



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

Song-Can Chen^{a,b,c,1}, Guo-Xin Sun^{a,1}, Yu Yan^d, Konstantinos T. Konstantinidis^{e,f}, Si-Yu Zhang^e, Ye Deng^a, Xiao-Min Li^{a,c}, Hui-Ling Cui^{a,c}, Florin Musat^b, Denny Popp^g, Barry P. Rosen^{h,2}, and Yong-Guan Zhu^{a,i,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, Huaqiao 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(oids), 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 biogeochemical cycles 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 biogeochemical cycles that emerged as a consequence of the GOE.

arsenic | detoxification | evolution | oxygen | biogeochemistry

One of life's earliest challenges was coping with the toxicity of harmful metal(oids) (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 arsenicals (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 arsenite

Significance

The oxygenation of the atmosphere about 2.4 billion years ago remodeled global cycles of toxic, redox-sensitive metal(oids), including that of arsenic, which must have represented a cataclysm in the history of life. Our understanding of biological adaptations surrounding this key transition remains unexplored. By estimating the timing of genetic systems for arsenic detoxification, we reveal an expansion of enzymes and pathways that accompanied adaptations to the biotoxicity of oxidized arsenic species produced by Great Oxidation Event. These include enzymes originated via convergent evolution and pathways that use oxygen for enzymatic catalysis. Our results illustrate how life thrived under the stress of metal(oid) toxicity and provide insights into environmental biogeochemical cycling and microbial evolution.

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.K., 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 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/).

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.

² To whom correspondence may be addressed. Email: brosen@fiu.edu 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 arsenite 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 ArsI C-Acylase that cleaves the carbon–arsenic bond in methylarsenite (16). Arsenic resistance genes are usually organized in *ars* operons, which are nearly always under control of an ArsR transcriptional repressor. Four different ArsRs, in which each an arsenical 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 incongruousness 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 autocorrelated 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 an oxic 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 the oxic 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 anoxygenic 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

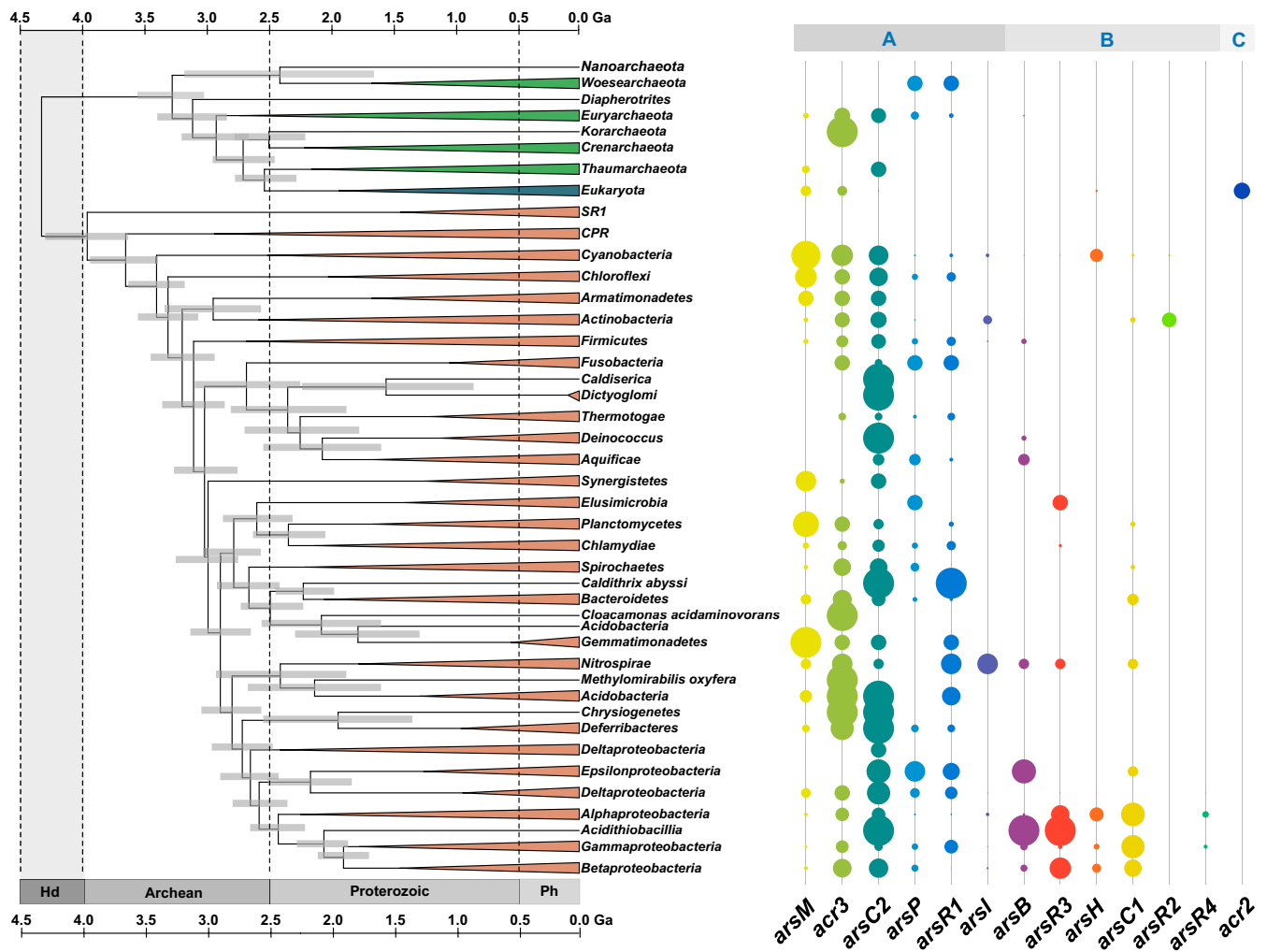


Fig. 1. Phylogenetic distribution of 13 arsenic detoxification genes. (Left) Reference phylogenetic trees of major lineages of Bacteria, Archaea, and Eukaryotes. (Right) Relative abundance of 13 arsenic detoxification genes present within each major lineage. The 13 genes were divided into three sets (A-C) according to their phyletic distribution patterns. The reference phylogeny was reconstructed from concatenate alignment of 16 ribosomal proteins, as previously reported (19). Divergent times and corresponding confidence intervals (95%) were estimated using PhyloBayes (analysis 7; Table 2). Timescale: Hd, Hadean; Ph, Phanerozoic; Ga, billions of years.

analysis on 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 *acr3*, 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 methylarsenite 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 methylarsenite 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, *ArsB*, 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 methylarsenite and

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*	Autocorrelated	Autocorrelated	Autocorrelated	Autocorrelated	Autocorrelated	Autocorrelated
Calibration [†]	Full set	Full set	–Cyanobacteria	–Rhodophyta	Full set	Full set
Root prior [‡]	U(3.35,4.38)	Γ(3.95;0.23)	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	MT	Three-domain tree
Gene age (Gyr) [¶]						
<i>arsM</i>	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)
<i>acr3</i>	2.97 (2.71–3.09)	3.10 (2.77–3.31)	3.18 (2.87–3.38)	2.99 (2.77–3.30)	3.08 (2.78–3.34)	3.00 (2.81–3.23)
<i>arsC2</i>	2.70 (2.34–2.89)	2.74 (2.39–2.95)	2.82 (2.44–2.98)	2.64 (2.41–2.91)	2.74 (2.42–3.01)	2.73 (2.50–3.01)
<i>arsR1</i>	2.79 (2.45–2.97)	2.83 (2.45–3.03)	2.91 (2.55–3.10)	2.73 (2.49–2.98)	2.83 (2.52–3.09)	2.82 (2.61–3.08)
<i>arsP</i>	2.79 (2.45–2.97)	2.83 (2.45–3.03)	2.91 (2.55–3.10)	2.73 (2.49–2.98)	2.83 (2.52–3.09)	2.82 (2.61–3.08)
<i>arsB</i>	2.07 (1.57–2.36)	2.10 (1.57–2.41)	2.16 (1.58–2.46)	2.03 (1.61–2.39)	2.10 (1.60–2.47)	2.10 (1.73–2.47)
<i>arsI</i>	2.26 (1.78–2.47)	2.27 (1.94–2.54)	2.36 (1.99–2.61)	2.19 (1.85–2.51)	2.26 (1.90–2.62)	2.28 (1.90–2.58)
<i>arsH</i>	1.91 (1.79–2.04)	1.92 (1.79–2.03)	1.93 (1.80–2.05)	1.90 (1.79–2.04)	1.82 (1.59–2.00)	1.91 (1.80–2.06)
<i>arsR2</i>	1.80 (1.70–1.91)	1.81 (1.70–1.91)	1.81 (1.70–1.92)	1.79 (1.70–1.91)	1.86 (1.50–2.14)	1.80 (1.70–1.93)
<i>arsR3</i>	1.71 (1.48–1.84)	1.71 (1.51–1.90)	1.77 (1.55–1.93)	1.63 (1.49–1.86)	1.91 (1.69–2.09)	1.71 (1.54–1.94)
<i>arsC1</i>	1.72 (1.48–1.85)	1.73 (1.53–1.94)	1.79 (1.57–1.92)	1.64 (1.51–1.89)	1.74 (1.54–1.97)	1.72 (1.58–1.95)
<i>acr2</i>	0.97 (0.80–1.11)	0.99 (0.80–1.16)	1.02 (0.78–1.11)	0.95 (0.79–1.13)	1.00 (0.85–1.16)	1.17 (0.99–1.40)
<i>arsR4</i>	1.14 (0.97–1.26)	1.15 (0.98–1.31)	1.18 (1.02–1.28)	1.08 (0.97–1.26)	1.00 (0.81–1.16)	1.14 (1.02–1.32)

*Autocorrelated, 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 where archaea and eukaryotes are sister group.

[¶]Median age estimates of gene birth nodes, with 95% confidence intervals in parentheses; Gyr, billion years.

dimethylarsenite, would be oxidized nonenzymatically by dioxygen into relatively innocuous methylarsenate and dimethylarsenate. However, the influence of dioxygen did not stop here. Our results further suggest that two new obligate oxygen-dependent methylarsenite resistance enzymes, ArsH and ArsI, arose during or after the GOE. Concurrent with the evolution of these new oxygen-dependent methylarsenite 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 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	–Cyanobacteria	–Rhodophyta	Full set	Full set
Root prior [‡]	U(3.35,4.38)	Γ(3.95;0.23)	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	MT	Three-domain tree
Gene age (Gyr) [¶]						
<i>arsM</i>	3.40 (3.23–3.61)	3.37 (3.23–3.72)	3.44 (3.03–3.68)	3.40 (3.24–3.73)	3.40 (3.24–3.72)	3.45 (3.24–3.76)
<i>acr3</i>	2.79 (2.55–2.96)	2.77 (2.59–3.05)	2.86 (2.51–3.06)	2.79 (2.56–3.04)	2.78 (2.60–3.04)	2.81 (2.59–3.04)
<i>arsC2</i>	2.39 (2.03–2.68)	2.39 (2.03–2.76)	2.45 (2.04–2.72)	2.38 (2.01–2.76)	2.40 (2.06–2.76)	2.39 (2.07–2.74)
<i>arsR1</i>	2.47 (2.12–2.75)	2.46 (2.09–2.86)	2.53 (2.12–2.79)	2.46 (2.06–2.84)	2.47 (2.12–2.83)	2.46 (2.17–2.82)
<i>arsP</i>	2.47 (2.12–2.75)	2.46 (2.09–2.86)	2.53 (2.12–2.79)	2.46 (2.06–2.84)	2.47 (2.12–2.83)	2.46 (2.17–2.82)
<i>arsB</i>	1.99 (1.70–2.21)	1.99 (1.73–2.34)	2.04 (1.71–2.30)	1.98 (1.68–2.30)	2.00 (1.73–2.30)	2.01 (1.75–2.28)
<i>arsI</i>	1.36 (0.84–2.02)	1.40 (0.84–2.04)	1.42 (0.79–2.01)	1.37 (0.78–2.00)	1.39 (0.82–2.15)	1.38 (0.79–2.05)
<i>arsH</i>	1.70 (1.61–1.81)	1.61 (1.43–1.82)	1.70 (1.61–1.82)	1.70 (1.61–1.84)	1.58 (1.41–1.78)	1.64 (1.43–1.83)
<i>arsR2</i>	1.63 (1.57–1.73)	1.63 (1.56–1.76)	1.63 (1.56–1.74)	1.63 (1.57–1.77)	1.63 (1.56–1.73)	1.63 (1.57–1.75)
<i>arsR3</i>	1.53 (1.33–1.70)	1.53 (1.34–1.74)	1.57 (1.31–1.73)	1.53 (1.37–1.73)	1.69 (1.53–1.88)	1.56 (1.36–1.77)
<i>arsC1</i>	1.31 (0.93–1.58)	1.29 (1.02–1.59)	1.35 (0.95–1.58)	1.31 (1.00–1.59)	1.30 (1.01–1.58)	1.34 (1.02–1.62)
<i>acr2</i>	1.11 (0.92–1.29)	1.11 (0.96–1.31)	1.17 (0.95–1.32)	1.09 (0.93–1.28)	1.10 (0.94–1.28)	1.18 (1.01–1.38)
<i>arsR4</i>	1.02 (0.89–1.14)	1.02 (0.92–1.19)	1.05 (0.89–1.17)	1.02 (0.90–1.19)	0.79 (0.65–0.94)	1.05 (0.93–1.19)

*Autocorrelated, 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 where archaea and eukaryotes are sister group.

[¶]Median age estimates of gene birth nodes, with 95% confidence intervals in parentheses; Gyr, billion years.

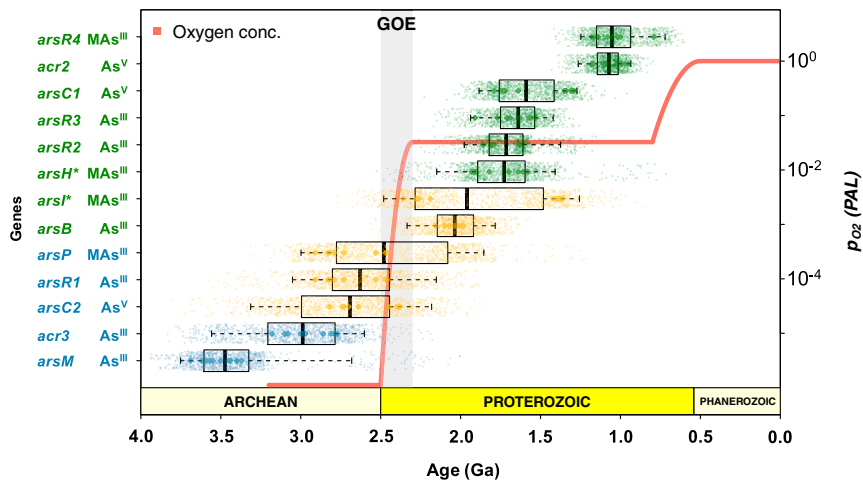


Fig. 2. Gene birth date for each of 13 arsenic detoxification genes. Gene ages were derived from reconciliation results (cycle), using fully dated species trees ($n = 1200$) sampled from 12 PhyloBayes analyses. The median age estimates under each analytical scenario (Tables 1 and 2) were shown as diamond. The uncertainties associated with the results from all PhyloBayes analyses were integrated as 95% composite confidence intervals (whisker of the boxplot). Age estimates of genes evolved before, around, and after GOE were shown as blue, yellow, and green, respectively. Atmospheric oxygen content throughout Earth's history was overlaid on the gene's age (red line) (9). Right y axis, pO_2 , relative to the present atmospheric level (PAL); left y axis, gene names. Genes found in both anaerobes and aerobes, or only in aerobes were denoted as blue and green, respectively (Fig. 3). Oxygen-dependent genes (*arsI* and *arsH*) were indicated by star. As^{III}, As^V, and MAs^{III} were used to delineate genes acting on inorganic arsenite, arsenate, or methylarsenite, respectively. Ga, billions of years.

of arsenic during the Archean 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 bootstrapping 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

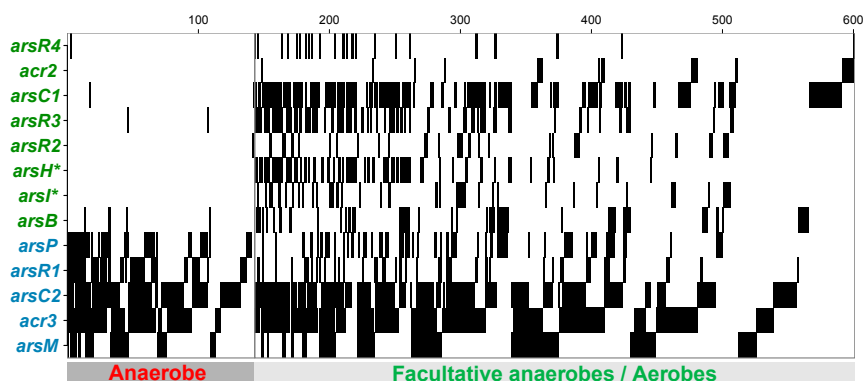


Fig. 3. Distribution of 13 arsenic detoxification genes among strict anaerobes and aerobes. Species were classified either as aerobes (including facultative anaerobes) or anaerobes based on their capability to use oxygen as a terminal electron acceptor. Each black tick indicated the presence of the corresponding gene in a taxon. Genes evolved before or at beginning of GOE were denoted as blue, and those after as green. Oxygen-dependent genes (*arsI* and *arsH*) were indicated with the star symbol.

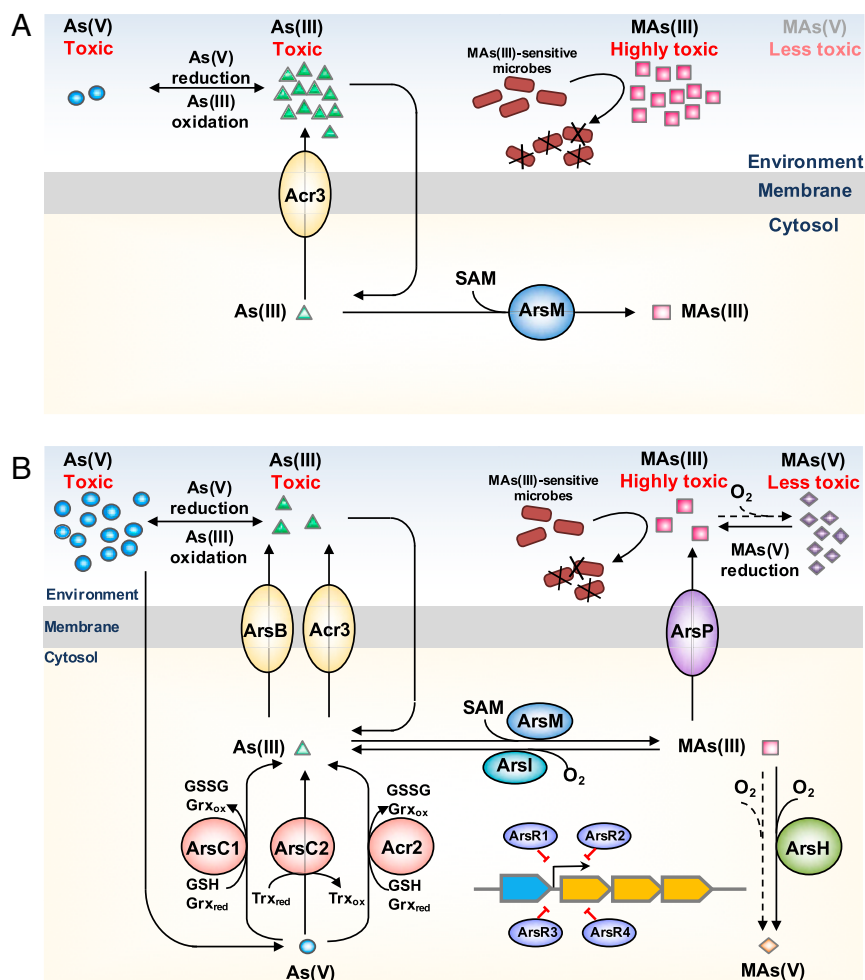


Fig. 4. Arsenic resistance systems before (A) and after (B) the GOE. As(III), arsenite; As(V), arsenate; MAs(III), trivalent methylarsenite; MAs(V), pentavalent methylarsenate; SAM, S-adenosylmethionine; GSH, reduced glutathione; GSSG, oxidized glutathione; Grx_{red}: reduced glutaredoxin; Grx_{ox}, oxidized glutaredoxin; Trx_{red}, reduced thioredoxin; Trx_{ox}, oxidized thioredoxin.

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 how inclusion of this constraint impacts the results. The time constraint on Rhodophyta was derived from the oldest fossil records of Bangiale red algae, which occurred in 1.20 Bya Hunting Formation (51). 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).

For all molecular clock analyses, two independent PhyloBayes Markov chain Monte Carlo (MCMC) chains were run in parallel up to 1 mo (~60,000 model cycles). The convergence of MCMC chains was checked by comparing the posterior distributions of independent runs, using *tracemcomp* program implemented in PhyloBayes (effective sizes >100, and maximum discrepancy between chains <0.3). A state of the MCMC chain was sampled every 20 cycles after 20% initial generations discarded as burn-in. All PhyloBayes analyses were also run under the prior conditions by removing the sequence data, to verify that the estimated divergence time is not solely driven by fossil records (SI Appendix, Fig. S18).

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 χ^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, internode certainty 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 v7.310 (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 *hmmsearch* 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 Acr3, ArsB, ArsH, Arsl, 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 from 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 *hmmsearch* 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 bootstraps. The results from these tree-building trails indicated that sequences 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 prokaryotic arsenate reductase genes (ArsC1 and ArsC2) by taking genomic contexts into account. The *hmmsearch* 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 (Acr2) were determined via a phylogenetic method. Branches within a well-supported clade consisting known Acr2 were selected as putative Acr2 (*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 (–add and –keeplength) and assigned to a reference tree with evolutionary placement algorithm in RAxML. Sequences that were placed within the corresponding clade of the reference tree were identified as ArsR (*SI Appendix, Fig. S31*).

Sequences retrieved here were further screened for presence of key catalytic residues (*SI Appendix, Table S1*). Homologs passed through these criteria were regarded as functional orthologs involved in arsenic resistance, which were used for subsequent analysis. The same identification pipeline

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 *-automated1* option. The best-fit evolutionary model for each gene family (Acr3: LG+I+G; ArsB: LG+I+G; ArsC1: WAG+I+G; ArsC2: LG+I+G; Acr2: LG+I+G; ArsH: LG+I+G; Arsl: WAG+I+G; ArsM: LG+I+G; ArsP: LG+I+G; 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 taxons 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 ecceTERA (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 amalgamate = 1). Fully dated species tree (option dated = 2) reconstructed by either PhyloBayes or TreePL was provided to restrict 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 (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B) for providing computational resources. Funding for this project is provided by the National Natural Science Foundation of China (41430858), the Strategic Priority Research Program of Chinese Academy of Sciences (XDB15020302 and XDB15020402), and NIH grants GM55425 and ES023779 to B.P.R.

1. T. Clarkson, Health effects of metals: A role for evolution? *Environ. Health Perspect.* **103** (suppl. 1), 9–12 (1995).
2. Y. G. Zhu, M. Yoshinaga, F. J. Zhao, B. P. Rosen, Earth abides arsenic biotransformations. *Annu. Rev. Earth Planet. Sci.* **42**, 443–467 (2014).
3. E. C. Fru *et al.*, Arsenic stress after the proterozoic glaciations. *Sci. Rep.* **5**, 17789 (2015).
4. E. Lebrun *et al.*, Arsenite oxidase, an ancient bioenergetic enzyme. *Mol. Biol. Evol.* **20**, 686–693 (2003).
5. T. R. Kulp *et al.*, Arsenic(III) fuels anoxygenic photosynthesis in hot spring biofilms from Mono Lake, California. *Science* **321**, 967–970 (2008).
6. R. S. Oremland, C. W. Saltikov, F. Wolfe-Simon, J. F. Stolz, Arsenic in the evolution of earth and extraterrestrial ecosystems. *Geomicrobiol. J.* **26**, 522–536 (2009).
7. R. S. Oremland, J. F. Stolz, The ecology of arsenic. *Science* **300**, 939–944 (2003).
8. S. Duval, A. L. Ducluzeau, W. Nitschke, B. Schoep-Cohenet, Enzyme phylogenies as markers for the oxidation state of the environment: The case of respiratory arsenate reductase and related enzymes. *BMC Evol. Biol.* **8**, 206 (2008).
9. T. W. Lyons, C. T. Reinhard, N. J. Planavsky, The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307–315 (2014).
10. E. C. Fru *et al.*, The rise of oxygen-driven arsenic cycling at ca. 2.48 Ga. *Geology* **47**, 243–246 (2019).
11. B. P. Rosen, Biochemistry of arsenic detoxification. *FEBS Lett.* **529**, 86–92 (2002).
12. R. Mukhopadhyay, B. P. Rosen, Arsenate reductases in prokaryotes and eukaryotes. *Environ. Health Perspect.* **110** (suppl. 5), 745–748 (2002).
13. J. Chen, M. Madegowda, H. Bhattacharjee, B. P. Rosen, ArsP: A methylarsenite efflux permease. *Mol. Microbiol.* **98**, 625–635 (2015).
14. J. Qin *et al.*, Arsenic detoxification and evolution of trimethylarsine gas by a microbial arsenite S-adenosylmethionine methyltransferase. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2075–2080 (2006).
15. J. Chen, H. Bhattacharjee, B. P. Rosen, ArsH is an organoarsenic oxidase that confers resistance to trivalent forms of the herbicide monosodium methylarsenate and the poultry growth promoter roxarsone. *Mol. Microbiol.* **96**, 1042–1052 (2015).
16. M. Yoshinaga, B. P. Rosen, A C-As lyase for degradation of environmental organo-arsenic herbicides and animal husbandry growth promoters. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7701–7706 (2014).
17. J. Qin *et al.*, Convergent evolution of a new arsenic binding site in the ArsR/SmtB family of metalloregulators. *J. Biol. Chem.* **282**, 34346–34355 (2007).

18. J. Chen, V. S. Nadar, B. P. Rosen, A novel MAs(III)-selective ArsR transcriptional repressor. *Mol. Microbiol.* **106**, 469–478 (2017).
19. L. A. Hug *et al.*, A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
20. S.-C. Chen *et al.*, Recurrent horizontal transfer of arsenite methyltransferase genes facilitated adaptation of life to arsenic. *Sci. Rep.* **7**, 7741 (2017).
21. E. Pennisi, Algae suggest eukaryotes get many gifts of bacteria DNA. *Science* **363**, 439–440 (2019).
22. L. A. David, E. J. Alm, Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469**, 93–96 (2011).
23. S. E. Hoefl *et al.*, *Alkalilimnicola ehrlichii* sp. nov., a novel, arsenite-oxidizing haloalkaliphilic gammaproteobacterium capable of chemoautotrophic or heterotrophic growth with nitrate or oxygen as the electron acceptor. *Int. J. Syst. Evol. Microbiol.* **57**, 504–512 (2007).
24. M. Fakraee, S. A. Crowe, S. Katsev, Sedimentary sulfur isotopes and Neoproterozoic ocean oxygenation. *Sci. Adv.* **4**, e1701835 (2018).
25. A. D. Anbar *et al.*, A whiff of oxygen before the great oxidation event? *Science* **317**, 1903–1906 (2007).
26. S. A. Crowe *et al.*, Atmospheric oxygenation three billion years ago. *Nature* **501**, 535–538 (2013).
27. B. Eickmann *et al.*, Isotopic evidence for oxygenated Mesoproterozoic shallow oceans. *Nat. Geosci.* **11**, 133–138 (2018).
28. N. J. Planavsky *et al.*, Evidence for oxygenic photosynthesis half a billion years before the Great Oxidation Event. *Nat. Geosci.* **7**, 283–286 (2014).
29. S. Silver, L. T. Phung, Bacterial heavy metal resistance: New surprises. *Annu. Rev. Microbiol.* **50**, 753–789 (1996).
30. D. H. Nies, Microbial heavy-metal resistance. *Appl. Microbiol. Biotechnol.* **51**, 730–750 (1999).
31. R. Riding, P. Fralick, L. Y. Liang, Identification of an Archaean marine oxygen oasis. *Precambrian Res.* **251**, 232–237 (2014).
32. J. Chen, M. Yoshinaga, B. P. Rosen, The antibiotic action of methylarsenite is an emergent property of microbial communities. *Mol. Microbiol.* **111**, 487–494 (2019).
33. J. Li, S. S. Pawitwar, B. P. Rosen, The organoarsenical biocycle and the primordial antibiotic methylarsenite. *Metallomics* **8**, 1047–1055 (2016).
34. M. C. Sforna *et al.*, Evidence for arsenic metabolism and cycling by microorganisms 2.7 billion years ago. *Nat. Geosci.* **7**, 811–815 (2014).
35. C. L. Dupont, S. Yang, B. Palenik, P. E. Bourne, Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 17822–17827 (2006).
36. C. L. Dupont, A. Butcher, R. E. Valas, P. E. Bourne, G. Caetano-Anollés, History of biological metal utilization inferred through phylogenomic analysis of protein structures. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 10567–10572 (2010).
37. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
38. E. Pruesse, J. Peplies, F. O. Glöckner, SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).
39. S. Mukherjee *et al.*, Genomes OnLine database (GOLD) v.6: Data updates and feature enhancements. *Nucleic Acids Res.* **45**, D446–D456 (2017).
40. N. Lartillot, T. Lepage, S. Blanquart, PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).
41. T. Lepage, D. Bryant, H. Philippe, N. Lartillot, A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.* **24**, 2669–2680 (2007).
42. H. C. Betts *et al.*, Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat. Ecol. Evol.* **2**, 1556–1562 (2018).
43. J. W. Valley *et al.*, Hadean age for a post-magma-ocean zircon confirmed by atom-probe tomography. *Nat. Geosci.* **7**, 219–223 (2014).
44. S. A. Wilde, J. W. Valley, W. H. Peck, C. M. Graham, Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago. *Nature* **409**, 175–178 (2001).
45. A. Hickman, *Regional Review of the 3426–3350 Ma Strelley Pool Formation* (Pilbara Craton, Western Australia, 2008).
46. D. Wacey, Stromatolites in the approximately 3400 Ma Strelley Pool Formation, Western Australia: Examining biogenicity from the macro- to the nano-scale. *Astrobiology* **10**, 381–395 (2010).
47. A. M. Satkoski, N. J. Beukes, W. Q. Li, B. L. Beard, C. M. Johnson, A redox-stratified ocean 3.2 billion years ago. *Earth Planet. Sci. Lett.* **430**, 43–53 (2015).
48. A. G. Cairnsmith, Precambrian solution photochemistry, inverse segregation, and banded iron formations. *Nature* **276**, 807–808 (1978).
49. K. O. Konhauser *et al.*, Could bacteria have formed the Precambrian banded iron formations? *Geology* **30**, 1079–1082 (2002).
50. S. A. Crowe *et al.*, Photoferrotrophs thrive in an Archaean ocean analogue. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 15938–15943 (2008).
51. N. J. Butterfield, *Bangiomorpha pubescens* n. gen., n. sp.: Implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology* **26**, 386–404 (2000).
52. S. A. Smith, B. C. O'Meara, treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**, 2689–2690 (2012).
53. K. D. Yamada, K. Tomii, K. Katoh, Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. *Bioinformatics* **32**, 3246–3251 (2016).
54. J. D. Thompson, T. J. Gibson, D. G. Higgins, Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinf.* **2.3.1–2.3.22** (2002).
55. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
56. L. S. Johnson, S. R. Eddy, E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinf.* **11**, 431 (2010).
57. A. Bateman *et al.*, The Pfam protein families database. *Nucleic Acids Res.* **30**, 276–280 (2002).
58. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
59. C. Notredame, D. G. Higgins, J. Heringa, T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
60. C. B. Do, M. S. P. Mahabhashyam, M. Brudno, S. Batzoglou, ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15**, 330–340 (2005).
61. P. Di Tommaso *et al.*, T-Coffee: A web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* **39**, W13–W17 (2011).
62. D. Darrriba, G. L. Taboada, R. Doallo, D. Posada, ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).
63. C. Scornavacca, E. Jacox, G. J. Szöllösi, Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics* **31**, 841–848 (2015).
64. E. Jacox, C. Chauve, G. J. Szöllösi, Y. Ponty, C. Scornavacca, ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics* **32**, 2056–2058 (2016).