes (arsI, arsB, arsR3, arsH, arsC1, arsR2, and arsR4) found mostly in aerobes that are more sparsely distributed compared with those in set A. Set C comprised a single gene, acr2, with homologs detected only in eukaryotes. The descent patterns suggest that the genes in set A may have emerged as the earliest arsenic detoxification systems, followed by those in sets B and C. However, promiscuous horizontal gene transfer (HGT) of arsenic resistance genes across species (20, 21), as exemplified by apparent incongruence between individual gene phylogeny and the organism backbone (SI Appendix, Figs. S1–S14 and Table S2), obscured our capability to coordinate these genes along the geological timeline with merely phyletic patterns (22).

Gene Birth Date of Arsenic Detoxification Genes. To estimate the timing of the origin of the 13 arsenic resistance genes, we conducted a series of Bayesian molecular clock analyses, using a tree reconciliation algorithm, which explicitly models HGT and generates gene birth dates by mapping gene phylogeny onto a chronogram of species. We tested gene ages against chronograms modeled with autorecorrelated rate (analyses 1 to 6) and independent rate clock (analyses 7 to 12). For each clock model, a set of six independent analyses were performed to evaluate the robustness of the results to prior assumptions of root age (analyses 1 and 7), subsampling of fossil calibrations (analyses 3, 4, 9, and 10), and alternative topologies (analyses 5, 6, 11, and 12). Median gene ages under 12 analytical scenarios are shown in Tables 1 and 2, and the uncertainties associated with the results from all these analyses were integrated over to provide composite credibility interval for each gene family (Fig. 2). Although the timing of arsM and acr3 varied under different prior assumptions, all analyses consistently recovered 95% credibility intervals entirely within the Archean eon, suggesting that they originated before the GOE. For arsC2, arsR1, and arsP, we estimate that the median gene ages are before or at the beginning of the Paleoproterozoic period, with composite 95% confidence intervals overlapping with the GOE. In contrast, arsB, arsI, arsH, arsR2, arsR3, arsC1, acr2, and arsR4 are estimated to have evolved near the end of or significantly after the GOE. To assess the sensitivity of our results to alternative species topologies, we also reconciled gene families against 100 reference trees reconstructed from ribosomal proteins or small subunit ribosomal RNA (SSU rRNA). The results show only slightly differences in estimates of gene ages (SI Appendix, Fig. S20), which further supports our initial interpretation of the data. Overall, our analyses are consistent with an expansion of microbial arsenic resistance systems in response to the rise of atmospheric oxygen.

Physiology Bears Out the Age of Arsenic Detoxification Genes. We attempted to further validate these conclusions by analyzing the physiology of the host microorganisms. Organisms were classified either as aerobes (including facultative anaerobes) or anaerobes, based on their capability to utilize oxygen as a terminal electron acceptor. We found that all the genes predicted to originate in anoxic environment after the GOE are over-represented in aerobes, but are nearly absent in strict anaerobes (Fig. 3). Furthermore, the genes predicted to have a more ancient origin were found among both anaerobes and aerobes, including the ancient lineages of methanogens and acetogens (Fig. 3). This implies an early origin of these genes in an anoxic or microaerobic environment before or at the beginning of the GOE. They dispersed into theoxic environment after the rise of oxygen, as predicted by our evolutionary model. To further probe the robustness of our predictions, we tested the correlation of arsenic resistance systems with the physiology of the host microorganisms on a more densely sampled set of taxa encompassing more than 2,000 species. We found similar patterns of gene distribution across anaerobes/aerobes, suggesting that our results are broadly conserved independent of taxonomic sampling (SI Appendix, Fig. S21).

Discussion
Arsenic Detoxification Systems before the GOE. Our molecular clock analyses indicate that enzymatic pathways acting on trivalent arsenite, including arsenite efflux and arsenite methylation, constituted the core of microbial arsenic resistance systems before the rise of atmospheric oxygen (Fig. 4). Our results are consistent with geochemical models that predict the predominance of reduced arsenic compounds in the anoxic Archean biosphere (2, 3, 6, 10). Formation of traces of arsenate in the Archean, creating a selective pressure before the GOE (6), could have occurred via microbial mediated arsenite oxidation processes such as anoxic photosynthesis (5) or nitrate-dependent respiration (23). Alternatively, arsenate could have been formed during transient atmospheric oxygenation events documented back to ∼3.0 Bya (9, 24–28). However, our molecular clock analyses placed the earliest origin of the arsenate resistance system coincident with the onset of GOE (Fig. 2). This is consistent with recent...
The rise of oxygen in Earth’s atmosphere since the GOE both triggered global-scale oxidation of reduced arsenic species and led to widespread bioavailability of arsenate (3, 10). Our analyses indicate that the ancient arsenic resistance networks, optimized for detoxification of reduced arsenic in the anoxic Archean Earth, expanded to accommodate these environmental shifts (Fig. 4). In the face of these challenges, components of arsenate reduction systems (including a new efflux permease, ArsM, and arsenate reductases) evolved independently through convergent evolution after the GOE. The recurrent innovation of counterparts of ancient arsenate resistance devices is in agreement with enhanced arsenate stress because of gradually increasing oxygen levels after the Archean (3, 8). With the appearance of molecular oxygen, the ancient arsenic detoxification pathways were remodeled for detoxification of inorganic arsenic. For example, arsenite methylation process catalyzed by ArsM appeared in marine shales, suggesting that arsenate began to accumulate in the ocean only after the Archean eon (10), and compatible with the causal role of the GOE in altering the arsenic chemistry on Earth’s surface and driving the genetic expansion of arsenic resistance system.

The early origin of the arsenite efflux permease encoded by arcR3, together with its wide distribution among living organisms (Fig. 1), underpins the fundamental role of efflux mechanisms in heavy metal resistance (29, 30). In contrast, the physiological function of arsenite methylation in anoxic Archean environments remains unclear. The higher toxicity of the trivalent methylated product methylearsenate calls into question the commonly held assumption that methylation is a detoxification process. An attractive hypothesis is that the transient oxygenation of the Archean atmosphere (25, 26) and the existence of oxygen oases in local, shallow marine settings (24, 31) could have provided niches where microbial arsenite methylation could have operated as a detoxification pathway. Alternatively, methylation has been proposed as an antibiotic-producing process in Archean environments, with methylearsenate being a primitive antibiotic (32, 33). Further studies will clarify the function of ArsM in anoxic environments and its contribution to arsenic cycling and overall toxicity in ancient ecosystems.

Expansion of the Arsenic Resistance Network as a Consequence of the GOE. The rise of oxygen in Earth’s atmosphere since the GOE both triggered global-scale oxidation of reduced arsenic species and led to widespread bioavailability of arsenate (3, 10). Our analyses indicate that the ancient arsenic resistance networks, optimized for detoxification of reduced arsenic in the anoxic Archean Earth, expanded to accommodate these environmental shifts (Fig. 4). In the face of these challenges, components of arsenate reduction systems (including a new efflux permease, ArsM, and arsenate reductases) evolved independently through convergent evolution after the GOE. The recurrent innovation of counterparts of ancient arsenate resistance devices is in agreement with enhanced arsenate stress because of gradually increasing oxygen levels after the Archean (3, 8). With the appearance of molecular oxygen, the ancient arsenic detoxification pathways were remodeled for detoxification of inorganic arsenic. For example, arsenite methylation process catalyzed by ArsM could be recruited as a detoxification pathway under oxic settings. Its products, the toxic trivalent methylearsenate and...
dimethylarsenite, would be oxidized nonenzymatically by dioxygen into relatively innocuous methylethyl and di- methylethylarsenate. However, the influence of dioxygen did not stop here. Our results further suggest that two new obligate oxygen-dependent methylethylarsenite resistance enzymes, ArsH and ArsI, arose during or after the GOE. Concurrent with the evolution of these new oxygen-dependent methylethylarsenite detoxification enzymes, recurrent expansion of ArsR families after the GOE resulted in formation of diverse *ars* operons present in extant prokaryotes and enabled regulatory fine-tuning of *ars* genes throughout different ages of the Earth evolution (17).

### Conclusion and Implications
The timing we propose for the birth of arsenic resistance gene-families supports a shifted marine arsenic cycle across Archean–Proterozoic boundary. We observed an early origin of metabolic functions including methylation and excretion

#### Table 1. Birth Age of 13 arsenic resistance genes estimated under analytical scenarios 1 to 6

| Analysis | 1 | 2 | 3 | 4 | 5 | 6 |
|----------|---|---|---|---|---|---|
| Model assumptions and calibrations | Rate model* | Auto-correlated | Auto-correlated | Auto-correlated | Auto-correlated | Auto-correlated | Auto-correlated |
| Calibration† | Full set | Full set | Full set | Full set | Full set | Full set | Full set |
| Root prior‡ | U(3.35, 4.38) | U(3.35, 4.38) | U(3.35, 4.38) | U(3.35, 4.38) | U(3.35, 4.38) | U(3.35, 4.38) |
| Topology§ | ML | ML | ML | ML | ML | ML |
| Gene age (Gyr)$^*$ | 3.55 (3.27-3.72) | 3.59 (3.31-3.79) | 3.69 (3.45-3.89) | 3.50 (3.28-3.76) | 3.57 (3.29-3.83) | 3.62 (3.40-3.86) |

#### Table 2. Birth Age of 13 arsenic resistance genes estimated under analytical scenarios 7 to 12

| Analysis | 7 | 8 | 9 | 10 | 11 | 12 |
|----------|---|---|---|---|---|---|
| Model assumptions and calibrations | Rate model* | Uncorrelated | Uncorrelated | Uncorrelated | Uncorrelated | Uncorrelated | Uncorrelated |
| Calibration† | Full set | Full set | Full set | Full set | Full set | Full set | Full set |
| Root prior‡ | U(3.35, 4.38) | U(3.35, 4.38) | U(3.35, 4.38) | U(3.35, 4.38) | U(3.35, 4.38) | U(3.35, 4.38) |
| Topology§ | ML | ML | ML | ML | ML | ML |
| Gene age (Gyr)$^*$ | 3.40 (3.23-3.61) | 3.37 (3.23-3.72) | 3.44 (3.30-3.66) | 3.40 (3.24-3.73) | 3.40 (3.24-3.72) | 3.45 (3.24-3.76) |

$^*$Auto-correlated, autocorrelated rate model; Uncorrelated, uncorrelated rate model.

$^†$–Cyanobacteria, subsampled calibration points without Cyanobacteria; –Rhodophyta, subsampled calibration points without Rhodophyta.

$^‡$U, uniform distribution (upper, lower); Γ: Gamma distribution (mean; SD).

$^§$ML, maximum likelihood tree of ribosomal proteins; MT: alternative topology reflecting minority bipartitions; Three-domain tree: tree topology wherearchaea and eukaryotes are sister group.

$^*$Median age estimates of gene birth nodes, with 95% confidence intervals in parentheses; Gyr, billion years.
of arsenic during the Archaean eon, which is in accord with the fossil evidence indicating the occurrence of microbial arsenic metabolism and cycling 2.72 Bya (34). Our prediction of continuous innovation of gene families toward detoxification of oxidized arsenic species is in agreement with recent analysis of marine shales that inferred a sharp increase of dissolved arsenate from ∼2.48 Bya onward (10). The persistence of *ars* genes among distinct microbial lineages over billions of years implies a temporal continuity of arsenic stress (2).

The genetic expansion of arsenic resistance systems across the GOE would have entailed fitness advantages leading to success and diversification of life in the new redox landscape, which in turn remodeled the transition of metal chemistry on the Earth’s surface. Our molecular analysis, together with the innovations of protective mechanisms against other elements (35, 36) (e.g., Cu and Zn), provides a crucial constraint on the response of global biosphere to the major transitions in cycles of toxic, redox-sensitive metals.

**Methods**

**Genomic Sampling and Reconstruction of Species Tree.** A previously reported tree of life was used as template for reconstruction of species tree (19). A total of 786 representative species with a completely sequenced genome were sampled from the original dataset (see Dataset S1 for accession number). The ribosomal protein tree was inferred with RAxML v8.4.1 (37), using the PROTGAMMALG evolution model. To reconstruct the SSU rRNA tree, an alignment was generated from SSU rRNA genes of the sampled organisms, using the SINA alignment algorithm (38). One representative SSU rRNA gene was selected for species with multiple copies. Phylogenetic trees were calculated under the GTRCAT model, using RAxML. A total of 204 and 300 bootstrap replicates were conducted for ribosomal protein and SSU rRNA gene phylogenies, respectively, according to extended majority-rule consensus (MRE)-based bootstopping criteria. The oxygen requirement for each selected species was retrieved from Genomes OnLine Database (GOLD) (39) and literature reviews.

**Molecular Dating of the Tree of Life.** The divergence time of species tree was estimated with PhyloBayes, using a fixed RAxML phylogeny of ribosomal proteins, a CAT20 substitutional model, a birth–death process, and four gamma categories (40). The CAT20 model was chosen because preliminary
tests showed that analyses using a full CAT model failed to converge within a reasonable time (>2 mo). Both the autocorrelated lognormal (-ln) and uncorrelated gamma multiplier (-ugam) relaxed clocks were applied to model the rate variation across lineages (41). Bayesian cross-validation implemented in PhyloBayes was used to test whether one of two clock models fits the data better.

The clocks were calibrated with eight sets of temporal constraints (SI Appendix, Fig. S15 and Table S4) that are directly linked to fossil and geochemical evidence, as described previously (22, 42). The age of the last universal common ancestor (root) was constrained between 4.38 Bya (approximating earliest habitability evidence) (43, 44) and 3.35 Bya (fossil records from the Strelley Pool Formation) (42, 45, 46), using a uniform distribution. Gamma-distributed root prior (3.95 ± 0.23 Bya), assuming the maximum probability of the root age falling in the midway between the calibrations, was applied to test the effects of root prior distribution (analyses 2 and 8). Geochemical evidence from the Manzimnyama Banded Iron Formation, Fig Tree Group, South Africa, indicates the presence of free oxygen being produced by Cyanobacteria before 3.2 Bya (42, 47), and this was used as a minimum age for total-group of Cyanobacteria. However, as the Banded Iron Formation at 3.2 Bya may have been also formed via anaerobic processes [i.e., UV oxidation (48) and anoxygenic photosynthesis (49, 50)], PhyloBayes analyses without the constraint on Cyanobacteria (analyses 3 and 9) were performed to test the effects of root prior distribution (analyses 2 and 8). The age of the last common ancestor of Rhodophyta was derived from the oldest fossil records of Bangiaceae red algae, which occurred in 1.20 Bya (23). To evaluate whether this assumption is so stringent to overdetermine the estimated divergence times, analyses were performed with reduced sets of calibrations by precluding constraints on Rhodophyta (analyses 4 and 10). Comparisons of estimated confidence intervals suggested that varying root priors or subsampling of calibrations resulted in minimal changes of estimated divergence times (SI Appendix, Fig. S19).

In addition, ribosomal protein phylogeny and SSU rRNA gene phylogeny were converted to ultrametric tree, using TreePL under a penalized likelihood model (52). The rate smoothing parameters were set to 10-based values between 1 and 10,000 with cross-validation procedure and the $\chi^2$ test enabled in TreePL. The full set of temporal constraints (SI Appendix, Fig. S16 and Table S4) was used.

To evaluate the effect of phylogenetic uncertainty on the results, alternative tree topologies reflecting alternative arrangements/bipartitions for taxa of uncertain relationships were generated. Conflicting bipartitions (n = 32) of RAxML ribosomal protein tree that are substantially represented (>40%) in bootstrap replicates were retrieved using RAxML (37) (option -f t, interslice analysis). The alternative minority-bipartition topology was obtained by editing the RAxML tree to reflect all conflicting bipartitions via subtree prune and regraft (analyses 5 and 11). A three-domain tree placing Archaea as a sister group of Eukaryotes was built similarly (analyses 6 and 12). Both alternative topologies were dated with full alignment of ribosomal proteins, using PhyloBayes. Furthermore, we built 100 alternative chronograms using
TreePL (SI Appendix, Fig. S20), based on alternative topologies containing 50% of randomly selected minority bipartitions (Bipartition-Jackknife analysis). Branch length of these alternative topologies were re-estimated by RAxML (option -f e), using full alignment of ribosomal proteins.

Identification of Arsenic Resistance Genes. A hidden Markov model (HMM)-based search was performed to identify arsenic resistance genes in selected genomes. To develop HMM profiles, reference protein sequences were downloaded from Uniprot or National Center for Biotechnology Information (NCBI) (SI Appendix, Table S3) and aligned using MAFFT v7.3100 (53) with linsi option. Sequence alignment was visualized by ClustalX (54), and the ambiguously aligned regions were removed using TrimAl v1.2 (55). HMM profiles were built on curated alignments using hmmbuild in HMMER v3.1b2 package (56).

To collect homologs of arsenic resistance genes, each HMM profile was searched against 786 genomes, using hmmssearch with an E-value cutoff of 0.1. Hit scores were retrieved, and the corresponding sequences were examined for conserved domains, using protein family (PFAM) database (57). With profile searches for ArsC3, ArsB, ArsR, ArsM, and ArsP (SI Appendix, Figs. S22–S27), the retrieved hits were partitioned into two distinct groups: one exhibited significantly higher scores that consist of reference proteins, and another showed a much lower score that included distant homologs. The separation of scoring values permitted us to distinguish these arsenic resistance genes and their remote relatives, and we annotated the sequences showing better scoring values as the target proteins. To determine whether these sequences are truly arsenic resistance proteins, hits from hmmssearch were aligned with MAFFT (multiple sequence alignment based on fast Fourier transform), and phylogenetic trees were constructed using RAxML under the PROTGAMMAAUTO model with 100 nonparametric bootstrap replicates from these trees. Clustering of homologs with significant higher scores formed a moderate- to strong-supported monophyletic clade among the functional characterized proteins (SI Appendix, Figs. S22–S27), which provided evidence that the arsenic resistance proteins were correctly annotated.

In contrast, HMM profiles showed lower ability to distinguish ArsCs from their distant relatives (SI Appendix, Figs. S28–S30), probably because of their short protein lengths and absence of highly conserved domains. Therefore, we identified arsenate reductase genes (ArsC1 and ArsC2) by taking genomic contexts into account. The hmmssearch scoring threshold for each arsenate reductase (ArsC1 and ArsC2) was optimized to include sequences from the phylogenetic clade containing both reference proteins and homologs located within ars operon (SI Appendix, Figs. S28 and S29). Eukaryotic arsenate reductases (Arc2) were determined via a phylogenetic method. Branches within a well-supported clade consisting known Arc2 were selected as putative Arc2 (SI Appendix, Fig. S30).

ArsR homologs were classified into four families on the basis of a reported phylogenetic tree (18). Reference alignment and phylogenetic tree of ArsRs were built as described previously (18). For each ArsR family, homologs extracted by HMM profiles were added to reference alignment using MAFFT (multiple sequence alignment based on the information criterion and Bayesian information criterion). Inference was further applied to fetch protein sequences of arsenic resistance genes in 2,031 organisms included in EggNOG Database (v4.5.1).

Phylogenetic Analysis of Arsenic Resistance Genes. The protein sequences of each arsenic detoxification gene family were aligned with five different methods [MUSCLE (58), ClustalW (54), T-Coffee (59), MAFFT (53) and ProbCons (60)]. Consensus alignment of genes was calculated on the basis of the consistency of output from individual alignment programs using M-Coffee, provided in the T-Coffee package (61). The poorly aligned regions were excised using TrimAl v1.2 (55) with -automated option. The best-fit evolutionary model for each gene family (ArsC3: LG+H+G; ArsB: LG+H+G; ArsC1: WAG+I+G; ArsC2: LG+H+G; Acr2: LG+I+G; ArsR: LG+H+G; Ars: WAG+I+G; ArsM: LG+G+I; ArsP: LG+H+I; ArsR1: LG+I+G; ArsR2: Dayhoff+G+F; ArsR3: LG+I+G; ArsR4: LG+G) was determined by ProtTest3 (62), according to Akaike information criterion and Bayesian information criterion. Inference of maximum likelihood tree was performed under best-fit evolutionary model, using RAxML. Nonparametric bootstrap analysis for each gene tree was conducted under a corresponding evolutionary model with 100 replicates. The pairwise phylogenetic distances were calculated by summing up all of the branches linking two taxa in maximum-likelihood phylogeny. The congruence between gene tree and species tree (ribosomal protein phylogeny) was assessed by scatterplots of pairwise phylogenetic distances calculated from corresponding trees.

Gene Birth Date Inference. Gene birth dates were inferred using a reconciliation algorithm implemented in ecoTEERA (63, 64). An ensemble (n = 10) of nonparametric bootstrapped trees were used as a gene tree set to resolve the uncertainty in deep-branching phylogenies, using amalgamation algorithm (option amalgamated = 1). Fully dated species tree (option dated = 1) was inferred by either ecoTEERA or TreePL, provided that the HGT events among only chronological overlapped lineages. Gene birth was parsed as the earliest split event that led to the gene clade. Posterior estimates of gene age (i.e., median and 95% highest posterior density interval) were calculated over the course of 1,200 reconciliation analyses, using fully dated species trees (n = 100) sampled from each of PhyloBayes MCMC analysis (Tables 1 and 2). To assess the sensitivity of our results to reconciliation algorithms, the gene ages were also estimated using the Analyzer of Gene and Species Trees (AnGST) program (22). AnGST was run with default parameters (event cost: HGT = 3.0, DUP = 2.0, and LOS = 1.0; ultrametric = True) with 10 bootstrapped gene trees. Due to computation limitations, AnGST was performed only on consensus species trees of 12 Bayesian molecular clock analyses (SI Appendix, Fig. S20 and Tables 1 and 2).

Data Availability. Accession numbers of all genomes used in this study are listed in Dataset S1. Protein sequence alignments and maximum-likelihood trees of 13 arsenic resistance genes are available in Dataset S2. Species trees based on alignment of concatenated ribosomal proteins or SSU rRNA are included in Dataset S3.

ACKNOWLEDGMENTS. We thank Lawrence A. David for the insightful discussion about data analysis and results interpretation. We acknowledge the Bundesministerium für Bildung und Forschung (BMBF)-funded German Network for Bioinformatics Infrastructure.de.NBI (031A357B, 031A353A, 031A358A, 031A353B, 031A353A, 031A357C, 031A354A, 031A352B) for providing computational resources. Funding for this project is provided by the National Natural Science Foundation of China (41430856), the Strategic Priority Research Program of Chinese Academy of Sciences (XDB15020302 and XDB15020402), and NIH grants GM55425 and E5023779 to B.P.R.
