

THE DISTRIBUTION AND DYNAMICS OF  
RESISTANCE GENES IN SOIL MICROBIOMES

By

Taylor Katherine Dunivin

A DISSERTATION

Submitted to  
Michigan State University  
in partial fulfillment of the requirements  
for the degree of

Microbiology – Environmental Toxicology – Doctor of Philosophy

2019

## ABSTRACT

### THE DISTRIBUTION AND DYNAMICS OF RESISTANCE GENES IN SOIL MICROBIOMES

By

Taylor Katherine Dunivin

The soil microbiome harbors immense microbial biodiversity that encodes important functions of interest to public health. These include functional genes that encode resistance to antibiotics and arsenic. In the case of antibiotic resistance, transfer from environmental strains to pathogens is a public health risk, and arsenic resistance and metabolisms are important for bioremediation as they impact the fate of arsenic in the environment. While these resistance genes are well-characterized *in vitro*, the full scope of their environmental distribution, diversity, and interspecies transfer is unknown. A better understanding of the diversity and distribution of resistance genes would provide insights into the potential for mitigation of public health problems such as arsenic contamination and antibiotic resistance.

The work in this dissertation used a combination of cultivation-dependent and -independent techniques to better understand the dynamics and distributions of antibiotic and arsenic resistance genes in the environment. The influence of a disturbance on microbial antibiotic resistance and arsenic related genes was investigated by examining soils overlaying an underground coal mine fire in Centralia, PA. Additionally, soil meta-analyses were used to determine broader distributions patterns of these genes. These data and methods not only provide insights into the distributions and dynamics of antibiotic resistance and arsenic related genes in soil microbiomes but also provide a framework for future studies of other functional genes.



Copyright by  
TAYLOR KATHERINE DUNIVIN  
2019

This thesis is dedicated to my sister, Hailey.  
Your love of nature inspired mine.

## ACKNOWLEDGEMENTS

This body of work would not have been possible without the support from many people. I am grateful for the scientific, professional, and personal guidance that I received throughout my graduate career.

First, I would like to thank my PhD mentor Dr. Ashley Shade. As a new professor at Michigan State University, Ashley trusted me to design a project before ever stepping into the lab. This pushed me to grow as a researcher early on in my graduate career, and I am forever grateful for it. Despite my non-traditional career goals, Ashley took me on as a student and embraced my interest in science policy. She supported my decision to join the Broadening Experiences in Scientific Training (BEST) program, introduced me to the executive director of Science Debate, and allowed me to travel to Washington DC for three months as a Mirzayan fellow at the National Academies of Sciences, Engineering, and Medicine. This time away from the bench (or the terminal) is unheard of in some labs, but it has proved to be invaluable for my professional development.

I would also like to thank the community of scientific support from Michigan State University. Thank you to the members of my guidance committee: Dr. Robert Hausinger, Dr. James Tiedje, and Dr. Hui Li. I am thankful for their thoughtful questions, and their high standards pushed me to become the scientist I am today. Thank you to the past and present members of the Shade Lab: Alan Bowsher, Fina Binarti, Jackson Sorensen, John Chodkowski, John Guitar, Keara Grady, Nejc Stopnisek, Pat Kearns, Sang-Hoon Lee, Siobhan Cusack, and Wendy Smythe. Thank you to Justine Miller and Susanna Yeh, two undergraduate students I had the honor of mentoring.

I am thankful for the professional development guidance I received at Michigan State University. I want to thank my the Department of Microbiology and Molecular Genetics and the Institute for Integrative Toxicology for their professional development activities over the years. I want to thank Poorna Viswanathan for guiding me through my first teaching experience and continuing to mentor me well beyond one semester. I want to thank the BEST program, specifically Julie Rojewski, for providing me with opportunities to explore careers outside of academia. I want to thank Sheril Kirshenbaum for being my science policy mentor and allowing me to participate so much in Science Debate's activities. I want to thank Scott Pohl for mentoring me on science communication and encouraging all of my story ideas.

Finally, I am grateful to my family and friends for their ceaseless support and sympathetic ears throughout all five years of my graduate career. I especially want to thank my parents, Kevin Dunivin and Carolyn Moore, as well as my sister, Hailey Dunivin, for supporting me in my work and for distracting me with many vacations together. I also want to thank my grandparents, John and Kathleen Moore, for always checking on how my work was going. I want to thank Mark Borowski for always understanding and encouraging me. I want to thank all of the friends who supported me, especially Lavidia Brooks for hundreds of subway lunches and countless laughs. Thank you all for the support and love.

## TABLE OF CONTENTS

LIST OF TABLES .....	iv
LIST OF FIGURES .....	v
KEY TO ABBREVIATIONS.....	viii
KEY TO SYMBOLS .....	xi
CHAPTER 1 : Introduction .....	1
Overview .....	1
Arsenic tolerance, resistance, and metabolism .....	1
Arsenite efflux pumps and accessory functions (acr3, arsB, arsD) .....	2
Cytoplasmic arsenate reductases (arsC (trx), arsC (grx)) .....	3
Arsenite methylation (arsM) .....	4
Bioenergetic arsenic related genes (aioA, arxA, arrA) .....	4
Arsenic biogeochemical cycling .....	5
Arsenic exposure and arsenic related genes.....	7
Abiotic gradients and arsenic related genes.....	9
Antibiotic resistance.....	10
Environmental Disturbance and Centralia, PA .....	11
Summary and Aims.....	12
CHAPTER 2 : Taxonomically-linked growth phenotypes during arsenic stress among arsenic resistant bacteria isolated from soils overlying the Centralia coal seam fire.....	14
Abstract .....	15
Introduction.....	16
Materials and Methods.....	19
Soil collection and site description .....	19
Cultivation-dependent soil bacterial community growth.....	19
Isolation of arsenic resistant bacteria .....	20
Morphological characterization and temperature maxima .....	20
DNA extraction and quantification .....	21
Endpoint PCR and amplicon sequencing.....	21
Phylogenetic analysis.....	23
Cultivation-independent 16S rRNA amplicon sequencing and analysis .....	24
Arsenic transformation capabilities .....	25
Minimum inhibitory concentrations (MICs).....	25
Results.....	27
Taxonomic diversity and composition of arsenic resistant isolates.....	27
Genetic characterization of arsenic resistance .....	27

Arsenic transformation.....	32
Incongruent phylogenies of arsenic resistance and 16S rRNA genes.....	33
MICs and growth phenotypes in arsenic.....	33
Discussion.....	40
 CHAPTER 3 : Community structure explains antibiotic resistance gene dynamics over a temperature gradient in soil .....	44
Abstract.....	45
Introduction.....	46
Materials and methods.....	49
Reference Database construction.....	49
Sample collection, sequencing, and quality control.....	49
Gene targeted assembly and quality control.....	50
Ecological analyses.....	51
Resistance gene comparison.....	51
Reproducibility, code, and data .....	52
Results and Discussion .....	53
Soil samples and gene targeted assembly .....	53
Detected ARGs and changes in their abundance with temperature.....	54
Diversity of ARGs .....	58
ARG distribution and sequence-specific biogeography .....	60
ARG Compositional shifts.....	65
Conclusions.....	68
 CHAPTER 4 : RefSoil+: A reference database for genes and traits of soil plasmids .....	69
Abstract.....	70
Introduction.....	71
Materials and Methods.....	74
Data availability .....	74
RefSoil plasmid database generation.....	74
Accessing RefSeq genomes and plasmids .....	74
Plasmid characterization .....	75
Antibiotic resistance gene detection .....	76
Results and Discussion .....	77
Plasmid characterization .....	77
ARGs on soil plasmids .....	81
RefSoil+ applications.....	88
 CHAPTER 5 : A global survey of arsenic related genes in soil microbiomes .....	90
Abstract.....	91
Introduction.....	92
Materials and Methods.....	95
Gene Selection and Functional Gene (FunGene) Database Construction .....	95
Arsenic related genes in cultivable soil microorganisms.....	95
Reference Database Construction.....	96
Sample collection and preparation.....	97

DNA extraction and metagenome sequencing.....	97
Public soil metagenome acquisition.....	98
Soil metagenome processing and gene targeted assembly .....	98
Soil metagenome comparison .....	99
Results.....	100
A bioinformatic toolkit for detecting and quantifying arsenic related genes.....	100
Phylogenetic distributions and genomic locations of arsenic related genes.....	100
Phylogenetic diversity of arsenic related genes: insights into vertical and horizontal transfer .....	104
Cultivation bias and environmental distributions of arsenic related genes.....	108
Arsenic related gene endemism .....	116
Discussion.....	117
A bioinformatic toolkit for detecting and quantifying arsenic related genes.....	117
Phylogenetic diversity and distribution of arsenic related genes.....	118
Cultivation bias and environmental distributions of arsenic related genes.....	120
Arsenic related gene endemism .....	121
Conclusions.....	122
CHAPTER 6 : Conclusions and future directions .....	123
Summary .....	124
Future Directions .....	125
APPENDICES .....	127
APPENDIX A: Supplementary Tables .....	128
APPENDIX B: Supplementary Figures.....	152
REFERENCES .....	179

## LIST OF TABLES

Table 2.1. Relative abundance of isolate 16S rRNA gene sequences from our amplicon survey of the same soil. ....	28
Table 3.1. Resistance genes tested in this study. ....	55
Appendix A Table 1. Soil geochemical data. ....	129
Appendix A Table 2. Degenerate primers used for end point PCR. ....	130
Appendix A Table 3. Isolates with short <i>arsC</i> sequences (< 200 bp). ....	131
Appendix A Table 4. Phenotypes of arsenic resistant isolates. ....	132
Appendix A Table 5. Comparison of arsenic resistance gene sequences with NCBI references. ....	133
Appendix A Table 6. Parameters for reference gene (FunGene) database construction and gene-targeted assembly for each protein of interest. ....	134
Appendix A Table 7. Sequencing depth and Nonpareil-estimated coverage of <i>Centrulia</i> metagenomes. ....	135
Appendix A Table 8. Sample site characteristics and measured soil geochemical data. ....	136
Appendix A Table 9. Spearman's rank correlation between relative abundance of ARGs and soil temperature. ....	137
Appendix A Table 10. Correlations between ARG phylum and Proteobacteria class normalized abundances and soil temperature. ....	138
Appendix A Table 11. Correlations between ARG phylum and Proteobacteria class relative abundances and soil temperature. ....	139
Appendix A Table 12. ResFams HMMs and antibiotic classifications. ....	141
Appendix A Table 13. Summary of reference arsenic resistance and metabolism protein sequences from FunGene databases. ....	146
Appendix A Table 14. Available metadata and accession numbers for soil metagenomes used in this study. ....	147
Appendix A Table 15. Phylum-level summary of arsenic related genes in RefSoil+ chromosomes and plasmids. ....	150
Appendix A Table 16. Summary of endemic arsenic related gene sequences. ....	151



## LIST OF FIGURES

Figure 1.1. Arsenic biogeochemical cycle.....	6
Figure 2.1. Phylogenetic tree of 16S rRNA sequences from <i>Centralia</i> arsenic resistant isolates.	29
Figure 2.2. As resistance genotypes and phenotypes of isolated bacterial strains.....	31
Figure 2.3. Comparison of arsenic resistance gene sequences and 16S rRNA gene sequences from arsenic resistant isolates. ....	35
Figure 2.4. Growth phenotypes of isolates in increasing concentrations of arsenic. ....	38
Figure 3.1. Negative correlations between normalized abundance of ARGs and soil temperature. ....	57
Figure 3.2. Observed richness (AB) and evenness (CD) of <i>rplB</i> (AC) and ARG (BD) along the <i>Centralia</i> temperature gradient.....	59
Figure 3.3. Presence of ARG sequences in <i>Centralia</i> metagenomes. ....	61
Figure 3.4. Normalized abundance of ARG sequences in <i>Centralia</i> metagenomes. ....	63
Figure 3.5. Relative abundance of taxonomically similar ARGs. ....	66
Figure 4.1. Summary of RefSoil plasmids.....	78
Figure 4.2. Plasmid size distributions. ....	79
Figure 4.3. Relationship between plasmid size and genome size.....	82
Figure 4.4. Distribution of ARGs in RefSoil genomes and plasmids.....	84
Figure 4.5. Proportion of ARGs on genomes and plasmids in RefSoil+ and RefSeq databases..	86
Figure 5.1. Arsenic resistance and metabolism gene toolkit schematic. ....	101
Figure 5.2. Arsenic resistance and metabolism genes in RefSoil+ organisms. ....	102
Figure 5.3. Phylogeny of arsenite efflux pumps in RefSoil+ organisms.....	105
Figure 5.4. Phylogeny of cytoplasmic arsenate reductases in RefSoil+ organisms. ....	107
Figure 5.5. Phylogeny of ArsM in RefSoil+ organisms. ....	109
Figure 5.6. Phylogeny of AioA, ArrA, and ArxA in RefSoil+ organisms. ....	110

Figure 5.7. Comparison of arsenic resistance and metabolism gene abundance between cultivation dependent and cultivation independent methods. ....	111
Figure 5.8. Arsenic resistance and metabolism gene biogeography. ....	114
Appendix B Figure 1. Average OD <sub>590</sub> over 72 h in TSB50 with increasing concentrations of arsenate A) or arsenite B). ....	153
Appendix B Figure 2. Lag time in TSB50 with increasing concentrations of arsenate and arsenite normalized to growth in TSB50 without arsenic. ....	155
Appendix B Figure 3. Growth rate in TSB50 with increasing concentrations of arsenate and arsenite normalized to growth in TSB50 without arsenic. ....	157
Appendix B Figure 4. Maximum OD <sub>590</sub> in TSB50 with increasing concentrations of arsenate and arsenite normalized to growth in TSB50 without arsenic. ....	159
Appendix B Figure 5. Dendrogram of isolate growth phenotypes in arsenic. ....	161
Appendix B Figure 6. Sampling strategy along the Centralia temperature gradient. ....	162
Appendix B Figure 7. Comparison of community structure assessed using two different methods. ....	163
Appendix B Figure 8. Pair-wise Spearman's correlations of normalized ARG abundances in Centralia. ....	164
Appendix B Figure 9. Relationship between normalized abundance of ARGs and soil temperature. ....	165
Appendix B Figure 10. Beta diversity of Centralia microbial communities with <i>rplB</i> and ARGs. ....	167
Appendix B Figure 11. Relationship between plasmid number and chromosome size. ....	168
Appendix B Figure 12. Proportion of ARGs on genomes and plasmids in RefSoil+ and RefSeq databases normalized to base pairs. ....	169
Appendix B Figure 13. Proportion of ARGs by classification in RefSoil and RefSeq databases. ....	170
Appendix B Figure 14. Phylogeny of Acr3 in RefSoil+ organisms. ....	172
Appendix B Figure 15. Phylogeny of ArsB in RefSoil+ organisms. ....	173
Appendix B Figure 16. Phylogeny of ArsC (trx) in RefSoil+ organisms. ....	174
Appendix B Figure 17. Phylogeny of ArsC (grx) in RefSoil+ organisms. ....	175
Appendix B Figure 18. Phylogeny of ArsM in RefSoil+ organisms. ....	176

Appendix B Figure 19. Histogram of arsenic related gene copy numbers in RefSoil+ organisms. .....	177
Appendix B Figure 20. Phylum-level community structure of soil metagenomes.....	178

## KEY TO ABBREVIATIONS

AgNO<sub>3</sub> – silver nitrate

AIC – Akaike information criterion

AMD – acid mine drainage

ARG – antibiotic resistance gene

As – arsenic

As<sup>3+</sup> – trivalent arsenic; arsenite

As<sup>5+</sup> – pentavalent arsenic; arsenate

BAM – binary alignment map

BLAST – basic local alignment search tool

bp – base pair

CA – California

CO – carbon monoxide

CO<sub>2</sub> – carbon dioxide

DNA – deoxyribonucleic acid

dPBS – Dulbecco's phosphate-buffered saline

EDTA - ethylenediaminetriacetic acid

EDTA•Na<sub>2</sub> – disodium ethylenediaminetriacetic acid

Gbp – gigabase pair

GC – guanine-cytosine

gDNA – genomic deoxyribonucleic acid

GOLD – genomes online database

GPS – global positioning system

HGT – horizontal gene transfer

HMM – hidden Markov model

ICP-MS – inductively coupled plasma mass spectrometry

IMG – integrated microbial genomes

iTOL – interactive tree of life

JGI – joint genome institute

KEGG – Kyoto encyclopedia of genes and genomes

KM – KEGG module

KO – KEGG ortholog

MAG – metagenome assembled genome

MG-RAST – metagenomic rapid annotations using subsystems technology

MIC – minimum inhibitory concentration

MiGA – microbial genomes atlas

NA – not applicable

$\text{Na}_2\text{HAsO}_4$  – sodium arsenate

$\text{NaAsO}_2$  – sodium arsenite

NCBI – national center for biotechnology information

$\text{NH}_4$  – ammonium

nr – non-redundant

$\text{OD}_{590}$  – optical density at wavelength of 590 nm

OTU – operational taxonomic unit

PA – Pennsylvania

PCR – polymerase chain reaction

RDP – ribosomal database project

rpm – rotations per minute

rRNA – ribosomal ribonucleic acid

SAM – sequence alignment map

TE – tris-EDTA

TSA50 – 50% trypticase soy agar

TSB50 – 50% trypticase soy broth

## KEY TO SYMBOLS

$<$  – less than

$=$  – equal to

$>$  – greater than

$\leq$  – less than or equal to

$\geq$  – greater than or equal to

$A$  – maximum optical density

$\gamma$  – lag time

$\mu$  – growth rate

$\rho$  – Spearman's rho

## **CHAPTER 1 : Introduction**



## **Overview**

Soil harbors immense microbial biodiversity. This diversity, which constitutes the soil microbiome, serves as a reservoir of functions that can affect public health. These functions include 1) arsenic resistance and metabolism and 2) antibiotic resistance. Arsenic resistance and metabolism impacts the biogeochemical cycling of arsenic (1), a toxic metalloid, and the dissemination of antibiotic resistance genes (ARGs) can lead to antibiotic resistant infections. Both gene groups are widespread in the environment and can be horizontally transferred. Additionally, arsenic related genes and ARGs have been found together on plasmids (2–4). Despite relevance to public health, the full scope of the environmental distribution, genomic distribution (chromosome vs. plasmid) diversity, and interspecies transfer of these genes is unknown. This knowledge gap is due, in part, to the immense diversity of the soil microbiome. This dissertation uses an environmental disturbance gradient and soil meta-analyses to examine the distribution and dynamics of ARGs and arsenic related genes in soil.

## **Arsenic tolerance, resistance, and metabolism**

Arsenic is a toxic metalloid included on the Environmental Protection Agency's list of priority pollutants (5). Soil arsenic levels are generally low ( $< 15$  ppm) (6), but anthropogenic activities, including application of arsenic-containing pesticides, burning fossil fuels, and mining increase environmental arsenic concentrations (7). All domains of life are sensitive to arsenic because it exhibits toxicity on a cellular level by inhibiting energy production, causing oxidative stress, and inappropriately binding proteins (8). In humans, acute arsenic exposure can be fatal, while chronic arsenic exposure causes gastrointestinal, cardiovascular, and kidney disease as well as a multitude of cancers (9). Chronic exposure from drinking water contamination alone currently impacts over 200 million people worldwide (10). Exposure prevention is complex

because the flow of arsenic through the environment depends on anthropogenic activities, geochemistry, and microbial activities (11). As an element, arsenic cannot be degraded, so exposure prevention is the most effective measure to reduce arsenic-related disease. The environmental fate of arsenic is thus an important health concern. The toxicity and mobility of arsenic varies between its two most common oxidation states with arsenate being less mobile and less toxic than arsenite (12). While the mobility of arsenic depends on several abiotic factors in the environment, including oxygen levels and pH (13), microorganisms are the main drivers of arsenic biogeochemical cycling (14).

Because arsenic is toxic (8), microbes have evolved a variety of mechanisms to cope with this element, including tolerance, resistance, and metabolism (15–21). Tolerance mechanisms include intracellular chelation, adjusting the cytoplasmic redox environment, and biofilm formation (21). These tolerance mechanisms provide transient protection and do not alter arsenic speciation, but arsenic resistance and metabolism, encoded by arsenic related genes, can impact the biogeochemical cycling of this metalloid. This dissertation will focus on four major functional classes of genes encoding arsenic resistance and metabolism in Bacteria and Archaea.

#### *Arsenite efflux pumps and accessory functions (*acr3*, *arsB*, *arsD*)*

Perhaps the most streamlined form of arsenic resistance is the extrusion of arsenite from the cell through arsenite efflux pumps. The minimum requirements for resistance to arsenite with this mechanism include a repressor protein, ArsR, along with an efflux pump (22, 23). Genes encoding arsenite efflux pumps are present in a variety of soil microorganisms and include *acr3* and *arsB*. Additional genes included in the operon include *arsA* and *arsD*. Gene *arsA* encodes an ATPase that drives arsenite efflux pumps (24). Gene *arsD* encodes a repressor and arsenite

chaperone (25). Arsenite efflux pumps can also function downstream of arsenate reductases, which will be discussed in the next section.

Arsenite permeases ArsB and Acr3 both remove arsenite from the cytoplasm, but they are not homologous (26). For example, COG0798 (from the COG database (27)) is entitled “Arsenite efflux pump ArsB, ACR3 family” (28), but it detects only Acr3 family sequences. *acr3* is diverse and requires two sets of primers, one for each clade: *acr3*(1) and *acr3*(2) (29, 30). Most sequences in clade *acr3*(1) belong to Actinobacteria (29, 30), but clade *acr3*(1) is not monophyletic. Studies have found sequences of *acr3*(1) that belong to Firmicutes as well as Proteobacteria and are distinct from Actinobacteria-related sequences (29–31). Clade *acr3*(2) is known to contain mostly Proteobacteria (29–31). Known *arsB* sequences have lower phylogenetic diversity (29, 30), and only one primer set is required to target *arsB* (29). Evidence exists for horizontal gene transfer (HGT) of both *acr3* and *arsB* (30), and both are found on plasmids (32, 33).

#### Cytoplasmic arsenate reductases (*arsC* (trx), *arsC* (grx))

In Bacteria and Archaea, cytoplasmic arsenate reduction is encoded by two non-homologous genes: *arsC* (trx) and *arsC* (grx). These genes are phylogenetically widespread, suggested to be evolutionarily ancient (16), and have been shown to undergo HGT (34). *arsC* (grx) likely evolved from a glutaredoxin family protein (35), and *arsC* (trx) likely evolved from a phosphotyrosine phosphatase (36). Additionally, they are thought to differ in their efficiency to reduce arsenate with *arsC* (trx) having greater efficiency than *arsC* (grx) (31, 37). *arsC* (grx) is typically associated with Gamma- and Alphaproteobacteria while *arsC* (trx) is associated with low GC Gram positive Bacteria (16). While some argue that both *arsC* genes evolved from a

single ancestral gene (16), more recent reports suggest that cytoplasmic arsenate reduction evolved multiple times throughout history (15).

#### Arsenite methylation (*arsM*)

The gene *arsM* encodes an arsenite S-adenosylmethionine (SAM) methyltransferase which adds methyl groups to arsenite (38). As its nomenclature suggests, *arsM* can be incorporated into *ars* operons and is regulated by the transcriptional repressor ArsR (39). In this dissertation, *arsM* is separated from general discussions of the *ars* operon due to its distinct impact on arsenic biogeochemical cycling. While the *ars* operon is commonly known for reduction of arsenate and subsequent efflux of arsenite, ArsM methylates arsenic, ultimately creating organoarsenical compounds. The evolutionary history of *arsM* was recently reported (19, 40), and revealed that ArsM and its eukaryotic homolog AS3MT are widespread throughout the tree of life. Two independent analyses concluded that several inter-domain transfers of *arsM* have occurred (19, 40). Chen and colleagues (2017) also suggest that ArsM evolved in Cyanobacteria roughly 2.5 billion years ago (19). Thus, the evolutionary history of *arsM* is both long and complicated.

#### Bioenergetic arsenic related genes (*aioA*, *arxA*, *arrA*)

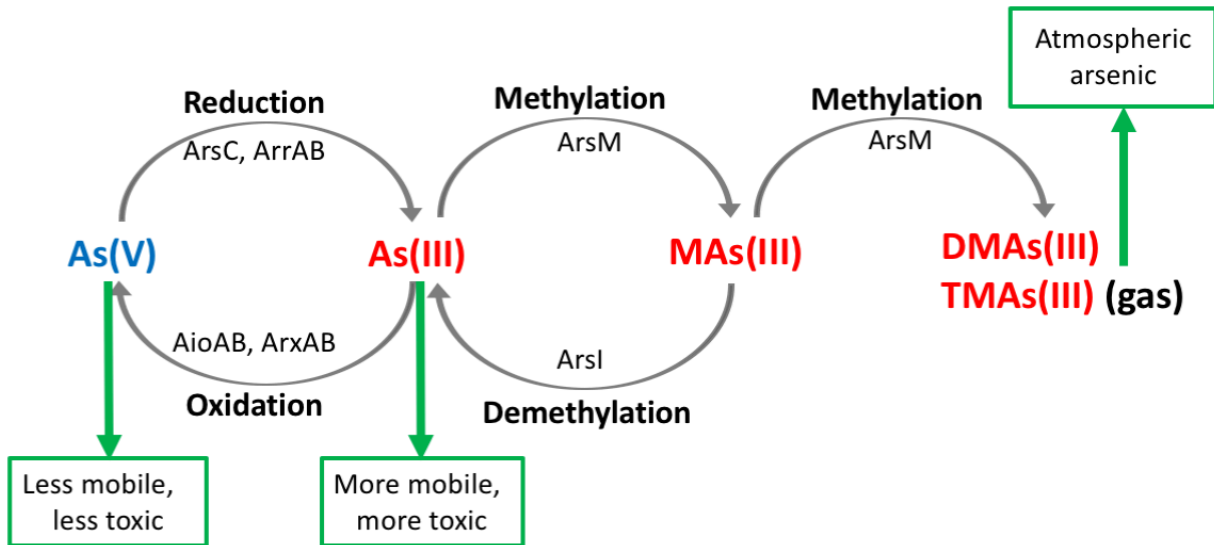
Arsenic bioenergetic enzymes include AioAB, ArxAB, and ArrAB. All three of these enzymes are molybdoenzymes and belong to the DMSO-reductase superfamily (18). Most known organisms harboring these genes belong to phylum Proteobacteria (21, 30, 31, 41–43). Despite these similarities, all three perform distinct functions and have distinct evolutionary histories.

AioAB is a periplasmic arsenite oxidase that functions in both aerobic and anaerobic environments (18). Nomenclature in this field was updated when Lett and colleagues (2012) replaced three circulating names for an arsenite oxidase (AroAB, AoxAB, AsoAB) with AioAB (44). AioAB versions have been found in Bacteria and Archaea (18, 21), and phylogenetic analysis of *aioA* shows a separation of Bacterial and Archaeal-derived sequences (18, 45, 46), suggesting that *aio* evolved before the divergence of Bacteria and Archaea (18).

Arx and Arr do not have known representatives from Archaea and thus are thought to have evolved in Bacteria (18). ArxAB is an anaerobic arsenite oxidase while ArrAB is a dissimilatory arsenate reductase (46). ArxAB is thought to have evolved in Bacteria (46), and Arr is thought to have evolved from Arx (18). Because ArxAB and ArrAB are phylogenetically similar and because ArxAB was not formally discovered until 2010 (47), papers published prior to 2010 could have improperly annotated ArxAB as ArrAB (48). *arrA* has representatives in Chrysiogenetes, Crenarchaeota, Deferribacteres, and Firmicutes in addition to Proteobacteria (21), but all known *arx* representatives are in phylum Proteobacteria (21).

### Arsenic biogeochemical cycling

Bacteria and Archaea impact arsenic biogeochemical cycling through arsenic reduction, oxidation, methylation, and demethylation (**Figure 1.1**). Reduction of arsenate produces the more toxic and more mobile arsenite. Cytoplasmic arsenate reduction (encoded by *arsC* (grx) and *arsC* (trx)) reduces arsenate to arsenite and allows subsequent extrusion of arsenite through efflux pumps (encoded by *acr3*, *arsB*) (8). Arsenate can also be used as a terminal electron acceptor through a dissimilatory arsenate reductase (encoded by *arrAB*) (49). The genes *aioAB* and *arxAB* encode arsenite oxidases (49, 50) which decreases arsenite mobility and is often



**Figure 1.1. Arsenic biogeochemical cycle.**

Species of arsenic are colored based on oxidation state: oxidized (blue) and reduced (red).

Reaction descriptions are labeled outside of grey arrows, and related genes are labeled within arrows. Green arrows and boxes highlight environmental consequences of different arsenic species.

considered a tool for bioremediation. Genes encoding biomethylation (*arsM*) (51) and demethylation (*arsI*) (52) are understudied but nonetheless occur in the environment and increase or decrease the volatilization of arsenic respectively. Microbial interactions with arsenic can therefore impact the distribution of arsenic in the environment. The abundance of these genes in the environment is thus an important consideration for bioremediation and risk assessment (21).

#### *Arsenic exposure and arsenic related genes*

Arsenic is common in the environment, but its concentrations range from < 1 ppm to > 4,000 ppm (53). A low-arsenic site (< 10 ppm) was shown to harbor cultivable arsenic resistant isolates (6), and arsenic related genes are present across a wide variety of arsenic concentrations (37, 54). Thus, arsenic related genes are a poor candidate for arsenic monitoring. The impact of arsenic on microbial communities, however, can be estimated using arsenic gradients. Arsenic gradients over small geographic scales can occur naturally from geological variations or unnaturally from anthropogenic activities such as mining. These gradients are useful for probing the influence of arsenic on arsenic related gene distributions. Despite being a known, strong selective pressure, the impact of arsenic on arsenic related gene content is convoluted.

Several studies have examined arsenic related genes along environmental arsenic gradients. Escudero and colleagues (2013) examined water and sediment samples from the Andean plateau, which has a naturally occurring arsenic gradient (0.1 – 10,000 ppm As) (37). They screened for the presence of *acr3*, *arsC* (grx), *arsC* (trx), *aioA*, and *arrA* using PCR. Genes *acr3*, *arsC* (trx), *aioA*, and *arrA* were detected in the majority of samples regardless of arsenic concentration; however, *arsC* (grx) was ubiquitous in low-arsenic samples (< 10 ppm) and only detected in one high-arsenic sample (37). Similarly, Luo and colleagues (2014) examined

metagenomes from soils with arsenic ranging from 30 – 820 ppm and observed arsenic related genes *acr3*, *arsB*, *arsC* (trx), *aioA*, and *arrA* in all samples (54). While end-point PCR only gives occurrence information, the authors also examined whether arsenic related genes changed in diversity or abundance along the arsenic gradient. Samples with higher arsenic concentrations had a greater diversity and abundance of arsenic related genes. Furthermore, arsenic concentration was correlated with *aioA* and *arsC* (trx) (54). Similarly, metagenomic analysis of arsenic related genes in paddy soils showed a relationship between arsenic concentration and genes *arsC*, *arsM*, *aioA*, and *arrA* (55).

Some studies have found gene-specific changes along arsenic gradients, but these are inconsistent across studies and environments. Desoeuvre and colleagues (2015) compared arsenic related genes from waters upstream, within, and downstream of acid mine drainage (AMD). Using PCR and clone libraries, the authors tested for the presence of *acr3*, *arsB*, *arsM*, and *arrA* (56). Reinforcing previous results, arsenic related genes were present in all samples. Gene *arsB*, however, was more diverse in upstream, pristine samples compared with AMD-contaminated samples. Notably, *acr3* diversity did not change, suggesting that *arsB* is less tolerant to high arsenic concentrations. Hu and colleagues (2019) estimated the arsenic related gene content of iron plaques from paddy soils with varying arsenic concentrations (6.7 – 210 ppm) (57). They observed greater relative abundance of *arsC* and *arsB* in samples with higher arsenic concentrations. Arsenic metabolism genes *arr*, *aio*, and *arsM* were generally detected across all samples but in low abundance (57). Costa and colleagues (2015) examined metagenomes from freshwater sediments with 85 and 297 ppm arsenic (58). They detected *arsABCDHR* and *arrAB* in at least one sample. *arsC* was not detected in the low arsenic site but was detected in the higher arsenic site, while *arrAB* and *arsD* were more abundant in the lower



arsenic site. Thus, while arsenic has a direct relationship with arsenic related genes, future work is needed to distinguish between background resistance and arsenic-driven changes.

#### *Abiotic gradients and arsenic related genes*

Environmental gradients are known to have strong influences on microbial community structure (59–61), which can impact functional potential. As discussed previously, arsenic related genes are widespread, and several are phylogenetically conserved. Thus, arsenic related gene content of the soil microbiome is subject to change with changing community membership regardless of changes in arsenic exposure. Several studies have documented this phenomenon by examining environmental gradients that were not characterized by changing arsenic concentrations (55, 62, 63).

For example, Zhang and colleagues (2015) examined 13 paddy soils across Southern China with low arsenic concentrations and found a relationship between soil pH and arsenic related gene distribution (55). Individual arsenic related genes were correlated with other geochemical parameters as well (55). A follow up analysis examined metagenomes from similar sites and also observed a relationship with pH (62). Similarly, a clone library study of arsenic related genes in AMD waters observed a relationship between arsenic related gene content and pH (56). Because there is a well-described relationship between soil pH and microbial community structure (64), the observation that a change in soil pH can result in a change in arsenic related gene content is expected.

Other factors can influence arsenic related gene content as well. Metal (and metalloid) resistance has been shown to co-occur with antibiotic resistance (65–67). While multiple

examples of metal pollution impacting ARG content in the environment (68–70), the reverse relationship could also be true especially because co-occurrence is more common on genomes than plasmids (66). Thus, geochemical and anthropogenic factors beyond arsenic are likely to impact arsenic related gene diversity and abundance in the environment.

### **Antibiotic resistance**

ARGs are another group of functional genes in soil microbiomes that are important for public health. The Centers for Disease Control and Prevention estimates that 23,000 people die each year due to infections from antibiotic resistant bacteria in the United States (71). While these infections are caused by ARGs in clinically-relevant strains, ARGs present in nonpathogenic strains are also important. This importance is because there is a potential for HGT of ARGs to pathogenic strains (72–74). This potential transfer is particularly relevant for the soil microbiome because soil is considered a reservoir for ARGs due to greater ARG diversity than the clinic (75).

While soil harbors a diverse array of ARGs, it is unclear how much this reservoir impacts clinical antibiotic resistance (76). Understanding the propagation and dissemination of ARGs in soil is difficult because multiple interacting factors influence their fate (77, 78). Like arsenic related genes, direct selective pressure (e.g. antibiotics) can influence ARG abundance (79). Additionally, several ARGs are thought to have evolved > 2 billion years ago (80) and can propagate vertically without direct selection pressure. For example, abiotic changes, such as nitrogen addition, have been shown to alter ARG distributions in soil (81, 82), and increased temperatures have been shown to reduce ARG abundance (83, 84). Furthermore, it is unclear what proportion of ARGs in soil are mobile, and HGT rates in bulk soil are estimated to be lower

than areas with higher population densities such as gut microbiomes and the phyllosphere (85). Additionally, a functional metagenomic study across soil types showed evidence for predominantly vertical, rather than horizontal, transfer of ARGs (82). Investigations of the dynamics of ARGs in soil microbiomes and the distribution of ARGs on soil-associated chromosomes and plasmids would improve understanding of their environmental dissemination.

### **Environmental Disturbance and Centralia, PA**

The abundance of ARGs and arsenic related genes depends on microbial community membership, direct selective pressure (e.g., from antibiotics or arsenic), and rates of HGT. Local antibiotic and arsenic concentrations select for organisms with functional ARGs and arsenic related genes, respectively, but relationships between abundance of arsenic related genes and some ARGs within a microbial community are nuanced, in part, due to their long evolutionary histories (86–88). Vertical transfer of arsenic related genes throughout evolutionary history implies that these genes are not necessarily correlated with contemporary arsenic concentrations (86, 87). Similarly, ARGs have been detected 30,000 year old permafrost samples (88) and in caves with minimal anthropogenic influence (89). Organisms harboring antibiotic and arsenic related genes also can increase in abundance in the absence of direct selection pressures (e.g. in response to some other biotic or abiotic driver). In present day, antibiotic and arsenic related genes are found on chromosomes and several mobile genetic elements (21), and HGT of ARGs and arsenic related genes has been shown (30, 34, 43, 90), indicating that the numbers and distributions of these genes are subject to change without parallel turnover in microbial community membership. The abundance of these functional genes and the impact of their encoded mechanisms are thus intimately linked with community dynamics.

Environmental disturbance gradients provide an appropriate framework to study functional gene dynamics. Disturbances such as increased temperature and pollution can impact diversification (91), population diversity (92) and rates of HGT (93, 94). While many disturbances are transient, a good system to test hypotheses about functional gene dynamics should include a multigenerational disturbance with a clear endpoint. Underground coalmine fires act as a multigenerational disturbance and expose soil microbial communities to increased temperatures as well as coal combustion pollutants. Soil surface temperatures of coal mine fire-affected areas range from 21-80°C (95). Steam from these fires emits gases including CO and CO<sub>2</sub> (95–97), and arsenic, along with lead, zinc, mercury, and copper have all been found surrounding active vents of underground coal fires (98). The Centralia, PA coalmine fire ignited in 1962 and has traveled along the coal seam ever since, creating a fire-impact gradient that can be linked to historical fire movement.. The soil microbial communities overlying the underground coal mine fire thus experience a multitude of stressors, including temperature and pollution. The influence of coalmine fires on soil microbial community antibiotic and arsenic related genes is unknown, but it may influence their distribution and diversity.

## **Summary and Aims**

This dissertation aims to improve understanding of the distribution and dynamics of two functional genes groups in soil: 1) antibiotic resistance and 2) arsenic resistance and metabolism. Using a model disturbance and soil meta-analyses, this dissertation investigates hypotheses about resistance gene ecology. The Centralia, PA ecosystem is an underground coal seam fire and was used as a model disturbance. This system is used to test whether microbiome functional potential changes with disturbance (Chapters 2 – 3). Furthermore, bioinformatic methods were developed to detect ARGs and arsenic related genes and then applied to soil microbiomes (Chapters 4 – 5). Chapter 2 describes the genetic and functional characterization of an isolate collection of arsenic

resistant bacteria cultivated from soil impacted by the Centralia disturbance. Chapter 3 tests the hypothesis that the Centralia disturbance reduced clinically-relevant ARGs in the soil. Chapter 4 compares the diversity and distributions of ARGs on plasmids among cultivable microorganisms from soil and other environments (e.g., clinic). Chapter 5 is a meta-analysis of the distribution and diversity of arsenic related genes in soil that leverages public metagenomes and metatranscriptomes. New bioinformatic tools were applied to soil microbiomes to elucidate the distributions, diversity, and dynamics of ARGs and arsenic related genes. Ultimately, this work contributes to understanding of ARGs and arsenic related genes in the environment and provides insights into mitigation of antibiotic resistance and arsenic bioremediation.

## **CHAPTER 2 : Taxonomically-linked growth phenotypes during arsenic stress among arsenic resistant bacteria isolated from soils overlying the Centralia coal seam fire**

Work presented in this chapter has been published as Dunivin TK, Miller J, and Shade A. Taxonomically-linked growth phenotypes during arsenic stress among arsenic resistant bacteria isolated from soils overlying the Centralia coal seam fire. *Plos ONE*. 2018; 13(1)

## Abstract

Arsenic, a toxic element, has impacted life since early Earth. Thus, microorganisms have evolved many arsenic resistance and tolerance mechanisms to improve their survival outcomes given arsenic exposure. We isolated arsenic resistant bacteria from Centralia, PA, the site of an underground coal seam fire that has been burning since 1962. From a 57.4°C soil collected from a vent above the fire, we isolated 25 unique aerobic arsenic resistant bacterial strains spanning seven genera. We examined their diversity, resistance gene content, transformation abilities, inhibitory concentrations, and growth phenotypes. Although arsenic concentrations were low at the time of soil collection (2.58 ppm), isolates had high minimum inhibitory concentrations (MICs) of arsenate and arsenite (>300 mM and 20 mM respectively), and most isolates were capable of arsenate reduction. We screened isolates (PCR and sequencing) using 12 published primer sets for six arsenic resistance genes. Genes encoding arsenate reductase (*arsC*) and arsenite efflux pumps (*arsB*, *ACR3(2)*) were present, and phylogenetic incongruence between 16S rRNA genes and arsenic resistance genes provided evidence for horizontal gene transfer. A detailed investigation of differences in isolate growth phenotypes across arsenic concentrations (lag time to exponential growth, maximum growth rate, and maximum OD<sub>590</sub>) showed a relationship with taxonomy, providing information that could help to predict an isolate's performance given arsenic exposure *in situ*. Our results suggest that microbiological management and remediation of environmental arsenic could be informed by taxonomically-linked arsenic tolerance, potential for resistance gene transferability, and the rare biosphere.

## Introduction

Arsenic, a toxic metalloid, is naturally present in soil, but levels are generally low ( $< 10$  ppm) (99). Because of the ubiquity of arsenic and its toxicity, bacteria have evolved a variety of arsenic-specific detoxification mechanisms (21). Bacterial strains have been shown to oxidize, reduce, methylate, and demethylate arsenic (50). The toxicity and mobility of arsenic can change depending on its oxidation state with arsenate ( $\text{As}^{5+}$ ) being less soluble and less toxic than arsenite ( $\text{As}^{3+}$ ) (100); thus, environmental bacteria are considered important constituents of the biogeochemical cycling of arsenic because the presence and transfer of the resistance genes encoding these activities affect the mobility of arsenic.

Arsenic resistance genes can be located on chromosomes, plasmids, or both (21). Several studies indicate that horizontal gene transfer (HGT) has occurred with arsenic resistance genes (16, 34, 43, 90, 101), suggesting the potential exists for arsenic resistance genes to propagate in a microbial community given a selective pressure of arsenic exposure; however, timing of HGT is difficult to determine (16). In addition to arsenic-specific mechanisms of resistance conferred by arsenic resistance genes, microorganisms can also employ nonspecific and transient cellular mechanisms to withstand arsenic exposure. Cell envelope permeability to arsenic, oxidative stress response, and heat shock proteins have all been shown to be differentially regulated in response to arsenic (21, 102–106). These are collectively referred to as arsenic tolerance mechanisms (103, 107). However, tolerance in the absence of resistance (i.e. arsenic resistance genes) is often not enough to enable cell survival given lasting arsenic exposure (107).

Much of the current understanding of arsenic resistance and tolerance has come from the detailed study of arsenic resistant isolates that have been cultivated from arsenic contaminated



sites (e.g., (30, 34, 101, 108–112). More broadly, culture-dependent approaches to improve knowledge of microbial diversity and functions are experiencing a renaissance in today's age of high-throughput meta 'omics (e.g., (113–115). In addition to direct assessment of physiology and functional capabilities, characterized isolates can provide high quality genome references for culture-independent metagenome and single-cell genome assemblies (116–118). Thus, culture-dependent approaches continue to offer opportunity to examine several aspects of arsenic resistance that are not captured with culture-independent approaches. For example, growth phenotypes in arsenic and minimum inhibitory concentrations (MICs) are best determined directly with isolates. Additionally, it is difficult to assess potential horizontal gene transfer (HGT) from culture-independent methods (119, 120), and HGT is an important consideration in arsenic resistance gene ecology. Finally, cultured isolates provide access to microorganisms that may be used to support applications like bioremediation of contaminated sites (e.g., (116, 118)). Though isolate collections do not provide comprehensive knowledge of microbial diversity and are limited by cultivation conditions, these collections can be used to inform isolate ecology in the context of their larger microbial community, especially when coupled with culture-independent approaches (e.g., (121)).

The underground coal seam fire in Centralia, PA ignited in 1962 and has been burning ever since. The soil microbial communities overlying the underground fire experience a multitude of fire-related stressors, including high temperatures and exposure to coal combustion products and CO and CO<sub>2</sub> gas emissions; these coal fire pollutants impact local biogeochemistry (95–97). Because arsenic is naturally present in coal, exposure to the coal seam fire is expected to influence soil microbial arsenic resistance and arsenic resistance gene transfer. Along with lead, zinc, mercury, and copper, arsenic has been documented in increased concentrations near

active vents, which are steaming surface fissures created by instability from the underground coal fire (98).

Our objective was to characterize arsenic resistant bacterial isolates from an active thermal vent (57.4°C) in Centralia in order to expand knowledge of the characteristics of arsenic resistant bacterial isolates from a coal mine contaminated site. This knowledge will improve metagenome analysis and genomic analysis of similar organisms, as there is a move towards expanding culture collections and knowledge of cultivated organisms (e.g., (118)). We aimed to gain insights into their genetic mechanisms of arsenic resistance, growth consequences under increasing arsenite and arsenate exposure, and potential for interspecies transfer of arsenic resistance. Our culture-dependent approach provided insights into isolate distinctions in growth phenotypes given arsenic exposure. Considering culture-independent information (16S rRNA gene amplicon sequencing) additionally allowed us to determine the relative contributions of these isolates to their larger community. These findings bring to light complexities of predicting microbial community-level response to arsenic.

## Materials and Methods

### Soil collection and site description

The Pennsylvania Department of Environmental Protection provided permission to access the field site. The field site is not a protected area. This work did not involve endangered or protected species. This study did not involve vertebrates. A soil surface core (20 cm depth and 5.1 cm diameter) was collected in October 2014 from an active vent (steam escaping) in Centralia, PA. This vent (site Cen13, GPS coordinates: 40 48.070, 076 20.574) was selected because it has had historical fire activity since at least 2007 (95) and was the hottest detected at the time of sampling with a measured surface temperature (10 cm depth) of 57.4°C (ambient air temperature was 13.3°C). Detailed soil geochemical data was assayed by the Michigan State University Soil and Plant Nutrient Laboratory (East Lansing, MI, USA, <http://www.spnl.msu.edu/>) according to their standard protocols, and total arsenic was measured by Element Materials Technology using the Environmental Protection Agency's method 3050B for sample preparation and ICP-MS (**Appendix A Table 1**). Upon sampling, the soil was kept on ice until transport to the lab where it was manually homogenized, sieved through 4 mm mesh, and stored at -80°C until further processing.

### Cultivation-dependent soil bacterial community growth

Five grams of soil were removed from -80°C and kept at 4°C for 48 h. The soil was warmed to room temperature for 1 h and then suspended in 25 mL of sterile Dulbecco's phosphate-buffered saline (ThermoFisher; dPBS), vortexed for 2 min, and allowed to settle for 2 min. The supernatant was plated onto 50% tryptic soy agar (Becton Dickinson and Company; TSA50) with 200 µg mL<sup>-1</sup> of cycloheximide added to inhibit fungal growth. Plates were incubated at 27°C for 24 h. To obtain a culture-dependent bacterial community representative of

these growth conditions, overgrown plates were scraped to make a 25% glycerol stock and stored at -80°C for future assays.

#### Isolation of arsenic resistant bacteria

Twenty mL of trypticase soy broth (TSB50) was inoculated with the bacterial community glycerol stock and grown for 6 h with shaking at 200 rpm and 12 mm amplitude. Arsenic was not included in the medium to avoid transfer of arsenic resistance genes. The culture was plated onto TSA50 with either 10 mM Na<sub>2</sub>HAsO<sub>4</sub> or 1 mM NaAsO<sub>2</sub> to screen for arsenate or arsenite resistant colonies, respectively. Ninety-four total colonies (35 from sodium arsenate; 59 from sodium arsenite) were streaked to purity (3x) on their respective media type; 69 pure isolates were recovered and made into 25% glycerol stocks for long term storage at -80°C. From these pure cultures, 25 distinct isolates were identified by genotype with 16S rRNA gene sequencing and by phenotype using MIC assays.

#### Morphological characterization and temperature maxima

Overnight cultures of isolates grown in 3 mL TSB50 were examined using a Nikon E800 Eclipse microscope. Cell morphology was visualized using a photometrics CoolSnap MYO microscope camera (Tuscan, AZ, USA) and Micromanager 4.22 (122) was used for image acquisition. Cell size was measured using Fiji image analysis software (123). Colony morphology on TSA50 plates was imaged after incubation at 27°C for 24 h. To measure growth temperature maxima, isolates (2% culture in fresh TSB50) were incubated in a T100 Thermo Cycler (BioRad) for 24 h with a thermal gradient (32-52°C). Optical density at 590 nm (OD<sub>590</sub>) was measured using an Infinite F500 plate reader (Tecan). The maximum temperature for growth

was determined as the highest temperature with an increase in OD<sub>590</sub> from background. This process was repeated for a minimum of two biological replicates per isolate.

#### DNA extraction and quantification

Freezer stocks of isolates were inoculated into 3 mL TSB50 and shaken at 27°C at 200 rpm with a 12 mm amplitude until turbid. Genomic DNA (gDNA) was extracted using the E.Z.N.A. Bacterial DNA Kit (Omega Bio-Tek) according to the manufacturer's instructions. Isolated gDNA was quantified with fluorometry using the Qubit dsDNA broad range assay kit (Invitrogen) and a Qubit 2.0 (Invitrogen) according to the manufacturer's instructions. DNA was stored in sterile Tris-EDTA buffer (Sigma; pH 8) at -20°C.

#### Endpoint PCR and amplicon sequencing

The near full length 16S rRNA gene was amplified for each isolate using the universal primer pairs Uni-27F and Uni-1492R (**Appendix A Table 2**). PCR amplification of 16S rRNA was carried out in a T100 Thermo Cycler (BioRad) using 25 µL total volume including 30 ng genomic DNA, 0.4 µM of each primer, 0.8 mM dNTPs (Sigma), 2.5 µL 10X Pfu Buffer (Promega), 2X high fidelity Pfu DNA Polymerase (Promega), and nuclease free water to a final volume of 25 µL. The 16S rRNA PCR reaction cycle included a 2 min initial denaturation at 95°C, 30 cycles of denaturation at 95°C for 30 s, annealing at 55°C for 30 s, extension at 72°C for 1 min, and a final extension at 72°C for 10 min. PCR products were run on a 1% agarose gel for 45 min at 700 mV. The PCR product of 1.4 kb from the 16S rRNA gene was gel extracted using the Wizard SV Gel and PCR Clean Up System (Promega) according to the manufacturer's instructions. Gel extraction products were quantified as described above. Purified 16S rRNA amplicons were sequenced using the ABI Prism BigDye Terminator Version 3.1 Cycle

sequencing kit by the Michigan State University Genomics Core Research Technology Support Facility. Forward and reverse 16S rRNA sequences were aligned using CAP3 (v. 3.0,(124)) to obtain near full length 16S rRNA gene sequences, except for isolates A2707, A2723, and A2735 which could not be sequenced using the 1492R primers. For these three isolates, primer U515F (125) was used to obtain a near-full length 16S rRNA sequence. Sequences were assigned taxonomy using both the Ribosomal Database Project (RDP) 16S rRNA database (v. 2.10, (126)) and the EzTaxon server (127).

Isolates were screened for the following arsenic resistance genes: *arsB*, *ACR3(1)*, *ACR3(2)*, *arsC*, *arrA*, *aioA*, and *arsM* using published primers that were chosen because of their continued use in the literature (**Appendix A Table 2**; (29, 90, 101, 128–130)). All PCRs were carried out with published reaction conditions in a T100 Thermo Cyclor (BioRad). While amplicons were obtained for all primer sets used, only products confirmed by sequencing were considered positive hits. Once a product was confirmed, the PCR was repeated using the confirmed isolate as a positive control. All amplicons were gel extracted and sequenced as described above. At least one forward and one reverse gene sequence was merged in CodonCodeAligner (v. 6.0.2, Codon Code Corporation) to create arsenic resistance gene contigs. All sequences >200 bp were submitted to NCBI, and sequences can be accessed from GenBank with the following accession numbers: 16S rRNA KX825887- KX825911, *arsC* KY405022- KY405029, *ACR3(2)* KY405030- KY405032, and *arsB* KY405033- KY405040. Four *arsC* contigs were < 200 bp and are included in **Appendix A Table 3**. Amino acid sequences for each protein-coding gene are also available in NCBI GenBank.

### Phylogenetic analysis

To compare the 16S rRNA phylogenetic diversity of *Centralia* arsenic resistant isolates to previous reports, isolates from existing literature were included in the phylogenetic analysis. Only studies with both 16S rRNA sequences > 700 bps and confirmed arsenic resistance (selection on arsenic-containing media) were included. Ultimately 6 studies (101, 110, 131–134) were included, and all sequences from relevant lineages were included in the final tree (55 sequences total). Closest 16S rRNA gene relatives deposited at the NCBI (<http://www.ncbi.nlm.nih.gov/>) were also included in the analysis. Sequences were aligned using the RDP aligner (135). RDP characters were removed from aligned sequences using BioEdit (v. 7.2.5, (136)). 16S rRNA gene trees were made with MEGA7.0 (137) and constructed with the Neighbor-joining algorithm using the Kimura 2 parameter model with 1000 bootstrap replications.

To examine the phylogeny of *arsC*, *arsB*, and *ACR3(2)* sequences, arsenic resistance gene sequences from the isolates were compared with homologous, chromosomal sequences from related organisms deposited at the NCBI. Sequences from phylogenetic relatives were found by searching chromosomes deposited at the NCBI, and closest NCBI matches for arsenic resistance gene sequences were determined using BLAST. A corresponding 16S rRNA tree was made using sequences from the isolates and their phylogenetic relatives. The sequences obtained from NCBI can be found with the following accession numbers: *Acinetobacter baumannii* strain A1 (CP010781.1), *Enterobacter cloacae* subsp. *cloacae* ATCC 13047 (296100371), *Pseudomonas aeruginosa* PAO1 (AE004091.2), *Enterobacter kobei* strain DSM 13645 (CP017181.1), *Escherichia coli* str. K-12 substr. MG1655 (NC\_000913.3), *Enterobacter asburiae* L1 (NZ\_CP007546.1), *Bacillus cereus* ATCC 10987 (AE017194.1), *Paenibacillus*

*terrae* HPL-003 (374319880), *Bacillus thuringiensis* strain Bc601 (CP015150.1), *Shewanella oneidensis* MR-1 (NC\_004347), *Stenotrophomonas maltophilia* K279a (AM743169.1), *Bacillus thuringiensis* strain 97-27 (CP010088.1), *Rhodoferrax ferrireducens* T118 (CP000267.1), *Cyclobacterium marinum* DSM 745 (CP002955.1) Trees were constructed using MEGA7.0 (137) and constructed with the maximum likelihood algorithm using the Kimura 2 parameter model with 100 bootstrap replications. Distances between arsenic resistance and 16S rRNA gene trees were calculated using the R environment for statistical computing (138) with the Phangorn package (139).

To further investigate evidence for HGT, the GC content of arsenic resistance gene sequences was compared with reference GC content from whole genomes of related species. Reference GC content was calculated by averaging the GC content of all organisms in NCBI “Genome Groups” for the related taxon.

#### Cultivation-independent 16S rRNA amplicon sequencing and analysis

Soil DNA was extracted, sequenced, and analyzed in a previous work (140) from the same sample used for isolation. Using BLAST (v. 2.2.26), a database of representative 16S rRNA gene sequences was constructed. Isolate 16S rRNA gene sequences from Sanger sequencing were used as queries against this database to find top hits and to estimate the relative abundance of our isolates in the microbial community. The top hit was determined as the hit with the highest percent identity for that isolate with a minimum percent identity of 96%, and the relative abundance of representative sequence (140) was used as the estimate of the relative abundance of each isolate.



### Arsenic transformation capabilities

The ability of the isolates to reduce arsenate or oxidize arsenite was measured using a slightly modified (described below) silver nitrate colorimetric assay as described previously (141). 0.1 M Tris-HCl (pH 7.3) was used as a reaction buffer instead of 0.2 M, and 1.33 mM sodium arsenate or sodium arsenite was used instead of 0.67 mM. Cells were inoculated in 3 mL TSB50 and incubated at 27°C for 15 h before plating. Cells were washed with sterile reverse osmosis (RO) water to remove culture media as indicated in Simeonova *et al.* (141), and 20 µL of the washed cell suspension was incubated with 80 µL of 0.1 M Tris-HCl and 1.33 mM As in a 96-well plate for 72 h at 27°C. Two standard curves with different ratios of sodium arsenate and sodium arsenite (0:100, 10:90, 25:75, 50:50, 75:25, 90:10, 100:0) were also included alongside the cells. After a 72 h incubation, cell viability was tested. Cells were patched onto fresh TSA50 plates to test cell viability. The reaction was initiated by adding 100 µL of sterile 0.1M AgNO<sub>3</sub> to each sample in the 96-well plate. After the silver nitrate reaction was initiated, plate photographs were taken, and colorimetric changes were assessed. This protocol was performed with at least two biological replicates plated in duplicate.

### Minimum inhibitory concentrations (MICs)

To determine the MICs of arsenate and arsenite as well as their growth phenotypes, isolates were inoculated from 25% glycerol stocks into 3 mL TSB50 and incubated with shaking at 200 rpm with a 12 mm amplitude at 27°C for 6 h. Inocula were added to a 96-well plate with arsenic-containing TSB50 to make a 1% inoculum. Concentrations tested include 0, 10, 50, 100, 150, 200, 250, and 300 mM sodium arsenate and 0, 1, 3, 5, 7, 10, 14, and 20 mM sodium arsenite. Plates were shaken continuously at 288 rpm with a 3 mm amplitude in an Infinite500 plate reader (Tecan) for 72 h at 27 ± 1°C. OD<sub>590</sub> was measured every 15 min. Growth

experiments were repeated with at least two biological replicates for each isolate, and growth curves for further analysis were made using technical triplicates.

The R environment for statistical computing (138) was used to plot growth curves and analyze key features of growth inhibition across the range of arsenate and arsenite concentrations tested using a modified script (<http://bconnelly.net/2014/04/analyzing-microbial-growth-with-r/>). Using the GroFit package (142), splining was used to extract growth parameters including time to exponential growth ( $\lambda$ ), maximum growth rate ( $\mu$ ), and maximum OD<sub>590</sub> (A). When splining was not appropriate (e.g. curves do not have a smooth fit), parameters were estimated parametrically using either Logistic, Gompertz, or Richards models informed by their Akaike information criterion (AIC) (143). Parameters for each isolate in TSB50 containing arsenic were normalized to arsenic-free controls. We used hierarchical clustering to examine similarities in growth phenotypes in arsenic for genera with more than two representatives ( $n > 2$ ). The clustering included growth parameters ( $\lambda$ ,  $\mu$ , and A) in 1 mM sodium arsenite and 10 mM sodium arsenate for each isolate. Only one arsenic concentration was used so that MIC NA values did not impact the clustering. All R scripts are available on GitHub ([https://github.com/ShadeLab/Arsenic\\_Growth\\_Analysis/tree/master/R\\_scripts](https://github.com/ShadeLab/Arsenic_Growth_Analysis/tree/master/R_scripts)) for future studies interested in isolate fitness in arsenic.

## Results

### Taxonomic diversity and composition of arsenic resistant isolates

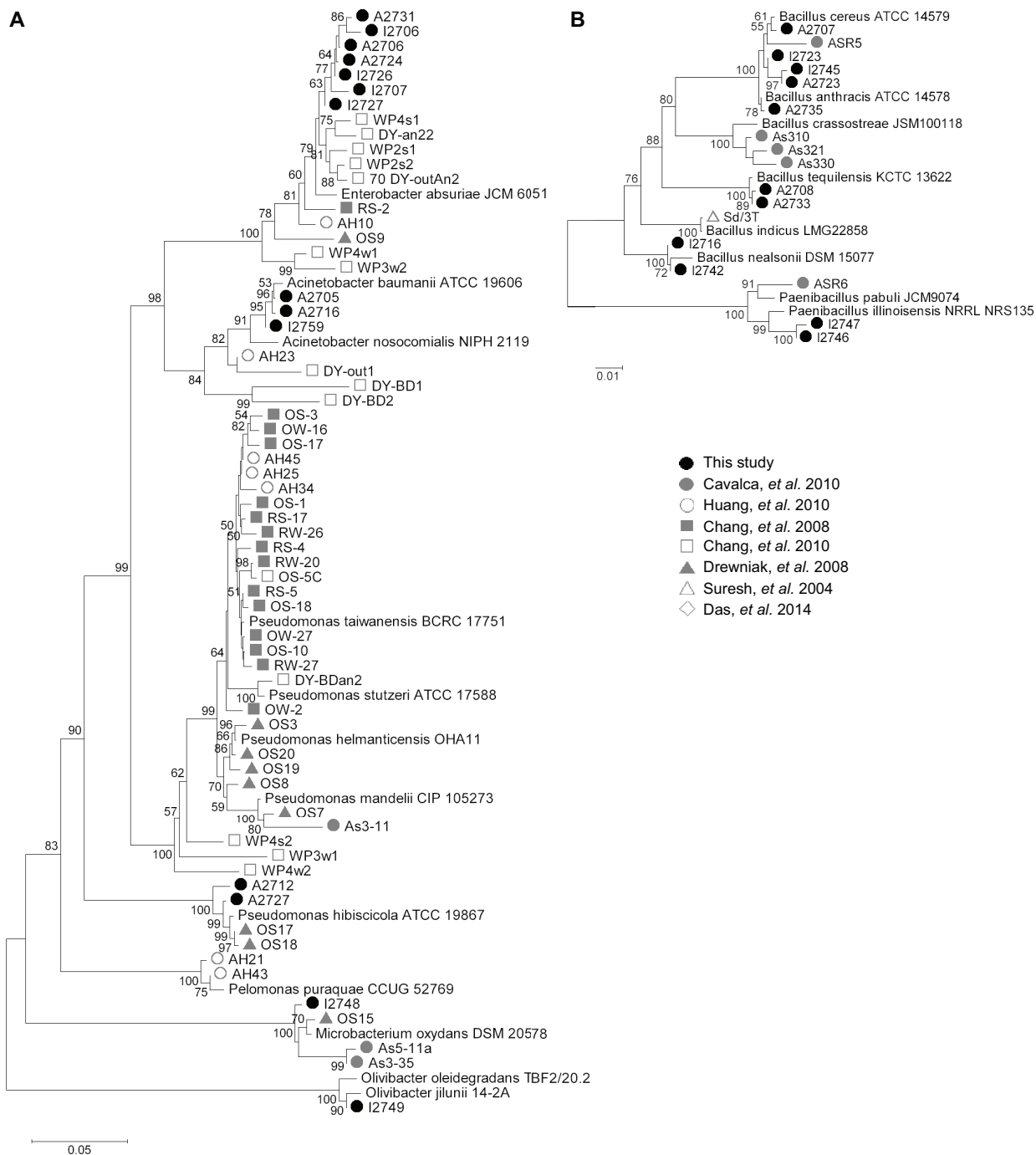
Arsenic resistant isolates were cultivated from soil near an active vent (**Appendix A Table 1**) of the Centralia coal seam fire with a low arsenic concentration (2.58 ppm) by screening for arsenic resistance on 10 mM sodium arsenate and 1 mM sodium arsenite. Isolates spanned seven genera, including *Acinetobacter*, *Bacillus*, *Enterobacter*, *Microbacterium*, *Olivibacter*, *Paenibacillus*, and *Pseudomonas* (**Figure 2.1** and **Appendix A Table 4**). The colony morphologies of the isolates aligned with expectations given 16S rRNA gene classification (near full length sequences were obtained), and all isolates grew in 24 h at or above 39°C (**Appendix A Table 4**). This cultivation effort resulted in an abundance of Firmicutes (48% of isolates). To determine the relative abundances of these arsenic resistant isolates within their larger community, we used BLAST to query isolate full-length 16S rRNA gene sequences against representative 16S rRNA gene sequences of operational taxonomic units from amplicon data (948,228 raw reads) obtained in our previous study (140). The relative abundance of top hits for each isolate ranged from  $6.23 \times 10^{-6}$  to  $1.59 \times 10^{-4}$  (**Table 2.1**), suggesting that all arsenic resistant isolates isolated in this study are rare members of this soil community.

### Genetic characterization of arsenic resistance

Arsenic resistance genotypes of the isolates were characterized using endpoint polymerase chain reaction (PCR) with a collection of published primers (**Appendix A Table 2**) specific for genes encoding resistance via diverse mechanisms, including arsenate reduction, arsenite oxidation, methylation, and arsenite efflux (**Figure 2.2A**). After endpoint PCR, all amplicons were sequenced to confirm their identities. Eight isolates (32%) had the gene encoding the arsenite efflux pump, *arsB*. The majority of *arsB*-positive isolates belonged to the

**Table 2.1. Relative abundance of isolate 16S rRNA gene sequences from our amplicon survey of the same soil.**

<b>Genus group</b>	<b>Isolates</b>	<b>Relative abundance</b>
<i>Acinetobacter</i>	I2759, A2705, A2716	6.23x10 <sup>-6</sup>
<i>Bacillus anthracis</i>	I2723, I2745, A2707, A2723, A2735	3.12x10 <sup>-6</sup>
<i>Bacillus subtilis</i>	A2708, A2733	1.03x10 <sup>-4</sup>
<i>Bacillus nealsonii</i>	I2716, I2742	1.59x10 <sup>-4</sup>
<i>Enterobacter</i>	I2706, I2707, I2726, I2727, A2706, A2724, A2731	3.12x10 <sup>-5</sup>
<i>Microbacterium</i>	I2748	3.12x10 <sup>-6</sup>
<i>Paenibacillus</i>	I2746, I2747	3.12x10 <sup>-6</sup>
<i>Pseudomonas</i>	A2712, A2727	9.35x10 <sup>-6</sup>
<i>Olivibacter</i>	I2749	2.49x10 <sup>-5</sup>

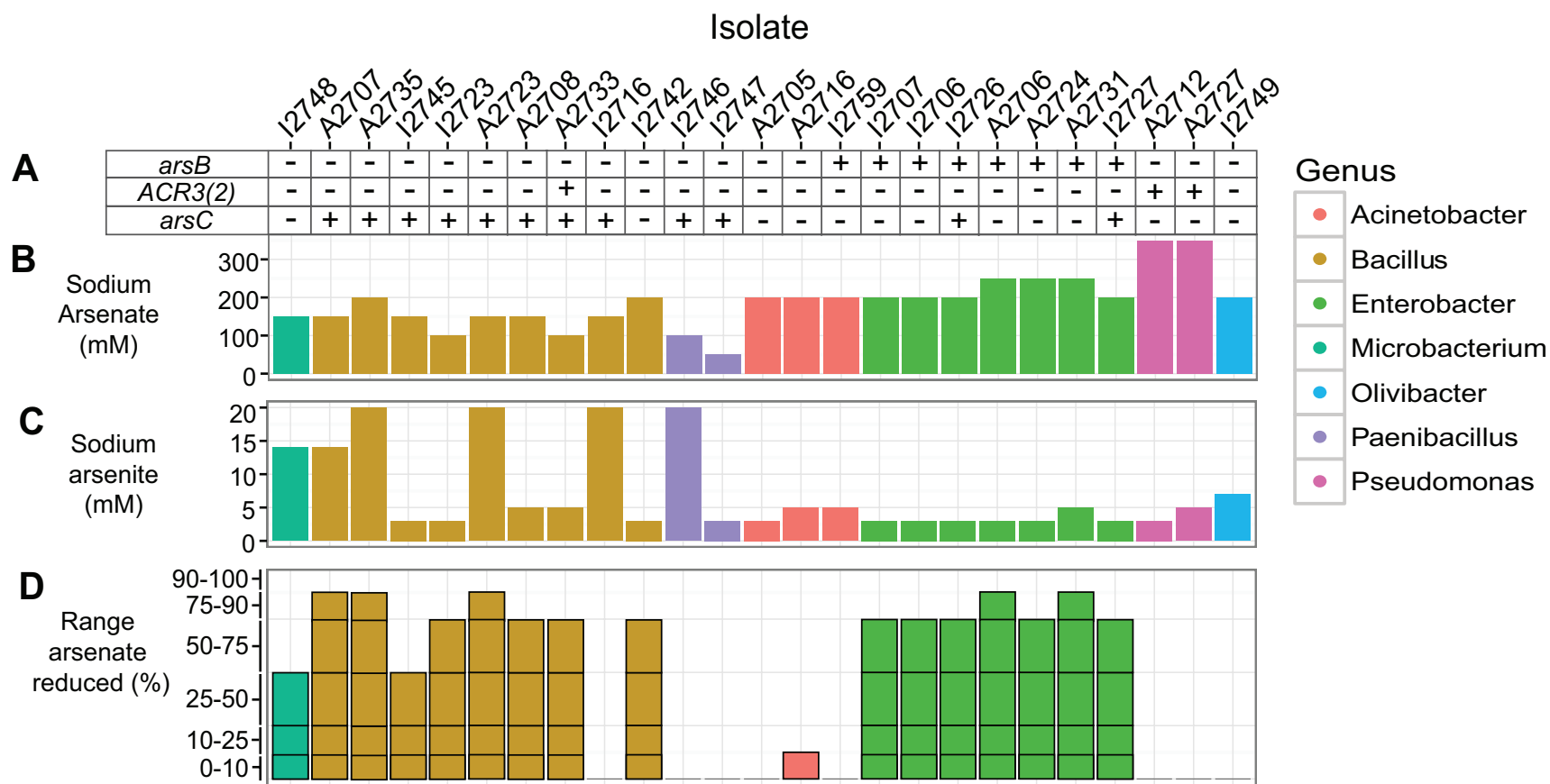


**Figure 2.1. Phylogenetic tree of 16S rRNA sequences from *Centralia* arsenic resistant isolates.**

Isolates from this study were compared with isolates from other studies that cultivated arsenic resistant isolates from soil. **A)** Actinobacteria, Proteobacteria, and

**Figure 2.1 (cont'd)**

Sphingobacteria. **B)** Firmicutes. Scale bars indicate the percent difference in nucleotide sequence.



**Figure 2.2. As resistance genotypes and phenotypes of isolated bacterial strains.**

A) Presence of arsenic resistance genes from end-point PCR are indicated (+). MICs of B) sodium arsenate and C) arsenite. D)

Categorical range of arsenate reduced based on standard curve of known ratios of arsenate and arsenite.

genus *Enterobacter* with the exception of one *Acinetobacter* isolate. Three isolates (12%) had the gene encoding arsenite efflux pump, *ACR3(2)*. Twelve isolates (48%) had the arsenate reductase gene, *arsC*. We did not find evidence for genes encoding other resistance mechanisms including dissimilatory arsenate reductase (*arrA*), arsenite oxidase (*aioA*), arsenite efflux pump (*ACR3(1)*), or arsenite methyltransferase (*arsM*) in the isolate collection. Thus, only genes related to arsenate reduction and arsenite extrusion were detected among these *Centralia* isolates using prominent primer sets. Notably, five isolates (20%) did not test positive for any arsenic resistance genes tested using published primers, suggesting sequence diversity of tested genes that are not captured with these primer sets, undescribed resistance genes, or resistance through general stress responses.

#### Arsenic transformation

We determined the abilities of isolates to transform arsenate and arsenite using a published semiquantitative measure of percent arsenic transformation without growth media (141). No isolates oxidized arsenite in this assay (data not shown). However, we observed a wide range of capabilities for arsenate reduction that generally corresponded to isolate taxonomy (**Figure 2.2D**). All isolates belonging to the genus *Enterobacter* had transformation capabilities at or above 50%. Isolates belonging to *Bacillus* had varied arsenate reduction capabilities ranging from 0-90%. The *Microbacterium* isolate (I2748) reduced 10-25% of arsenate in solution, and *Acinetobacter* isolates reduced 0-10% of arsenate. While nine isolates (36%) reduced arsenate *in vitro* and tested positive for *arsC*, there were discrepancies between the *in vitro* and genetic data. Isolates belonging to genera *Olivibacter*, *Paenibacillus*, and *Pseudomonas* did not reduce arsenate in this assay (**Figure 2.2D**). An additional three isolates (12%) tested positive for *arsC* but did not reduce arsenate in this assay. It is possible that *arsC* is nonfunctional in these



bacterial strains, not active in these conditions, or that arsenate reduction occurred but was below the limit of detection of this assay. Additionally, eight isolates (32%) reduced arsenate in this assay but did not test positive for the genes encoding arsenate reductases (*arsC* or *arrA*). These isolates may contain less characterized arsenate reductase genes (144).

#### Incongruent phylogenies of arsenic resistance and 16S rRNA genes

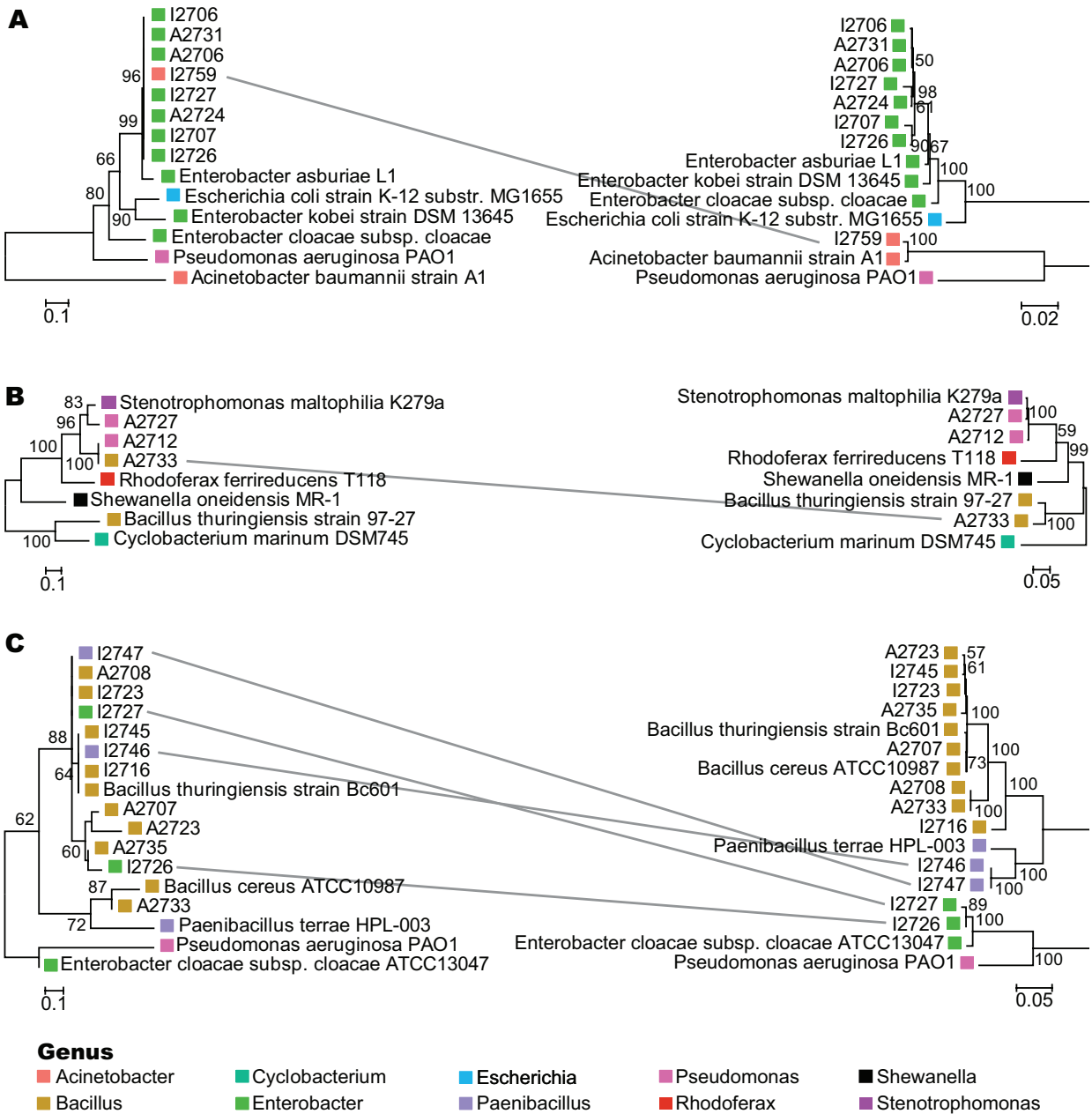
Maximum likelihood trees of detected arsenic resistance genes were compared with their corresponding 16S rRNA gene trees, and there was incongruence in all instances (**Figure 2.3**). All *arsB* sequences were related to *Enterobacter*, including those from an *Acinetobacter* isolate (**Figure 2.3A**). Three isolates spanning two genera (*Pseudomonas*, *Bacillus*) tested positive for *ACR3(2)*, and all had high sequence homology to *Stenotrophomonas*-derived *ACR3(2)* (**Figure 2.3B**). Comparing the *arsC* and 16S rRNA phylogenetic trees revealed several inconsistencies between gene sequence and phylogeny (**Figure 2.3C**). Twelve isolates spanning three genera (*Bacillus*, *Paenibacillus*, and *Enterobacter*) had high sequence homology to *Bacillus*-derived *arsC*, suggesting HGT. Closest NCBI BLAST hit and GC content for each arsenic resistance gene and corresponding taxa further suggested incongruence (**Appendix A Table 5**). Collectively, these data suggest past, and potential future, movement of these arsenic resistance genes via HGT.

#### MICs and growth phenotypes in arsenic

In parallel to characterization of genetic mechanisms of arsenic resistance, we determined the MICs of arsenate and arsenite for each isolate (**Figure 2.2BC**). MIC phenotypes ranged from 50 mM to >300 mM for sodium arsenate and from 3 to 20 mM for sodium arsenite. Both *Pseudomonas* isolates could withstand >300 mM sodium arsenate, which is typical for

previously reported pseudomonads resistant to arsenic (110, 145). High sodium arsenate resistance ( $>200$  mM) (87) was observed in 20% of the isolates. High sodium arsenite resistance ( $>15$  mM) (110) was observed in 16% of the isolates, all of which belong to phylum Firmicutes.

We also analyzed growth phenotypes (lag time, maximum growth rate, and maximum OD<sub>590</sub>) in arsenic, and our results highlight a nuanced relationship between growth in arsenic and taxonomy that was more informative than the observed MIC data alone (**Figure 2.4, Appendix B Figure 1– Appendix B Figure 4**). Limited conclusions can be made about *Paenibacillus*, *Microbacterium*, *Olivibacter*, and *Pseudomonas* isolates due to the small sample size ( $n \leq 2$ ) of these genera. Maximum growth rate ( $\mu$ ) and maximum OD<sub>590</sub> (A) showed similar patterns in each isolate, so we only report  $\mu$  here and provide A in supporting materials (**Appendix B Figure 2– Appendix B Figure 4**). In general, relative growth phenotypes were similar between arsenate and arsenite. Firmicutes isolates maintained basal growth rates in the presence of arsenic. Here we offer a qualitative description of the isolates' growth phenotypes in arsenic. More work will be needed to understand how general these growth phenotypes may be within lineages. While *Paenibacillus* isolates had the lowest MICs, they showed the least overall growth phenotype change in arsenic. *Bacillus* isolates, however, exhibited larger increases in lag time ( $\lambda$ ) as compared with *Paenibacillus* isolates. Conversely, the *Olivibacter* isolate showed an increase in lag time along with reductions in growth rate.



**Figure 2.3. Comparison of arsenic resistance gene sequences and 16S rRNA gene sequences from arsenic resistant isolates.**

Maximum likelihood trees for arsenic resistance genes (left) **A)** *arsB*, **B)** *ACR3(2)*, and **C)** *arsC* are shown alongside trees of corresponding 16S rRNA genes (right). Incongruence is highlighted with grey lines between the two trees. Scale bars indicate the percent difference in

**Figure 2.3 (cont'd)**

nucleotide sequence. Bootstrap values greater than 50% are indicated at the corresponding node, and boxes are colored based on isolate genus.

Again, because there was only one *Olivibacter* isolate, we cannot know how general its growth trends in arsenic are. Members of *Enterobacter* had reductions in growth rate as well as increased lag time with increasing arsenic concentrations despite their high MICs. Hierarchical clustering of growth phenotypes in genera with more than two isolates revealed clustering based on taxonomy rather than genotype or MIC (**Appendix B Figure 5**). Despite variability in  $\lambda$  in *Acinetobacter* isolates, they clustered apart from *Enterobacter* and *Bacillus* and had comparably higher values. Similarly, *Bacillus* strains clustered together despite variability in  $\mu$  observed within genus. Again, because we have limited representatives of *Paenibacillus*, *Pseudomonas* and *Microbacterium*, future studies should investigate the generality of their growth phenotypes in arsenic. These results suggest that, aside from the concentration of arsenic exposure, growth changes in lag time, rate, and maximum OD<sub>590</sub> may impact an isolate's survival outcomes *in situ*. More work is needed to determine if collective growth phenotype changes among arsenic resistant isolates within a soil community may be in part predicted by taxonomy and by occurrence of HGT.

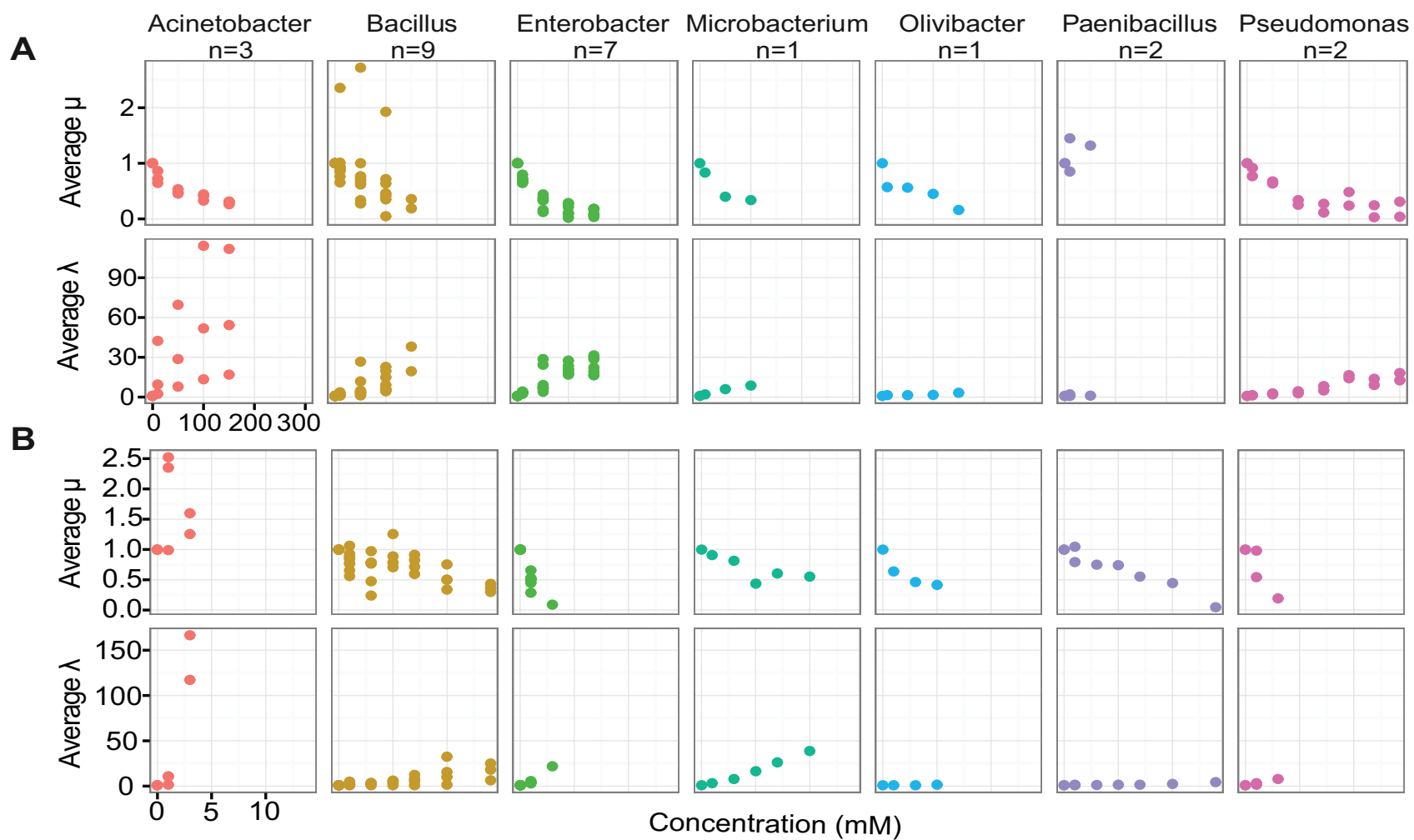


Figure 2.4. Growth phenotypes of isolates in increasing concentrations of arsenic.

**Figure 2.4 (cont'd)**

Lag time (  $\lambda$  ) and maximum growth rate (  $\mu$  ) of isolates in TSB50 with increasing concentrations of **A)** arsenate and **B)** arsenite normalized to growth without arsenic.

## Discussion

Our results from characterizing this modest isolate collection of arsenic resistant soil bacteria expose two considerations regarding the microbial community ecology of arsenic exposure. First, our data show that members of the rare biosphere harbor arsenic resistance genes that appear to be transferred via HGT in the past and therefore could have potential for transfer in the future. Second, our results suggest that nuanced growth phenotypes in arsenic may be predictable by the taxonomic identity of the microorganism that has not been described previously. This has implications for understanding a microbial community's response to arsenic, as it suggests there are differential growth responses, and therefore different competitive abilities, of resistant taxa. Thus, while the distribution and transfer of arsenic resistance genes in the microbial community have implications for filtering of community members given arsenic exposure, knowledge of arsenic growth phenotypes could be used to predict the compositional outcome (re-structuring) of an arsenic-exposed community; however, more work examining consistency of growth phenotypes in arsenic within and among lineages would inform the feasibility of such forecasting.

In this study, we described a collection of 25 aerobic arsenic resistant bacterial strains isolated from soils of an active vent from an underground coal seam fire in Centralia, PA, a unique terrestrial environment. We subsequently determined that, despite the fire activity at this particular site, the soil had relatively low arsenic concentrations at the time of soil collection (2.58 ppm). This is not surprising, given that 1) the fire is dynamic and past arsenic concentrations at the vent may have been higher given the natural occurrence of arsenic as a byproduct of coal combustion (97, 98) and 2) the widespread observation of microbial arsenic resistance from soils that have generally low contamination (29, 42, 87, 146, 147). While our



isolation resulted in an abundance of Firmicutes, this is not surprising since members of phylum Firmicutes have been shown to be resistant to arsenic previously with varied MICs (87, 148). Additionally, we acknowledge cultivation bias and that freezing soil prior to cultivation may have influenced our ability to resuscitate some strains (149). Accordingly, all 25 isolates were rare within their soil microbial community (**Table 2.1**). Previous studies have shown that cultivation from soil can isolate rare community members (121), but this investigation is the first specific documentation of enrichment of arsenic resistant bacteria from the rare biosphere. This finding is relevant to the Centralia community because soil arsenic concentrations may increase due to coal combustion (97, 98). While we cannot determine the response of the general community to additional arsenic deposition, our results suggest that members of the rare biosphere are capable of surviving arsenic stress and have potential to transfer resistance genes.

We also found that growth phenotypes in arsenic provided richer context for tolerance than MICs. Our results are consistent with previous reports that Proteobacteria often have high MICs of arsenic (**Figure 2.2B**) (30, 101); however, when simultaneously analyzing reductions in growth with arsenic, our results show distinct growth strategies among lineages, in both arsenate and arsenite (**Figure 2.4** and **Appendix B Figure 5**). While other reports have examined growth reduction in the presence of arsenic to find suitable strains for bioremediation (109, 146, 147, 150, 151), a suite of growth parameters are not typically investigated. Our full characterization of growth in increasing concentrations of arsenic showed a modest relationship between growth phenotype and taxonomy and highlights discrepancies between fitness in arsenic and MIC. This taxonomic delineation of growth phenotypes may be attributed to lineage-distinct mechanisms of arsenic tolerance; however, limited conclusions can be made for genera with small sample sizes (*Paenibacillus*, *Microbacterium*, *Olivibacter*, and *Pseudomonas*). Jobby and colleagues (102)

found an increased lag time with arsenic addition in an *Enterobacter* isolate from Navi Mumbai, which is similar to the lag times observed for *Enterobacter* isolates from Centralia, PA. This further implicates taxonomy as an important factor in an organism's tolerance to arsenic in liquid culture. Accounting for tolerance mechanisms may explain some of the discrepancies between MIC and arsenic resistance genotype (29) and between MIC and isolate abundance in contaminated sites (148). Valverde and colleagues (148) observed an increase in Firmicutes with increasing arsenic concentrations despite their lower MICs *in vitro*. Our findings suggest that arsenic resistant Firmicutes, in general, had modest changes in growth phenotypes in arsenic. Generally, this result questions the precision of MICs in predicting the success of a microorganism in the presence of arsenic. While this report is descriptive and not an exhaustive look at the relationship between growth phenotype in arsenic and taxonomy, consideration of both growth phenotype and taxonomy may offer additional predictive value and future studies should further examine growth phenotypes in arsenic.

Microbial arsenate reduction and the transfer of associated functional genes are important environmental health concerns because these processes increase the mobility of environmental arsenic (100). Incongruence between the phylogenetic alignment of *arsC*, *arsB*, and *ACR3(2)* and the 16S rRNA gene within this isolate collection suggests horizontal transfer of arsenic resistance genes (**Figure 2.3**), despite a low arsenic concentration and therefore low direct-selection pressure at this site. Determining the genetic environment of these arsenic resistance genes (chromosomal location or plasmid-borne) through whole genome sequencing would further determine whether these genes were horizontally transferred and provide insights into mechanisms of transfer. These results further emphasize the potential HGT of genes encoding arsenite efflux pumps and arsenate reductase seen previously (30, 34). Specifically, HGT of the

gene encoding an arsenite efflux pump (*arsB*) has been seen in environments with low arsenic concentrations (30). Notably, these data indicate potential HGT from multiple species, suggesting community-level contributions to arsenic resistance rather than a limited source of resistance genes. Investigating interactions among community members in the context of arsenic contamination may provide insights into the sources and sinks underlying the movement of resistance genes.

Finally, we observe multiple discrepancies between genetic and functional assays when characterizing the isolates' arsenic resistance. Despite using twelve published and commonly used primer sets to screen for arsenic resistance genes, three isolates with relatively high MICs did not test positive for any arsenic resistance genes screened in this study, highlighting a caveat of using primers for detection (29, 101). We also observe inconsistencies between genetic results and arsenate transformation capabilities, suggesting divergent gene sequences, presence of untested arsenic resistance genes (including the possibility of novel genes (111)), or general stress responses. A wider breadth of arsenic resistance gene diversity is likely to be captured using complementary cultivation-independent methods.

Our focus on growth phenotypes in arsenic revealed a relationship with taxonomy that has not been described previously. Additionally, our data show that rare community members can exhibit arsenic resistance and contain arsenic resistance genes. These observations have implications not only for arsenic tolerance but also for mechanisms supporting general microbial community robustness to arsenic stress.

### **CHAPTER 3 : Community structure explains antibiotic resistance gene dynamics over a temperature gradient in soil**

Work presented in this chapter has been published as Dunivin TK and Shade A. Community structure explains antibiotic resistance gene dynamics over a temperature gradient in soil. *FEMS Microbiology Ecology*. 2018; 94

## Abstract

Soils are reservoirs of antibiotic resistance genes (ARGs), but environmental dynamics of ARGs are largely unknown. Long-term disturbances offer opportunities to examine microbiome responses at scales relevant for both ecological and evolutionary processes and can be insightful for studying ARGs. We examined ARGs in soils overlying the underground coal seam fire in Centralia, PA, which has been burning since 1962. As the fire progresses, previously hot soils can recover to ambient temperatures, which creates a gradient of fire impact. We examined metagenomes from surface soils along this gradient to examine ARGs using a gene-targeted assembler. We targeted 35 clinically-relevant ARGs and two horizontal gene transfer-related genes (*intI* and *repA*). We detected 17 ARGs in Centralia: AAC6-Ia, *adeB*, *bla\_A*, *bla\_B*, *bla\_C*, *cmlA*, *dfra12*, *intI*, *sul2*, *tetA*, *tetW*, *tetX*, *tolC*, *vanA*, *vanH*, *vanX*, and *vanZ*. The diversity and abundance of *bla\_A*, *bla\_B*, *dfra12*, and *tolC* decreased with soil temperature, and changes in ARGs were largely explained by changes in community structure. We observed sequence-specific biogeography along the temperature gradient and observed compositional shifts in *bla\_A*, *dfra12*, and *intI*. These results suggest that increased temperatures can reduce soil ARGs but that this is largely due to a concomitant reduction in community-level diversity.

## Introduction

The dissemination of antibiotic resistance genes (ARGs) is a pressing public health concern. The One Health initiative recognizes the intrinsic link between evolution of bacterial resistance in clinical and environmental settings (152). Clinically relevant antibiotic resistance genes (ARGs) have been detected in “pristine environments” (153) as well as a variety of marine, plant, and soil microbiomes (81, 154–156). Soil is considered to be an environmental reservoir of ARGs, with greater ARG diversity than observed in the clinic (75). Despite our ability to easily detect ARGs in soil, the dynamics of soil ARGs are not fully understood (78). Understanding of the dissemination of ARGs in the environment is impeded by our limited information on their diversification, maintenance, and dissemination (157).

Investigating the propagation and dissemination of ARGs in soil is difficult because multiple interacting factors influence their fate (77, 78). Perhaps most obviously, ARGs can be selected when there is environmental exposure to antibiotic (79). Environmental exposure can result from the anthropogenic use of antibiotics, for example in agriculture or via wastewater treatment outputs (158, 159), or it can result from environmental antibiotic production by microorganisms *in situ* (75). Antibiotic exposure can kill sensitive populations and allow for propagation of resistant strains. Additionally, ARGs can be horizontally transferred (157) and are often detected on plasmids and other mobile genetic elements (66, 160). Thus, ARGs on mobile genetic elements may be disseminated more rapidly than through population growth alone. Furthermore, several ARGs are thought to have evolved >2 billion years ago (80), and these may be maintained in the absence of selective pressure from antibiotics and transferred vertically. Another complicating factor for understanding ARG dissemination is the influence of the dynamics of soil microbial communities. While interspecies competition can impact ARG

abundance, one study of many habitats showed that abiotic soil conditions can be important drivers of ARG profiles (81). Anthropogenic influences, such as nitrogen addition to the soil, also can impact ARGs (82). Similarly, studies with changing abiotic conditions, such as increased temperatures, have reported subsequent reductions in ARG abundance (83, 84). In these examples and others, environmental disturbance can alter soil microbial community structure (161–163), and then can impact local ARGs and their dissemination.

Long-term disturbances that impact multiple microbial generations can provide opportunities to investigate the changes in ARGs in response to environmental stress. One such disturbance is Centralia, PA, the site of an underground coal seam fire that ignited in 1962. Because this town was evacuated in 1984, it also represents a post-urban ecosystem of minimal contemporary anthropogenic influence. This fire continues to advance along the coal seam, creating a gradient of contemporary and historical fire impact and allowing for observation of multiple microbial generations' responses to disturbance and their potential recovery. Surface soil microbial communities in Centralia are exposed to elevated temperatures (21 – 57°C) (140) and coal combustion pollutants (97) which include trace elements such as arsenic, copper, aluminum, and lead (96, 97). While temperature increases are large, deposition of coal combustion pollutants occurs at a slow rate and varies based on the subsurface structure and geochemical properties of the burning coal (97). The depth of the coal seam varies from surface-level to 46 m (95). Furthermore, surface temperatures cool to ambient levels as the fire progresses, but coal combustion pollutants are not necessarily removed. Previously, we observed changes in bacterial and archaeal community structure with fire history that was well explained by temperature rather than soil properties such as arsenic concentration (140).

We leveraged the long-term disturbance in Centralia to examine ARG biogeography given both the abandonment of human habitation and the presence of a multigenerational stressor for the microorganisms. We investigated 12 metagenomes of microbial communities from surface soils along the Centralia temperature gradient for 35 clinically-relevant ARGs conferring resistance to eight classes of antibiotics, as well as multi drug efflux pumps and two HGT-relevant genes *repA* and *intI*. We used gene targeted assembly of the metagenomes to capture a breadth of ARG diversity. To examine the potential extent of HGT in Centralia, we asked whether changes in community structure explained any changes in ARG profiles. Because we previously identified changes in community structure along the stressor (140), we also asked whether functional redundancy (e.g., different ARG sequences belonging to the same resistance class) within the soil microbial community moderated the impact of a disturbance on ARG profiles. Functional redundancy allows for changes in community structure to occur without subsequent change in ARG abundance. Also, because we focused on clinically relevant ARGs rather than potentially novel ARGs from thermophilic lineages, we hypothesized that ARG abundance would decrease with temperature, as observed in other studies (Diehl and Lapara 2010; Qian *et al.* 2016; Tian *et al.* 2016). We were, however, also interested in the biogeography of specific gene sequences and hypothesized that they may have unique responses, even within the same resistance class.



## Materials and methods

### Reference Database construction

Reference gene databases of diverse, near full length sequences were constructed using selected sequences from FunGene databases (164) for the following genes: AAC6-Ia, *adeB*, *ANT3*, *ANT6*, *ANT9*, *bla\_A*, *bla\_B*, *bla\_C*, *CAT*, *cmlA*, *dfra1*, *dfra12*, *ermB*, *ermC*, *intI*, *mexC*, *mexE*, *qnr*, *repA*, *strA*, *strB*, *sul2*, *tetA*, *tetD*, *tetM*, *tetQ*, *tetW*, *tetX*, *tolC*, *vanA*, *vanC*, *vanH*, *vanT*, *vanW*, *vanX*, *vanY*, and *vanZ*. Seed sequences and Hidden Markov Models (HMMs) for each gene were downloaded from FunGene, and diverse protein and corresponding nucleotide sequences (reference sequences) were selected with gene-specific search parameters (**Appendix A Table 6**). Briefly, minimum size amino acid was set to 70% of the HMM length; minimum HMM coverage was set to 80% as is recommended by Xander software for targeted gene assembly (165); and a score cutoff was manually selected based on a notable score reduction between consecutive sequences, as suggested by the Ribosomal Database Project (personal communication). Reference sequences were de-replicated before being used in subsequent analysis, and final sequence numbers are included in **Appendix A Table 6**.

### Sample collection, sequencing, and quality control

Study site, soil sampling and soil biogeochemistry were all performed as described (140). Briefly, surface soils were sampled along a gradient of fire-impact that was determined from historical characterizations of the site (95): fire-affected (n = 6), recovered (n = 5), and reference (n = 1). Fire-affected soils had elevated temperatures due to fire; recovered soils were at ambient temperature but historically had elevated temperatures from the fire; and the reference soil was never impacted by the fire. The reference sample was used as a qualitative control and is not intended as an quantitative and definitive comparison to non-impacted soils. Microbial

community DNA was obtained using a phenol chloroform extraction (Cho et al., 1996) and purification with MoBio DNEasy PowerSoil kit without vortexing. All samples were sequenced on the Illumina HiSeq 2500 platform with 2 x 150 bp paired end format at the Joint Genome Institute (JGI) and quality filtered using BBduk (<https://sourceforge.net/projects/bbmap/>). Metagenome coverage was estimated using Nonpareil (166).

### Gene targeted assembly and quality control

A gene targeted metagenome assembler (165) was used to assemble antibiotic resistance genes of interest from quality-filtered metagenomes. For each gene of interest, seed sequences, HMMs, and reference gene databases, as described above, were included. The *rplB* reference gene database, seed sequences, and HMMs from the Xander package were used. In most instances, default assembly parameters were used, except to incorporate differences in protein length (i.e. if the protein was shorter than 150 aa (default), as was the case for *dfra1*, *dfra12*, *AAC6-Ia*, *ermB*, *ermC*, *qnr*, *vanX*, and *vanZ*) (**Appendix A Table 6**). While the assembler includes chimera removal, additional quality control steps were added. Specifically, final assembled sequences (contigs) were searched against the reference gene database as well as the non-redundant database (nr) from NCBI (August 28, 2017) using BLAST (v. 2.2.26,(167)). Genes were re-examined if the top hit had an e-value  $> 10^{-5}$  or if top hit descriptors were not the target gene. Genes with low quality results were re-assembled with adjusted parameters. Aligned sequences from each sample were dereplicated and clustered at 90, 97, and 99% amino acid identity using the RDP Classifier (168). Our quality control analyses can be accessed on GitHub ('assembly\_assessments' repository in [https://github.com/ShadeLab/PAPER\\_Dunivin\\_Antibiotics\\_2017/tree/master/assembly\\_assessments](https://github.com/ShadeLab/PAPER_Dunivin_Antibiotics_2017/tree/master/assembly_assessments)).

### Ecological analyses

Phylum-level *rplB* relative abundance was used to examine differences in community structure. Relative abundance for each site was averaged among samples of the same fire classification (i.e. fire-affected, recovered, reference) and compared to 16S rRNA gene sequence data from a previous work (140). For subsequent ecological analyses, the RDP Classifier was used to generate an OTU table from 90, 97, and 99% amino acid identities. We refer to contigs clustered at 99% identity as “ARG sequences” throughout the remainder of the text. The OTU tables were analyzed in R (138). OTU tables were separated based on the gene(s) of interest (*rplB* and ARGs). Due to Nonpareil-estimated differences in coverage, *rplB* and ARG OTU tables were rarefied to an even sampling depth (258 and 180 assembled sequences, respectively) using the vegan package (169). Pielou’s evenness was calculated, and richness was estimated using PhyloSeq (170). The Psych package was used to calculate Spearman’s rank correlations between alpha diversity (richness and evenness) and soil temperature for both *rplB* and ARGs. Bray-Curtis distance was used to obtain dissimilarity matrices, and principal component analysis was used to visualize beta diversity. Distance matrices of rarefied, relativized data were analyzed using Mantel tests with Spearman’s rank correlations. Mantel tests were performed on *rplB*, ARG, and spatial distance matrices of sample locations.

### Resistance gene comparison

We assessed ARG biogeography at the gene, taxonomic class, and sequence levels. To compare the abundance of ARGs among data sets, total counts of *rplB* were used to normalize the abundance of each ARG sequence. Total counts of each ARG were calculated as the sum of the relative abundance of each ARG sequence. The Psych package (171) was used to calculate Spearman’s rank correlations between soil geochemical properties and total gene counts for each

ARG. Pairwise correlations for the total abundance of each resistance gene were also calculated. For taxonomic analysis of each ARG, the top BLAST result and the taxize package (172) were used to assign taxonomy to each ARG sequence. When the top hit was an uncultured bacterium, the second or third hit was used, and when all three top hits were unknown, the taxonomy was labeled unknown. Total counts of each taxonomic class were summed for each ARG, and Spearman's rank correlations were used to test for correlations between class abundance and temperature for all ARGs with representatives from at least three taxonomic groups. Spearman's rank correlations were performed on normalized and relativized abundance information, but only relativized abundance is shown because it agreed with normalized data and also had unique features. Furthermore, we examined biogeography of individual ARG sequences. A Venn analysis was performed between ARGs in fire affected and recovered samples using the VennDiagram package (173). The mean normalized abundance for each ARG sequence among samples was plotted against the number of sites it was observed in (occurrence). ARG sequences present in only one site were subsequently removed, and we used hierarchical cluster and heatmap analysis with the pheatmap package (174) to examine similar sequence biogeography along the temperature gradient.

#### Reproducibility, code, and data

Our computing workflows and R script can be accessed on GitHub ([https://github.com/ShadeLab/PAPER\\_Dunivin\\_Antibiotics\\_2017](https://github.com/ShadeLab/PAPER_Dunivin_Antibiotics_2017)). Metagenomes are available from IMG/GOLD study ID: Gs0114513.

## Results and Discussion

### Soil samples and gene targeted assembly

We previously collected soils along the Centralia temperature gradient (140). We submitted DNA extracted from twelve soils (temperature range = 12.1 - 54.2°C) to the Joint Genome Institute for a small-scale Community Science Project; we did not submit all 18 originally collected samples because there was a 12-sample limit with the small-scale award, and so we chose samples for sequencing that were representative of the thermal gradient. We sequenced metagenomes from soils that had elevated temperatures due to the fire (fire-affected,  $n = 6$ ), those that were historically impacted (recovered,  $n = 5$ ), and those with no documented impact (reference,  $n = 1$ ) (**Appendix B Figure 6**). Quality filtered metagenome size ranged from 21-51 Gbp, and Nonpareil-estimated coverage (166) varied from 29.12 to 89.96% (**Appendix A Table 7**). Though we measured a suite of geochemical data (**Appendix A Table 8**), our previous work found temperature to be the strongest driver of community structure (140), and we found that ARGs only correlated with temperature (**Appendix A Table 9**).

We used a gene-targeted metagenome assembler to probe Centralia metagenomes for ARGs. While this gene-centric methodology does not permit analysis of entire gene cassettes or flanking regions, it improves detection of low abundance genes, increases the length of assembled gene sequences, and is capable of detecting strain-level sequence variation (165). In addition to assembling ARGs of interest, we assembled *rplB*, a single copy gene and phylogenetic marker. We recently reported changes in community structure in surface soils along the Centralia coal seam fire, and this conclusion was based on analysis of 16S rRNA gene amplicon data (140). In this work, we used *rplB* community structure to compare ARG profiles because both were determined by the same annotation and assembly methods from shotgun

metagenomes. Thus, we first asked whether patterns observed using *rplB* sequences were similar to patterns we observed previously with 16S rRNA gene amplicons. Overall, patterns in community structure were consistent between these analyses (**Appendix B Figure 7**). This was verified based on significant Mantel tests between *rplB* and 16S rRNA genes (Mantel's  $r = 0.5877$ ,  $p = 0.001$  on 999 permutations, at the OTU level. There was no relationship between spatial proximity of soils and *rplB* community structure (Mantel's  $r = -0.14$ ,  $p > 0.05$  on 999 permutations), confirming our previous report that community structure is not strongly driven by local dispersal. *rplB* evenness was negatively correlated with temperature ( $\rho = -0.66$ ;  $p < 0.05$ ), and *rplB* richness also trended negatively ( $\rho = -0.55$ ;  $p = 0.05$ ). Decreased alpha diversity with increased temperature was expected because of the complex and extreme fire stressor (e.g., exposure to high temperature and coal combustion pollutants, Janzen and Tobin-Janzen 2008), and, again, is in agreement with our previous study (Lee and Sorensen et al. 2017). The only obvious difference was that the *rplB* dataset had a greater abundance of Firmicutes than the 16S rRNA gene dataset, which may be due to differences in DNA extraction methods (175) or marker gene target. Thus, we found that *rplB* assembled using these methods was comparable to 16S rRNA gene data (**Appendix B Figure 7**), showing that gene targeted assembly produced results consistent with previous work.

#### *Detected ARGs and changes in their abundance with temperature*

We examined a suite of genes encoding resistance to aminoglycosides, beta-lactams, chloramphenicol, sulfonamides, tetracyclines, trimethoprim, and vancomycin, as well as plasmid-related and genes encoding multidrug efflux pumps (**Table 3.1**). From Centralia metagenomes, we assembled 1,165 unique ARG clustered at 99% amino acid identity. Though

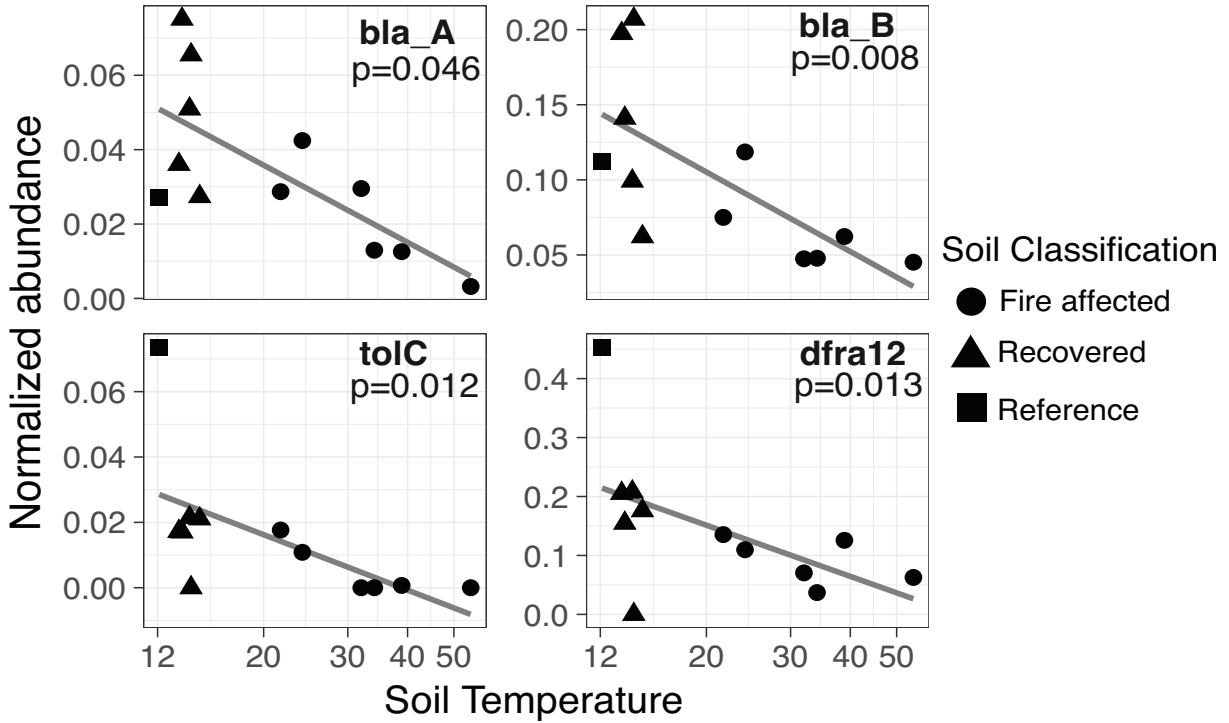
**Table 3.1. Resistance genes tested in this study.**

<b>Antibiotic specificity</b>	<b>Gene</b>
Aminoglycoside	AAC6-Ia, ANT3, ANT6, ANT9, strA,B
$\beta$ -Lactams	Class A (bla_A), Class B (bla_B), Class C (bla_C)
Chloramphenicol	CEP, cmlA
Macrolide	ermB,C, qnr
Multidrug efflux	adeB, mexC,E, tolC
Plasmid	intl, repA
Sulfonamide	sul2
Tetracycline	tetA,D,M,Q,W,X
Trimethoprim	dfra1, dfra12
Vancomycin	vanA,C,H,T,W,X,Y,Z

we targeted 35 distinct types of ARGs and two HGT-related genes, only 17 of these could be assembled from Centralia metagenomes. The genes *ANT3*, *ANT6*, *ANT9*, *CAT*, *dfra1*, *ermB*, *ermC*, *mexC*, *mexE*, *qnr*, *repA*, *strA*, *strB*, *tetD*, *tetM*, *tetQ*, *vanC*, *vanT*, *vanW*, and *vanY* were not observed, suggesting that they were either below the detection limit or absent. For detected ARGs, we found positive correlations between *vanA*, *H*, and *X* genes and between *tolC* and *dfra12* (**Appendix B Figure 8**). *vanAHX* genes are known to be associated with one another in VanA-type operons (176), and genes *tolC* and *dfra12* have previously been observed in isolates (177). While *sul2* and *intII* have been previously shown to be correlated (178), we did not observe a significant correlation between these genes. This discrepancy could be because our analysis does not distinguish between integron classes. Several ARGs in Centralia were negatively correlated with soil temperature (**Figure 3.1; Appendix A Table 9**), but no ARGs were correlated with other measured soil geochemical properties (results not shown, **Appendix A Table 8**). The most abundant ARGs detected in Centralia were *adeB*, *bla<sub>B</sub>*, and *dfra12* (**Figure 3.1, Appendix B Figure 9**). We note that the highest ARG normalized abundance was typically in Cen04 (13.3°C) but that this is due to low *rplB* abundance in the sample.

Our results are generally in agreement with other studies of ARGs in soils. For example, Fitzpatrick and Walsh (2016) also reported low abundance or absence of *qnr*, *tet* and *van* genes in soil. Several studies also reported that genes encoding dihydrofolate reductases and/or beta-lactamases were abundant in soils (67, 82, 154). Previous studies reported reductions in clinically-relevant ARGs with increased temperatures in digesters and compost (83, 84, 179). Diehl and Lapara (2010) observed a negative relationship between temperature and genes encoding tetracycline resistance and class 1 integrons in anaerobic digesters, but not aerobic ones. This may be further relevant to Centralia soils, as there likely are pockets of anaerobic





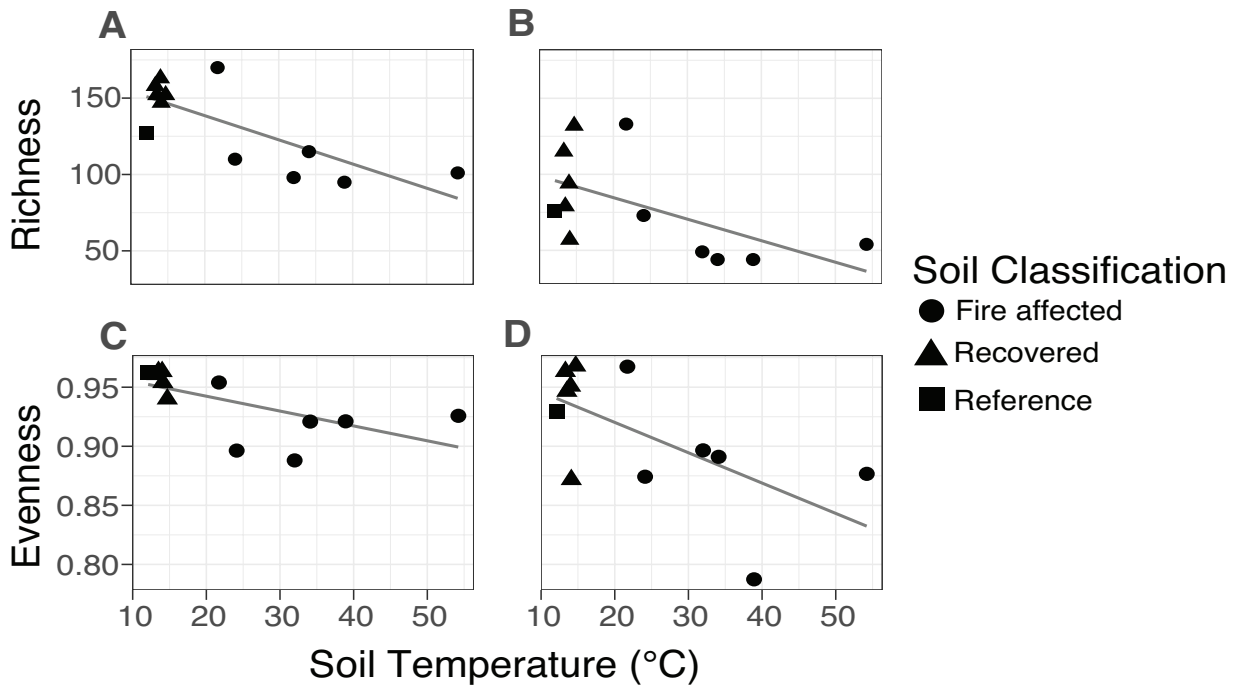
**Figure 3.1. Negative correlations between normalized abundance of ARGs and soil temperature.**

Coverage-adjusted abundance for *bla\_A*, *bla\_B*, *tolC*, and *dfra12* was normalized to total abundance of the single copy gene *rplB*. Normalized abundance is plotted against soil temperature. Note the differences in y-axes. The linear trend line and p-value corresponding to the Spearman's rank correlation are shown. Symbol indicates soil classification based on fire history.

activity in hot soils, especially at venting sites, which have measurably higher percent moisture content due to steam escaping (**Appendix A Table 8**). To our knowledge, this is the first description of a reduction in ARG abundances with temperature *in situ* with soil. These results suggest that ARGs may be reduced in soil environments by increasing temperature. Thus, we speculate that increases in temperatures expected to reduce microbial community diversity may result in decreased clinically relevant ARGs in the environment.

### Diversity of ARGs

We also examined the amino acid-level diversity of ARGs in *Centralia* metagenomes. We tested sequence cutoffs of 90, 97, and 99% amino acid identity, but overarching patterns did not vary based on sequence cutoff (results not shown). Thus, our subsequent diversity analysis applied the most stringent cutoff (99% amino acid identity), as was applied in the original gene targeted assembly paper (165). ARG richness was negatively correlated with temperature ( $\rho = -0.57$ ;  $p < 0.05$ ), but evenness had a variable response with temperature ( $\rho = -0.47$ ;  $p > 0.05$ ) (**Figure 3.2BD**). ARG alpha diversity (within-sample) trends were thus similar to *rplB* and 16S rRNA gene diversity trends (**Figure 3.2AC**), highlighting the influence of community structure on soil ARG profiles. In addition, overall differences in the composition of ARGs among sites were related to differences in *rplB* community structure (Mantel's  $r = 0.54$ ;  $p < 0.05$  on 999 permutations; **Appendix B Figure 10**). This result also supports the proposal that compositional shifts in membership among *Centralia* sites were driving the observed differences in ARGs, not propagation of ARGs by gene transfer. These results agree with a recent analysis that reported congruence between community structure and ARG profiles in soils (82). Similar to patterns in



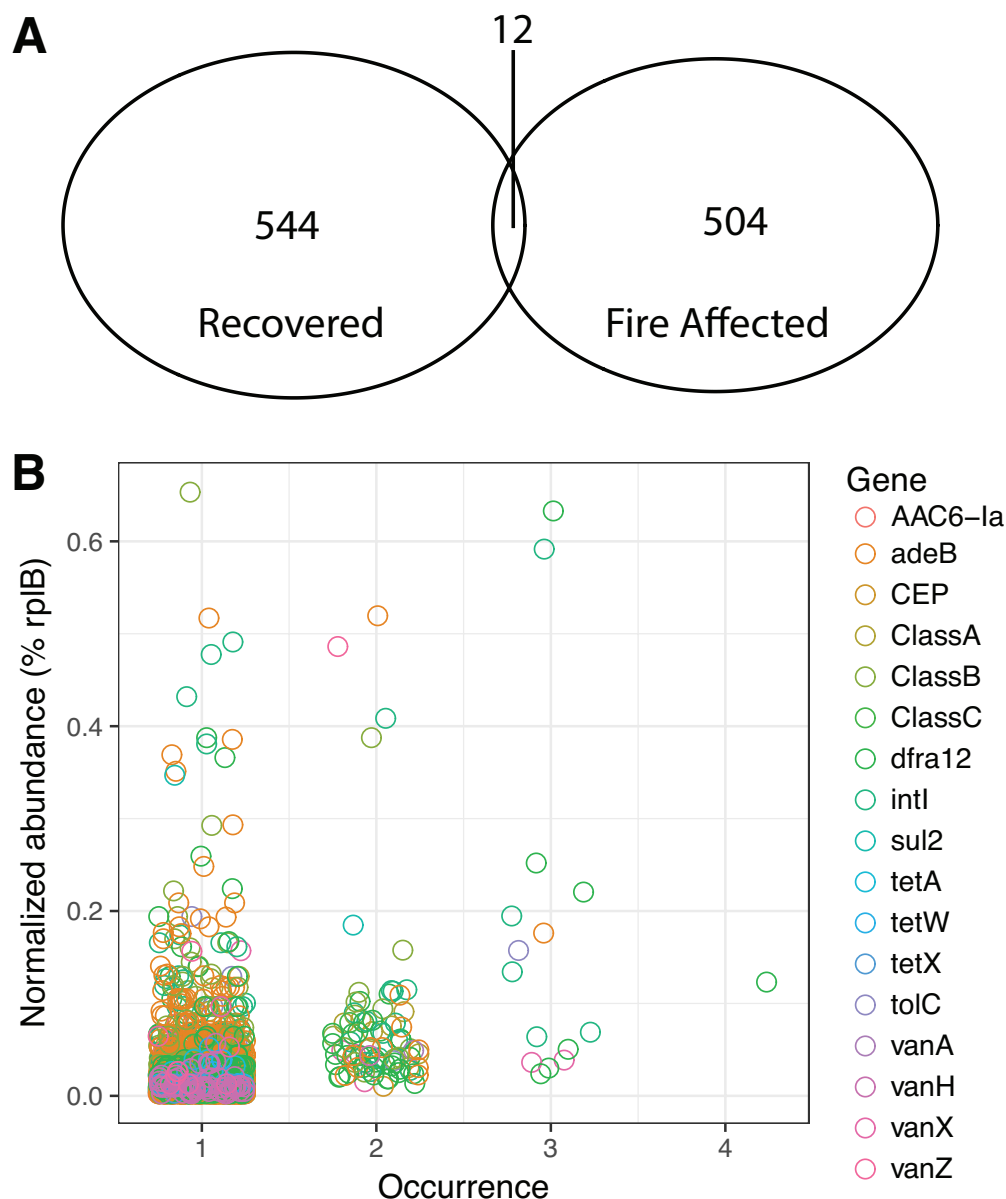
**Figure 3.2. Observed richness (AB) and evenness (CD) of *rplB* (AC) and ARG (BD) along the *Centralia* temperature gradient.**

Assembled sequences were clustered at 99% amino acid identity and rarefied to an even sampling depth. Observed number of sequences (richness) and Pielou's evenness is plotted against soil temperature. Symbol indicates soil classification based on fire history.

*rplB* and 16S rRNA genes, ARG profiles could not be explained by distance between sample sites (Mantel's  $r = 0.01$ ,  $p > 0.05$  on 999 permutations). This result suggests that local dispersal of ARGs, which could be indicative of HGT, is not a common mechanism of ARG dissemination in this system. However, when we considered fire-affected and recovered metagenomes separately, we found that *rplB* community structure explained ARG composition in fire-affected soils (Mantel's  $r = 0.71$ ;  $p < 0.05$  on 719 permutations), but not in recovered soils (Mantel's  $r = 0.30$ ;  $p > 0.05$  on 119 permutations). We determined that this result was not driven by one anomalous sample by performing iterative "leave-one-out" Mantel tests with four of five recovered soils, and all tests showed no correlation between *rplB* and ARGs (results not shown). The reason for no relationship between *rplB* and ARG in recovered soils is unclear (one hypothesis is that there is no signal given higher diversity), but this observation very indirectly suggests a potential larger influence of HGT in recovered soils than fire-affected soils that could be explored in future work.

#### ARG distribution and sequence-specific biogeography

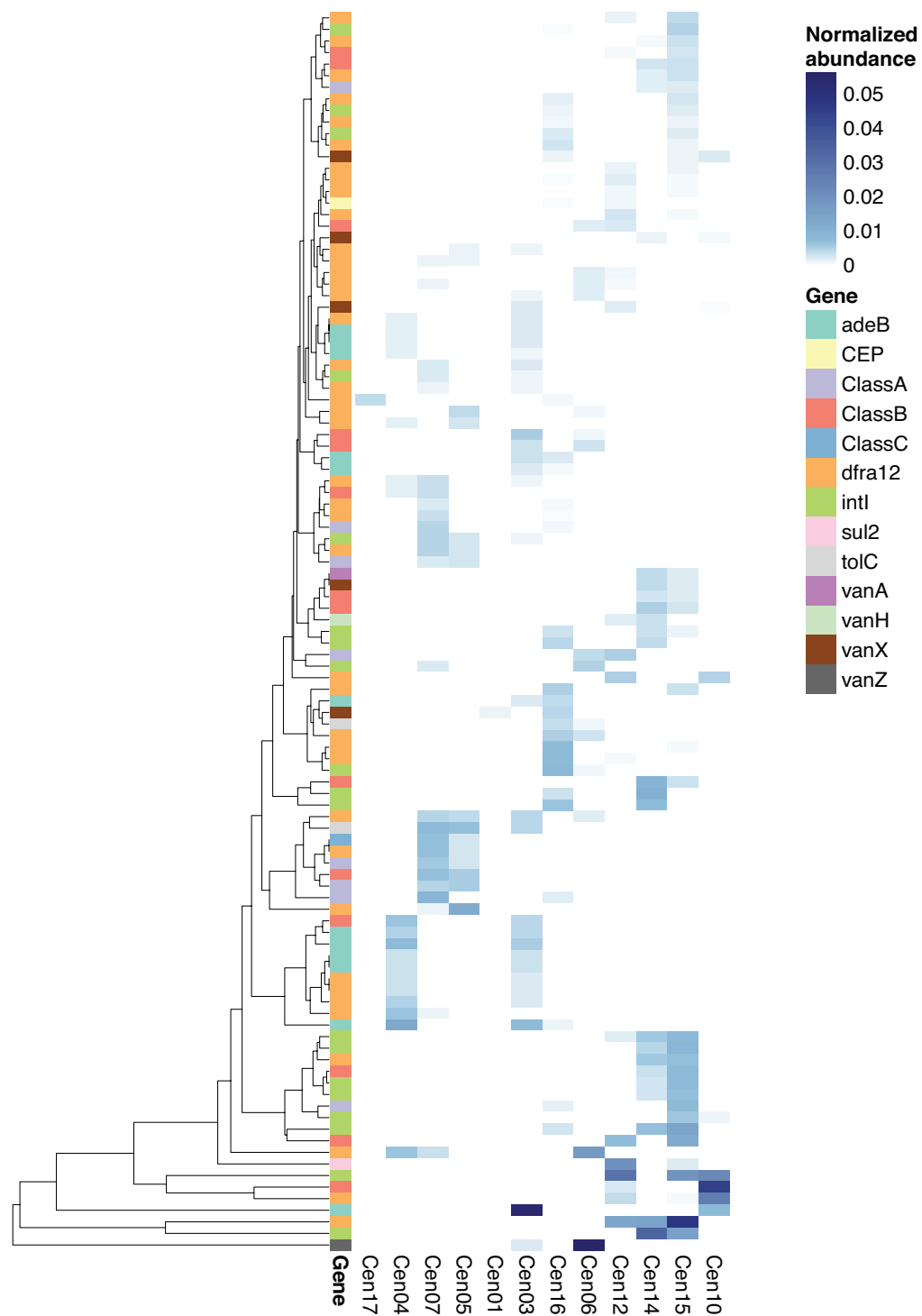
Only twelve ARG sequences were shared between fire-affected and recovered soils (**Figure 3.3A**). On one hand, this is expected because soils are heterogeneous and have high ARG diversity (154). Forsberg and colleagues (2014) observed 2,895 ARG sequences in a functional antibiotic resistance screen from 18 agricultural and grassland soils. Of these, only 2.6% were present in two or more soils, which is comparable to our data (1.1%). Similarly, the distinction between fire-affected and recovered soil in our study is in part explained by generally high ARG diversity, with minimal overlap of ARG sequences detected between all sites. Furthermore, most ARG sequences (94.16%), whether they were rare ( $< 1.5\%$  normalized abundance to *rplB*) or prevalent, were detected only in one metagenome (**Figure 3.3B**). Though



**Figure 3.3. Presence of ARG sequences in *Centrاليا* metagenomes.**

**A)** Venn diagram of ARG sequences observed in recovered and fire-affected soils. **B)** ARG abundance-occurrence patterns in *Centrاليا* metagenomes. Percent normalized abundance of ARG sequences was averaged among 12 metagenomes and plotted against the number of sites in which each sequence occurs in. Each point represents one cluster, and color indicates gene.

the gene-targeted assembly approach maximizes observation of diversity given metagenome coverage, it is possible that even greater coverage of these metagenomes could result in detection of more shared ARG sequences between samples. There were 13 distinct biogeographical dynamics that indicated genes sensitive to the fire, and these were classified into two categories based on their prevalence and patterns of detection: abundant-transient, and rare-transient sequences (**Figure 3.4**). Abundant-transient ARG sequences belonged to genes *adeB*, *bla\_B*, *dfra12*, *intI*, *sul2*, and *vanZ*. These sequences had a *rplB*-normalized abundance of  $\geq 1.5\%$  of the total community within at least one metagenome. Rare-transient biogeographic patterns were observed for ARG sequences belonging to *adeB*, *bla\_A*, *bla\_B*, CEP, *dfra12*, *intI*, *tolC*, *vanA*, *vanX*, and *vanH*. Rare-transient sequences represented those with  $\leq 1.5\%$  of the total community. However, step-wise relationships with temperature were observed for several ARG sequences, suggesting the potential enrichment by fire for microbes harboring these ARG sequences. Two clusters of rare-transient sequences with no temperature relationship were observed based on differences in normalized abundance (**Figure 3.4**), suggesting that they had no relationship with fire or temperature. Thus, we observed sequence-specific biogeography for ARG sequences along the temperature gradient, showing that the average changes in ARG abundance does not always fully explain the biogeography of each unique resistance gene sequence detected within that gene family.



**Figure 3.4. Normalized abundance of ARG sequences in Centralia metagenomes.**

Abundance of each gene sequence (clustered at 99% amino acid identity) present in  $\geq 2$

**Figure 3.4 (cont'd)**

metagenomes was normalized to *rplB*. Complete-linkage clustering was calculated with the *rplB*-normalized abundance of each ARG sequence. Heatmap shows normalized abundance on a blue scale. Soil sites (column) are ordered by increasing soil temperature. Each row represents one ARG sequence, and ARG is noted by color.

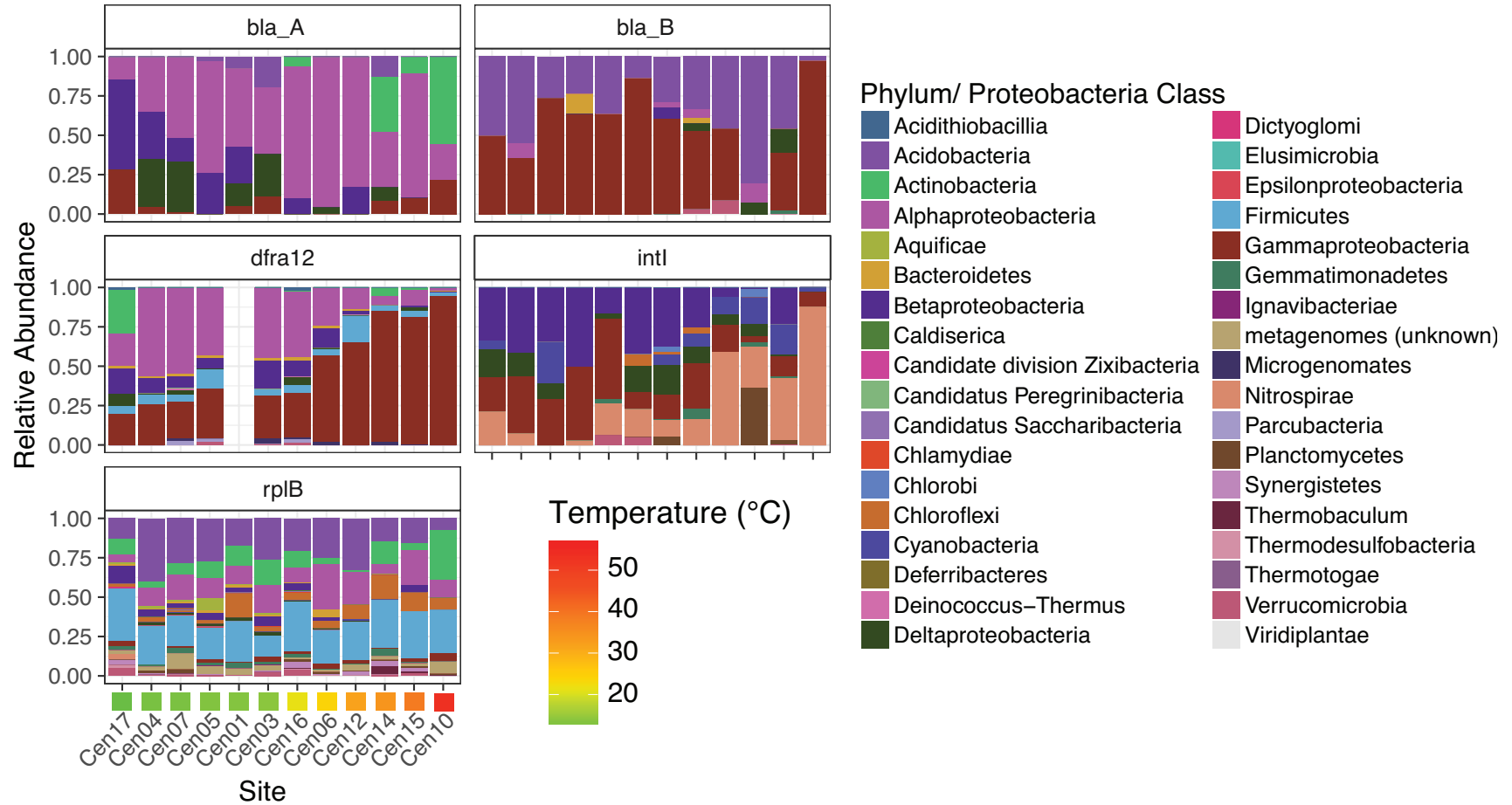


### ARG Compositional shifts

We examined both *rplB*-normalized and relativized abundance patterns to compare changes in composition of ARGs and changes in proportional contributions of ARGs. For this analysis, composition was considered at the phylum or *Proteobacteria* class levels based on top BLAST hits. For ARGs that represented more than three phyla or *Proteobacteria* classes, (*bla\_A*, *bla\_B*, *dfra12*, *intI*; **Appendix A Table 10**; **Appendix A Table 11**), we explored for correlations with temperature and observed changes in ARG composition with temperature for *bla\_A*, *dfra12*, and *intI* (**Figure 3.5**).

Generally, community structure was associated with ARG composition. *rplB*-level reduction in *Betaproteobacteria* corresponded with reductions in *Betaproteobacteria*-related ARG. *Betaproteobacteria*-related *bla\_A* and *dfra12* genes decreased with temperature (**Figure 3.5**; **Appendix A Table 11**). Thus, reductions in total *bla\_A* and *dfra12* counts is largely explained by a reduction in *Betaproteobacteria*. This pattern does not extend to *bla\_B* since *Betaproteobacteria*-related *bla\_B* genes were only detected in one soil (Cen16). We did not detect changes in *Gammaproteobacteria* based on *rplB*. This corresponded with consistent relative abundances of *Gammaproteobacteria*-related *bla\_A*, *bla\_B*, *dfra12*, and *intI* (**Appendix A Table 11**). *Gammaproteobacteria*-related *dfra12* increased in relative abundance with soil temperature ( $\rho = 0.95$ ,  $p < 0.05$ ), further highlighting that a reduction in total *dfra12* relative abundance is not due to changes in *Gammaproteobacteria*-related sequences. Phylum-level community structure, therefore, corresponded with compositional changes in ARGs, highlighting the influence of the underlying community on soil ARGs.

We observed evidence for functional redundancy of ARGs in Centralia through



**Figure 3.5. Relative abundance of taxonomically similar ARGs.**

Phylum-level taxonomy for *bla\_A*, *bla\_B*, *dfra12*, *intl*, and *rplB* for each site is shown. Color indicates phylum- and *Proteobacteria* class-level taxonomy of ARGs, and sites are ordered by increasing soil temperature. *dfra12* was not detected in Cen01.

compositional shifts along the temperature gradient. Total *bla<sub>A</sub>* relative abundance decreased with temperature (**Figure 3.1**); however, taxonomic groups of *bla<sub>A</sub>* were differentially impacted along the temperature gradient (**Figure 3.5; Appendix A Table 11**). Both normalized and relativized abundance of *Actinobacteria*-related *bla<sub>A</sub>* genes increased ( $\rho > 0.6$ ,  $p < 0.05$ ) while *Betaproteobacteria*-related *bla<sub>A</sub>* genes decreased ( $\rho < 0.6$ ,  $p < 0.05$ ) with temperature (**Appendix A Table 11**). Thus, fire impacted the abundance and composition of *bla<sub>A</sub>*. A decrease in total *bla<sub>A</sub>* (**Figure 3.1**) was accompanied by an increase in *Actinobacteria*-related *bla<sub>A</sub>*. This asymmetric response with temperature suggests an impact of functional redundancy on soil ARG profiles. We also observed a shift in *intI* composition despite consistent *intI* abundance along the temperature gradient. The relative abundance of *Beta*- and *Gammaproteobacteria*-related *intI* decreased with temperature ( $\rho < 0.6$ ,  $p < 0.05$ ), but the relative abundance of *Nitrospirae*-related *intI* increased with temperature ( $\rho > 0.6$ ,  $p < 0.05$ ) (**Figure 3.5; Appendix A Table 11**). We therefore observed changes in composition of *intI* with fire despite a lack of change in total *intI* abundance. Notably, previous studies have described *Nitrospirae*-related *intI*. Oliveira-Pinto and colleagues (2016) isolated an *intI* gene cassette related to *Nitrospirae* from a metal-rich stream, and Goltsman and colleagues (2009) identified both integrase and ARGs on chromosomes of *Nitrospirae* strains isolated from acid mine drainage. It is unclear, however, whether *Nitrospirae*-related *intI* genes are associated with ARG transfer. As *intI* encodes for a DNA integrase, this result suggests that *Nitrospirae* might contribute more to HGT in fire affected soils, but we cannot determine whether this putative gene transfer would include ARGs. We posit that reductions in ARG abundance due to increased temperature could increase subsets of clinically relevant ARGs, and studies using temperature as a control for ARGs should consider sequence-level ARG dynamics within the system.

## Conclusions

This case study of ARG biogeography over a long-term, severe thermal disturbance demonstrates the importance of community structure on soil ARG abundance and composition. Despite the stressor and the withdrawal of human activity, the diversity of ARG observed in *Centuria* is comparable to other soil systems (82, 154). For several clinically relevant ARGs, we observed a reduction in total abundance with increased temperature. While this has been reported in anthropogenic systems (83, 84, 179), we further probed *Centuria* datasets for compositional and sequence-specific ARG biogeography and found nuanced results. Generally, the reduction in ARG abundance could be explained by indirect effects (i.e. compositional shifts in the community). We posit that increased temperatures could result in a reduction in the diversity and abundance of ARGs in the environment, but our data also suggest that this reduction will not impact all ARG sequences similarly. ARG biogeographical dynamics in soil are thus largely dependent on community structure, which may also drive the observed fine-scale abundance-occurrence patterns.

## **CHAPTER 4 : RefSoil+: A reference database for genes and traits of soil plasmids**

Work presented in this chapter has been published as Dunivin TK, Choi J, Howe AC, and Shade A. 2019. RefSoil+: A reference database for genes and traits of soil plasmids. *mSystems*. 4(1) e00349-18

## **Abstract**

Plasmids harbor transferable genes that contribute to the functional repertoire of microbial communities, yet their contributions to metagenomes are often overlooked. Environmental plasmids have the potential to spread antibiotic resistance to clinical microbial strains. In soils, high microbiome diversity and high variability in plasmid characteristics present a challenge for studying plasmids. To improve understanding of soil plasmids, we present RefSoil+, a database containing plasmid sequences from 922 soil microorganisms. Soil plasmids were relatively larger than other described plasmids, which is a trait associated with plasmid mobility. There was a weak relationship between chromosome size and plasmid size and no relationship between chromosome size and plasmid number, suggesting that these genomic traits are independent in soil. We used RefSoil+ to inform the distributions of antibiotic resistance genes among soil microorganisms as compared to non-soil microorganisms. Soil-associated plasmids, but not chromosomes, had fewer antibiotic resistance genes than other microorganisms. These data suggest that soils may offer limited opportunity for plasmid-mediated transfer of described antibiotic resistance genes. RefSoil+ can serve as a reference for the diversity, composition, and host-associations of plasmid-borne functional genes in soil, a utility that will be enhanced as the database expands. Our study improves understanding of soil plasmids and provides a resource for assessing the dynamics of the genes that they carry, especially genes conferring antibiotic resistances.

## Introduction

Soil harbors immense microbial biodiversity, and the soil microbiome has functional consequences for ecosystems, like supporting plant growth (182, 183) and mediating key biogeochemical transformations (184). It also serves as a reservoir of microbial functional genes of interest to human and animal welfare. Within microbial genomes, important functions can be encoded on both chromosomes and extrachromosomal mobile genetic elements such as plasmids. Plasmids can be laterally transferred among community members, both among and between phyla (93, 119, 185). This capability causes the propagation of plasmid functional genes and allows for them to spread among divergent host strains. Within microbial communities, plasmids influence microbial diversification (186) and contribute to functional gene pools (119). Plasmids can alter the fitness of individuals in a community as they can be gained or lost in the environment, which alters functional gene content in the community and can have consequences for local competitiveness.

Antibiotic resistance genes (ARGs) provide a prime example of the importance of functional genes encoded on plasmids. ARGs can undergo plasmid-mediated horizontal gene transfer (160, 187). There is particular concern about the potential for spread of ARGs between environmental and clinically-relevant bacterial strains. Studies of ARGs in soil have shown overlap between environmental and clinical strains that suggests horizontal gene transfer (HGT) (72–74). For example, plasmid-encoded quinolone resistance (*qnrA*) in clinical Enterobacteriaceae strains likely originated from the environmental strain *Shewanella algae* (73). The extent of the impact of environmental reservoirs of ARGs is unknown (76), but studies have shown evidence for predominantly vertical, rather than horizontal, transfer of these genes (82).

Additionally, it is speculated that rates of transfer in bulk soil are low compared to environments with higher population densities such as the rhizosphere, phyllosphere, and gut microbiomes of soil microorganisms (85). In the case of antibiotic resistance, mobilization is a public health risk. Broadly, the ability of plasmids to rapidly move genes both between and among membership is linked to diversification in complex systems, especially soils (186).

Despite their ecological and functional relevance, plasmids are not well characterized in soil. Plasmids vary in copy number, host range, transfer potential, and genetic makeup (119, 188), making them difficult to assemble and characterize from complex soil metagenomes that contain tens of thousands of bacteria and archaea (189). Plasmid extraction from soil is biased towards smaller plasmids and excludes linear plasmids (119). Additionally, the mosaic gene content on plasmids makes their assembly from metagenomes difficult (119). Though new methods for plasmid assembly from metagenomes are being developed (190, 191), the resulting contigs represent a population average of plasmid gene content and size because they are very likely not derived from an individual cell. Thus, the size ranges of plasmids in soils is largely unknown, but of consequence because size is one factor reported to contribute to plasmid potential for transferability (185). Furthermore, “plasmidome” analysis and plasmid assembly from metagenomes do not provide host information. New methods, such as single-cell analysis and proximity ligation of chromosomes to plasmids prior to sequencing (192), are still expected to assemble plasmids with some degree of mosaicism. However, whole genomes sequenced from soil associated microorganisms, inclusive of both chromosomes and plasmids, could provide plasmid host and size information. A database including this information could also provide



information as to how much overlap there is as to functional genes encoded on plasmids within the host cell chromosome(s).

To aid in the study of plasmids and their associated functional genes in soil, we establish a resource to compare genetic locations of functional genes in soil microorganisms. We extended the RefSoil database (193) of 922 soil microorganisms to include their plasmids. We used this database to test whether soil-associated plasmids are distinct from plasmids from a broad, general database of microorganisms, RefSeq (194). We focused our comparisons on plasmid size and the content, diversity, and location of ARGs on plasmids and chromosomes. We used hidden Markov models from the ResFams database (155) to search for ARGs in the extended soil database, RefSoil+, and RefSeq. RefSoil+ provides insights into the range of plasmid sizes and their functional potential within soil microorganisms. RefSoil+ can be used to inform and test hypotheses about the traits, functional gene content, and spread of soil-associated plasmids and can serve as a reference for plasmid assembly from metagenomes.

## Materials and Methods

### Data availability

All data and workflows are publicly available on GitHub ([github.com/ShadeLab/RefSoil\\_plasmids](https://github.com/ShadeLab/RefSoil_plasmids)). A table of all RefSoil microorganisms with genome and plasmid accession numbers is available on GitHub in the DATABASE\_plasmids repository. This repository also hosts amino acid and nucleotide sequences for RefSoil+ genomes and plasmids. Plasmid retrieval workflows are included in the BIN\_retrieve\_plasmids directory. All workflows are included on Github as well in the ANALYSIS\_antibiotic\_resistance repository.

### RefSoil plasmid database generation

Accession numbers from RefSoil genomes were used to collect assembly accession numbers for all 922 strains. Assembly accession numbers were then used to obtain a list of all genetic elements from the assembly of one strain. Because all RefSoil microorganisms have completed genomes, all plasmids present at the time of sequencing are included in the assembly. Plasmid accession numbers were compiled for each strain and added to the RefSoil database to make RefSoil+. Plasmid accession numbers were used to download amino acid sequences, coding nucleotide sequences, and GenBank files. To ease comparisons between genome and plasmid sequence information, sequence descriptors for plasmid protein sequences were adjusted to mirror the format used for bacterial and archaeal RefSoil files.

### Accessing RefSeq genomes and plasmids

Complete RefSeq genomes and plasmids were downloaded from NCBI to compare with RefSoil. All RefSeq bacteria and archaea protein sequences were downloaded from release 89

(<ftp://ftp.ncbi.nlm.nih.gov/refseq/release>). All GenBank files for complete RefSeq assemblies were downloaded from NCBI. A total of 10,270 bacterial and 259 archaeal assemblies were downloaded. GenBank files were used to extract plasmid size and to compile a list of chromosomal and plasmid accession numbers. GenBank information was read into R and accession numbers for plasmids and chromosomes were separated. Additionally, all RefSoil accession numbers were removed from the RefSeq accession numbers. Ultimately, 10,335 chromosomes and 8,271 plasmids were collected to represent non-RefSoil microorganisms. Protein files were downloaded and tidied using the protocol for RefSoil plasmids as described above.

### Plasmid characterization

We summarized the RefSoil+ and RefSeq plasmids in several ways. Plasmid size was extracted from GenBank files for each RefSoil genome and plasmid. For comparison, size was also extracted from RefSeq plasmids. These data were compiled and analyzed in the R statistical environment for computing (195). The RefSoil metadata, which contains host information for each plasmid, was used to calculate proportions of RefSoil microorganisms with plasmids. Both the number of plasmids per organism and the number of RefSoil microorganisms with one plasmid were examined. Plasmid size distributions were compared using Mann Whitney U tests, Hartigan's dip test (196), and bimodality coefficients (197). The environmental abundances of RefSoil plasmids were calculated using estimations of RefSoil organism environmental abundance (193). Only soil orders with the most Refsoil+ representatives (Alfisol, Mollisol, Vertisol; (193)) were included in the analysis.

### Antibiotic resistance gene detection

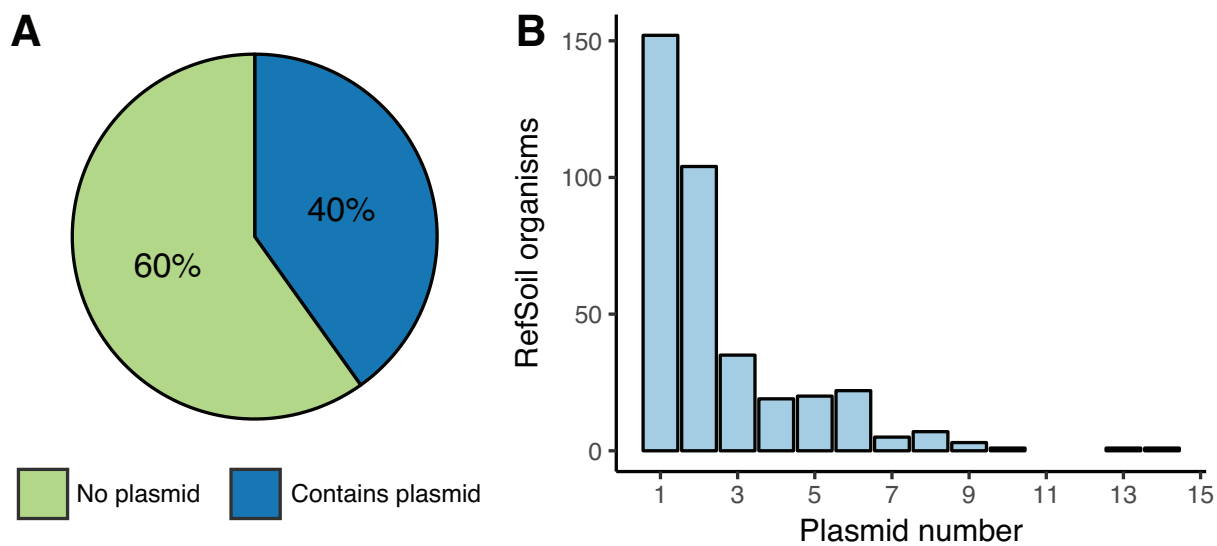
We examined ARGs from the ResFams database (174 total (155) in RefSoil+ (**Appendix A Table 12**). We then used HMMs from the ResFams database (155) to search amino acid sequence data from RefSoil genomes and plasmids with a publicly available, custom script and HMMER (198). To perform the search, `hmmsearch` (198) was used with `-cut_ga` and `-tblout` parameters. These steps were repeated for protein sequence data from the complete RefSeq database (accessed 24 July 2018). Tabular outputs from both datasets were analyzed in R. Quality scores and percent alignments were plotted to determine quality cutoff values for each gene (**Appendix B Figure 11**). All final hits were required to be within 10% of the model length and to have a score of at least 30% of the maximum score for that gene. When one amino acid sequence was annotated twice (i.e. for similar genes), the hit with the lower score was discarded. The final, quality filtered hits were used to plot the distribution of ARGs in RefSoil genomes and plasmids.

## Results and Discussion

### Plasmid characterization

RefSoil+ is an extension of the RefSoil database inclusive of soil-associated plasmids. RefSoil+ includes taxonomic information, amino acid sequences, coding nucleotide sequences, and GenBank files for a curated set of 922 soil-associated microorganisms. In total, 928 plasmids were associated with RefSoil microorganisms, and 370 RefSoil microorganisms (40.1%) had at least one plasmid (**Figure 4.1A**). This percentage is high compared to the proportion of non-eukaryotic plasmids in the general RefSeq database (34%; Mann-Whitney U  $p < 0.01$ ). The mean number of plasmids per RefSoil organism was 1.01, but the number of plasmids per organism varied greatly (variance = 3.2; **Figure 4.1B**). For example, strain *Bacillus thuringiensis* serovar *thuringiensis* (RefSoil 738) had 14 plasmids, ranging from 6,880 to 328,151 bp. The mean number of plasmids per RefSoil organism was also greater than RefSeq (Mann-Whitney U  $p < 0.01$ ). The abundance of plasmids found in RefSoil genomes highlights plasmids as an important component of soil microbiomes (186, 199).

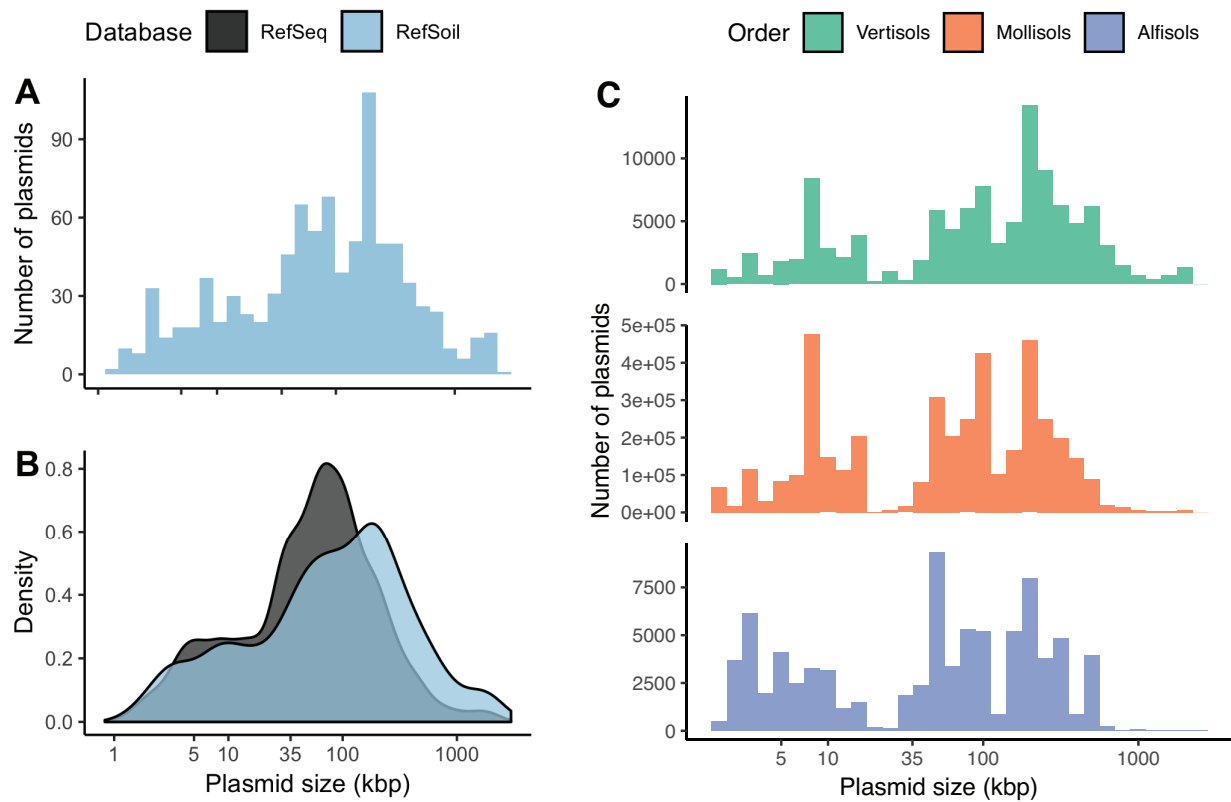
Soil-associated plasmids tended to be larger than plasmids from other environments (Mann-Whitney U  $p < 0.01$ ). Plasmid size in RefSoil microorganisms ranged from 1,286 bp to 2.58 Mbp (**Figure 4.2A**), which rivals the range of all known plasmids from various environments (744 bp – 2.58 Mbp) (16). In the distribution of plasmid size, both upper and lower extremes had representatives from soil. Plasmids from all habitats were previously shown to have a characteristic bimodal size distribution with peaks at 5 kb and 35 kb (15–17). In this analysis, the subset RefSeq plasmids had a multimodal distribution (Hartigans' dip test  $p < 0.01$ ; Bimodality coefficient = 0.745) and modes at 3 kb and 59 kb (**Figure 4.2**). Soil-associated



**Figure 4.1. Summary of RefSoil plasmids.**

A) Percentage of RefSoil microorganisms with (blue) and without (green) detected plasmids.

B) Distribution of the number of plasmids per RefSoil microorganism.



**Figure 4.2. Plasmid size distributions.**

**A)** Histogram of plasmid size (kbp) from RefSoil plasmids. **B)** RefSoil (blue) and RefSeq (gray) plasmid size distributions. **C)** Estimation of plasmid size distribution in three soil orders. Color indicates soil order and n indicates the community size.

plasmids in RefSoil+ also had a multimodal size distribution (Hartigans' dip test  $p < 0.05$ ; Bimodality coefficient = 0.800)) but had modes at 1 kb, 3 kb, 49 kb, and 183 kb. Additionally, RefSoil+ plasmids were larger than RefSeq plasmids (Mann Whitney U  $p < 0.01$ ) (**Figure 4.2**). Specifically, RefSoil+ proportionally contained more plasmids  $> 100$  kb (**Figure 4.2B**). Thus, while soil-associated plasmids vary in size, they are, on average, large. This is of particular importance because of the established differences in mobility of plasmids in different size ranges (185). Smillie and colleagues (2010) showed that mobilizable plasmids, which have relaxases, tend to be larger than non-transmissible plasmids, with median values of 35 and 11 kbp respectively (185). The majority of soil-associated plasmids (68.2%) were  $> 35$  kbp (**Figure 4.2**), suggesting they are more likely to be mobile. Additionally, conjugative plasmids, which encode type IV coupling proteins, have a larger median size (181 kbp) (185). Similarly, RefSoil+ plasmids had a mode of 183 kb (**Figure 4.2**), suggesting that these soil-associated plasmids are more likely to be conjugative. Future works should examine genetic potential for transfer of plasmids associated with different ecosystems to test this hypothesis.

Plasmid size may vary in the environment. To estimate the environmental size distributions of plasmids, we used estimates of the environmental abundance of RefSoil microorganisms (193). We focused on soil orders previously shown to include the most RefSoil representatives (Alfisols, Mollisols, Vertisols) (193). We found that plasmid size distributions varied based on soil order (Kruskal-Wallis  $p < 0.01$ ; **Figure 4.2C**). True environmental abundance may vary based on plasmid copy number within individuals and plasmids from uncultivated microorganisms, but this estimation gives a rough idea of plasmid size distributions

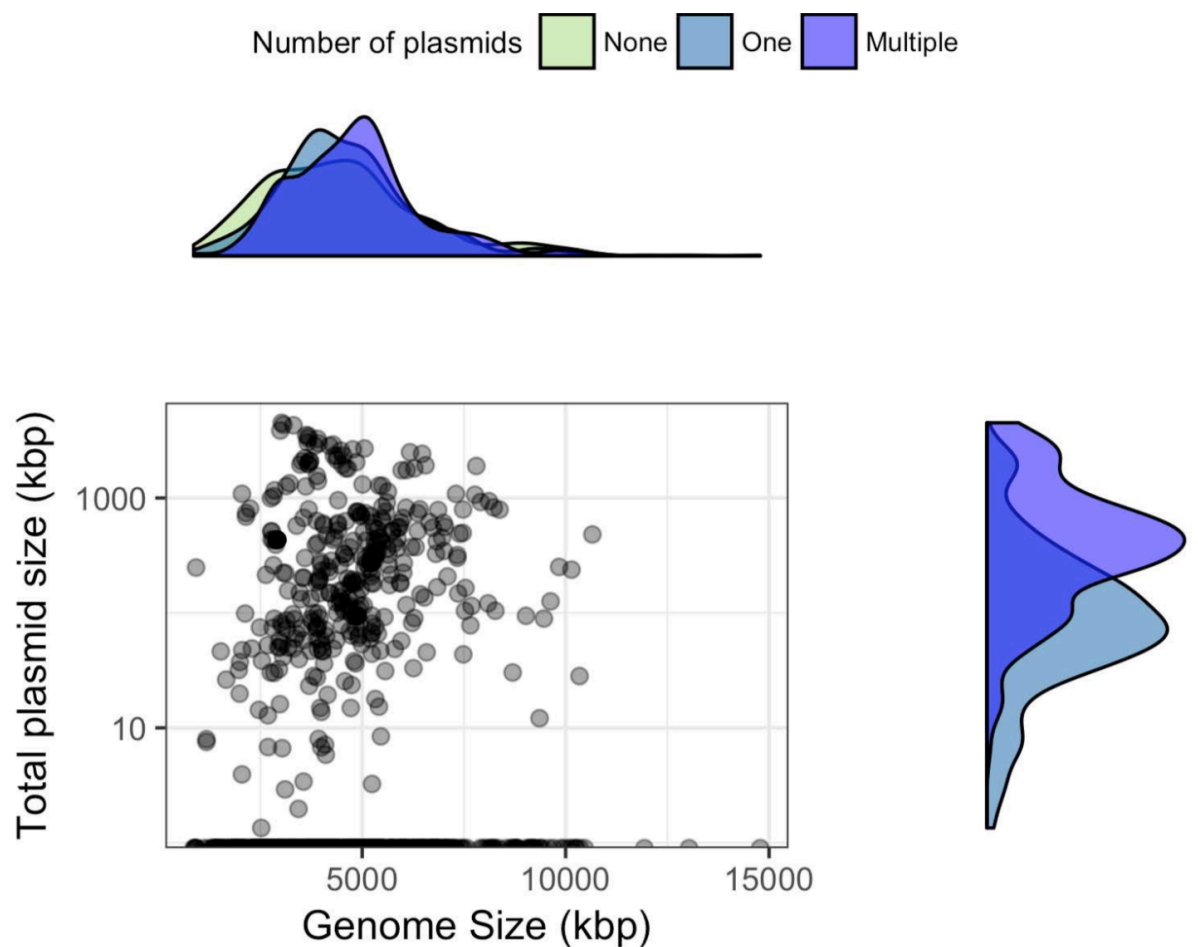


in the environment, and provides some baseline information because there are methodological challenges to accurately measuring plasmid size *in situ* (119, 190, 191).

Genome size, inclusive of chromosomes and plasmids, is an important ecological trait that is difficult to estimate from metagenomes (202). Due to incomplete assemblies, genome size must be approximated based on the estimated number of individuals through single-copy gene abundance (203). Extrachromosomal elements, however, inflate these estimated genome sizes because they contribute to the sequence information of the metagenome often without contributing single-copy genes (204). While our methodologies do not account for plasmid copy number (205), we examined the relationship between genome size and plasmid size in soil-associated microorganisms and found a weak but significant correlation (Spearman's  $\rho = 0.12$ ;  $p < 0.001$ ; **Figure 4.3**). Additionally, chromosome size was not predictive of the number of plasmids (**Figure 4.3**; **Appendix B Figure 11**). For example, *Bacillus thuringiensis* subsp. *thuringiensis* strain IS5056 had the most plasmids in RefSoil+, but these plasmids spanned the size range of 6.8 - 328 kbp. This strain's plasmids make up 19% of its coding sequences (206), but its chromosome (5.4 Mbp) is average for soils (204). Despite the weak relationship between genome size and plasmid characteristics within these data, the plasmid database can be used to inform estimates of average genome sizes from close relatives detected within metagenomes.

#### ARGs on soil plasmids

It is unclear whether soil ARGs are predominantly on chromosomes or mobile genetic elements. While mobile gene pools are not static, there is evidence to suggest low transfer of



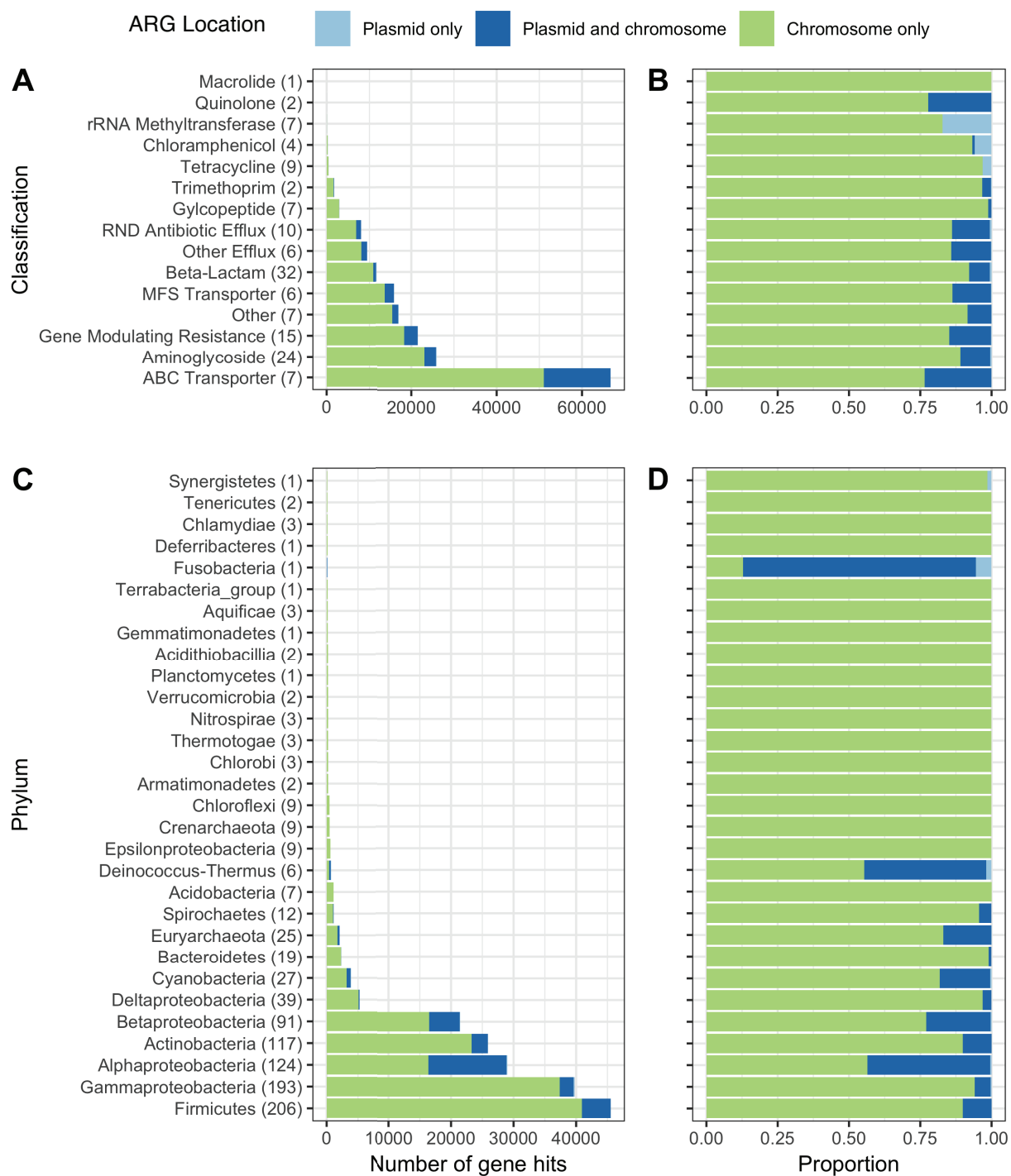
**Figure 4.3. Relationship between plasmid size and genome size.**

Total plasmid size (sum of all plasmids in an microorganism, kbp) is plotted on a log scale against total genome size for each RefSoil microorganism. Density plots are included for each axis to represent the distribution of RefSoil microorganisms with different numbers of plasmids (none (green), one (blue), or multiple (purple)).

ARGs in soil (82, 85, 207). For example, bulk soils are not a “hot spot” for HGT because they are often resource-limited (208), and surveys of ARGs in soil metagenomes have suggested a predominance of vertical transfer, rather than horizontal transfer, of ARGs (82, 207). Using RefSoil+ sequences and ResFams HMMs (155), we examined 174 genes encoding resistance to beta-lactams, tetracyclines, aminoglycosides, chloramphenicol, glycopeptides, macrolides, quinolones, and trimethoprim. After quality filtering, we detected 154,392 ARG sequences in RefSoil chromosomes and plasmids (**Figure 4.4**).

Adding plasmids to the RefSoil database increased the number of functional gene types, or genes that have functional potential (164), represented in the database, as 7 ARGs (16S rRNA methyltransferase, AAC6-Ib, ANT6, CTXM, ErmC, KPC, TetD) were only detected on plasmids. Notably, these functional genes would be missed if only chromosomes were considered. However, the majority of ARGs were chromosomally encoded in RefSoil+ microorganisms (**Figure 4.4AB**; chromosome v. plasmid Mann Whitney U  $p < 0.01$ ). We next examined the genomic distributions of ARGs in RefSoil+ based on taxonomy (**Figure 4.4CD**). Proteobacteria had the most plasmid-associated ARGs, which has been reported previously (66).

We were curious whether ARGs were more commonly detected on chromosomes than plasmids in general, or if this trend was specific to soil microorganisms. We found that the number of ARGs per genome was comparable for RefSoil and RefSeq (Mann Whitney U  $p > 0.05$ ), but RefSoil plasmids had fewer ARGs than RefSeq plasmids (Mann Whitney U  $p < 0.05$ ; **Figure 4.5**). Normalizing to individual microorganisms is biased towards chromosomes,

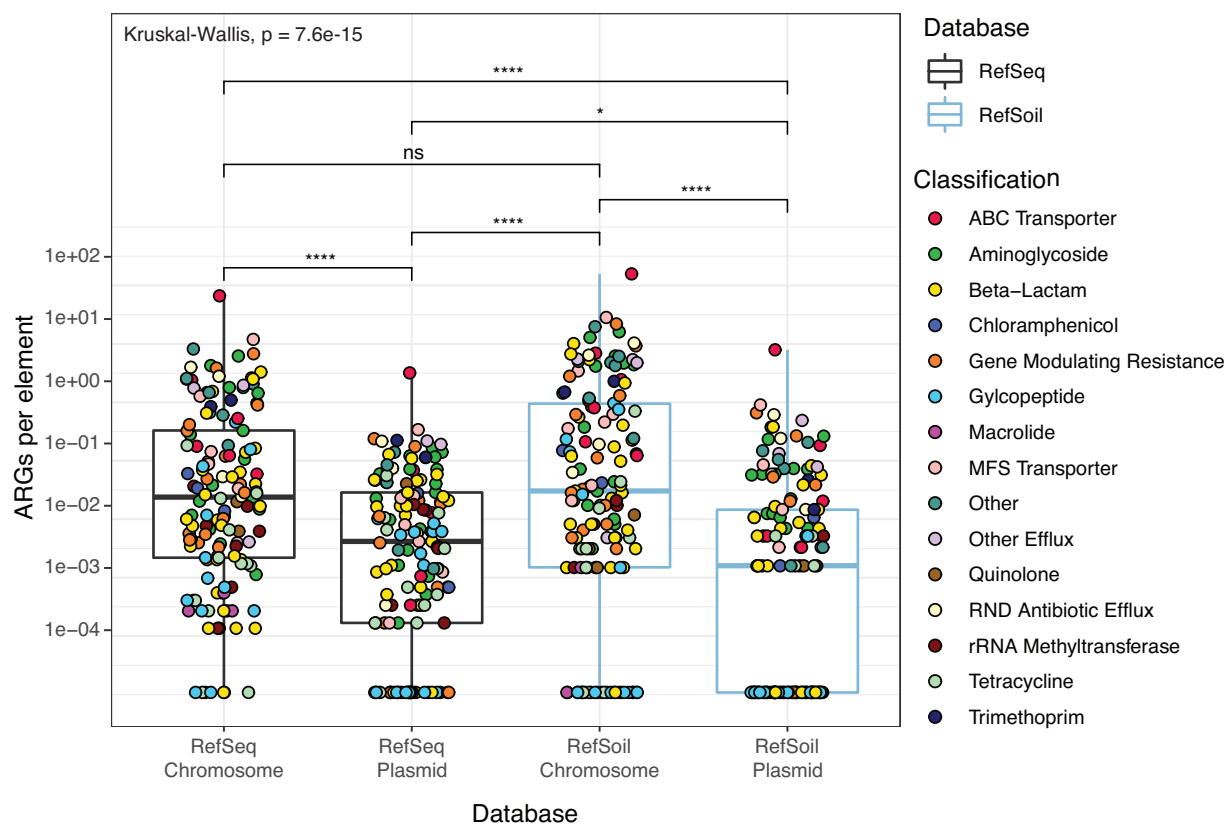


**Figure 4.4. Distribution of ARGs in RefSoil genomes and plasmids.**

**A)** The raw numbers and **B)** proportions of ARGs on plasmids (light blue), genomes (green) or

**Figure 4.4 (cont'd)**

both (dark blue) in RefSoil+ microorganisms by antibiotic resistance gene group. The number of genes included in each group is shown in parentheses. **C)** The raw numbers and **D)** proportions of detected ARGs on plasmids (light blue), genomes (green) or both (dark blue) in RefSoil+ microorganisms by phylum-level taxonomy. The number of taxa included in each phylum is shown in parentheses.



**Figure 4.5. Proportion of ARGs on genomes and plasmids in RefSoil+ and RefSeq databases.**

Number of ARGs was normalized to number of genetic elements. Boxplots are colored by database. Points represent individual ARGs and are colored based on classification. Kruskal-Wallis test results are shown in addition to significant results from pairwise Mann Whitney U tests.

however, because chromosomes typically have more base pairs than plasmids. To account for this, we also normalized ARGs to base pairs, and plasmids from both databases had more ARGs compared to chromosomes (Mann Whitney U  $p < 0.05$ ). Notably, RefSoil+ had less ARGs compared with RefSeq (Mann Whitney U  $p < 0.01$ ) (**Appendix B Figure 12**). This finding suggests that plasmid-mediated HGT rates of ARGs may be relatively low in these soil microorganisms. We note that the RefSoil database is limited in representatives of Verrucomicrobia and Acidobacteria which may change these estimates (193); however, this deficiency will improve as the database grows.

We examined this trend for each antibiotic class and observed a greater proportion of ARG sequences on plasmids in RefSeq compared with RefSoil+ for genes encoding glycopeptide and tetracycline resistance (**Appendix B Figure 13**). Gibson and colleagues (2015) also found a lack of tetracycline resistance genes in soil-associated isolates compared to water and human-associated strains (155). By determining whether ARGs were encoded on plasmids or chromosomes, our analysis suggests that these patterns were due to chromosomal genes and more likely vertically transferred (**Figure 4.5**). Thus, these soil bacteria harbor relatively fewer ARGs on plasmids, suggesting that RefSoil+ microorganisms have limited capacity for plasmid-mediated transfer of these genes. Future assessments of functional gene content on chromosomes and plasmids together will help to delineate changes in transfer potential and reveal selective or environmental factors that impact transfer potential.

While genome data from isolates cannot speak to environmental abundance of ARGs, our data support observations of ARGs in mobile genetic elements in soil from cultivation-independent studies as well. Luo and colleagues (2016) observed a low abundance of

chloramphenicol, quinolone, and tetracycline resistance genes in soil mobile genetic elements (199), and Xiong and colleagues (2015) also observed low abundance of *qnr* genes. Similarly, we observed fewer plasmid-encoded tetracycline resistance genes in soil-associated microorganisms than RefSeq microorganisms (**Appendix B Figure 13**). We did not observe significant differences for genes encoding quinolone or chloramphenicol resistance; however, these genes had small sample sizes ( $n = 2$  and  $3$  respectively). Mobile genetic elements in soil have also been shown to have an abundance of genes encoding multidrug efflux pumps and resistance to beta-lactams, aminoglycosides, and glycopeptides (199). Genes encoding beta-lactam and aminoglycoside resistance were comparable between RefSoil+ and RefSeq (Kruskal-Wallis  $P > 0.05$ ; **Appendix B Figure 13**). However, plasmid-borne glycopeptide resistance genes were less common in RefSoil+ plasmids (Mann Whitney U  $P < 0.05$ ).

#### RefSoil+ applications

RefSoil+ is publicly available on GitHub ([github.com/ShadeLab/RefSoil\\_plasmids](https://github.com/ShadeLab/RefSoil_plasmids)). It includes an excel file linking RefSoil+ organism taxonomy with accession numbers for corresponding chromosomes and plasmids. It also contains several fasta files with protein coding sequences and amino acid sequences. These files can be downloaded directly from GitHub. RefSoil+ has been used to better estimate genome sizes in soil (209) and to estimate the distribution of arsenic resistance genes in soil-associated chromosomes and plasmids (210).

Our results show that soil-associated plasmids have distinctive traits and can harbor functional genes that are not encoded on host chromosomes. RefSoil+ expands knowledge of functional genes with potential for transfer among soil microorganisms and offers insights into plasmid size and host ranges in soil (and improves accuracy of estimates of their genome sizes).



Because it is populated by the chromosomes and plasmids of isolates, RefSoil+ links host taxonomy to plasmid content. This linkage is important especially for heterogeneous ecosystems with high microbial richness like soils, which rely heavily on cultivation-independent methods for observing microbial diversity. RefSoil+ can guide assembly and support annotation of plasmids from soil metagenomes, and also direct hypotheses of host identity (190, 211). Notably, plasmid gene content is not static (212), and individuals can gain or lose plasmids (213, 214). Despite this potential issue, historical data of the genetic makeup and host range of plasmids can be used to better understand plasmid ecology, and to serve as an important reference to understand by how much host plasmid numbers and contents change in the future. This information contributes to information needed to understand patterns of plasmid dissemination, both across environments and among hosts.

RefSoil+ can be used as a reference database or as a database for primer design to target plasmids in the environment. Advances in microbiome sequencing methods such as pre-sequencing proximity linkage (e.g. Hi-C) (192), long-read technology (215), or single cell sequencing (216) could add to and leverage RefSoil+ to improve characterization of plasmid-host relationships in soil. As movement of ARGs are observed in the clinic and the environment, RefSoil+ can also serve as a reference for comparison with legacy plasmid and chromosome content and distributions. Novel genomes and plasmids could be added in future RefSoil+ versions, and plasmid-host relationships as well as encoded functions could be compared between cultivation-dependent and –independent methodologies. RefSoil+ provides a rich community resource for research frontiers in plasmid ecology and evolution within wild microbiomes.

## **CHAPTER 5 : A global survey of arsenic related genes in soil microbiomes**

Work presented in this chapter has been accepted in *BMC Biology* as Dunivin TK, Yeh SS, and Shade A. A global survey of arsenic related genes in soil microbiomes.

## Abstract

Environmental resistomes include transferable microbial genes. One important resistome component is resistance to arsenic, a ubiquitous and toxic metalloid that can have negative and chronic consequences for human and animal health. The distribution of arsenic resistance and metabolism genes in the environment is not well understood. However, microbial communities and their resistomes mediate key transformations of arsenic that are expected to impact both biogeochemistry and local toxicity. We examined the phylogenetic diversity, genomic location (chromosome or plasmid), and biogeography of arsenic resistance and metabolism genes in 922 soil genomes and 38 metagenomes. To do so, we developed a bioinformatic toolkit that includes BLAST databases, hidden Markov models and resources for gene-targeted assembly of nine arsenic resistance and metabolism genes: *acr3*, *aioA*, *arsB*, *arsC* (grx), *arsC* (trx), *arsD*, *arsM*, *arrA*, and *arxA*. Though arsenic related genes were common, they were not universally detected, contradicting the common conjecture that all organisms have them. From major clades of arsenic related genes, we inferred their potential for horizontal and vertical transfer. Different types and proportions of genes were detected across soils, suggesting microbial community composition will, in part, determine local arsenic toxicity and biogeochemistry. While arsenic related genes were globally distributed, particular sequence variants were highly endemic (e.g., *acr3*), suggesting dispersal limitation. The gene encoding arsenic methylase *arsM* was unexpectedly abundant in soil metagenomes (median 48%), suggesting that it plays a prominent role in global arsenic biogeochemistry. Our analysis advances understanding of arsenic resistance, metabolism, and biogeochemistry, and our approach provides a roadmap for the ecological investigation of environmental resistomes.

## Introduction

Microbial communities drive global biogeochemical cycles through diverse functions. The biogeography of functional genes can help to predict and manage the influence of microbial communities on biogeochemical cycling (217). These trait-based analyses require that the functional genes are well-characterized from both evolutionary and genetic perspectives (218). The arsenic resistance and metabolism genes exemplify a suite of well-characterized functional genes that have consequences for biogeochemistry. Arsenic is a toxic metalloid that, upon exposure, can have negative effects for all life, including humans, livestock, and microorganisms. The toxicity and mobility of arsenic depends, in part, on its oxidation state: the trivalent arsenite is more mobile and more toxic than the pentavalent arsenate (14). The toxicity of methylated arsenic species varies with oxidation state and number of methyl groups (monomethyl, dimethyl, trimethyl). Pentavalent methylarsenicals are progressively less toxic than inorganic arsenate, while trivalent methylarsenicals are progressively more toxic than inorganic arsenite with the exception of trimethylarsine which is the least toxic arsenic species (219, 220). Additionally, volatilization of arsenic can occur through methylation (221), which has varied impacts. Methylated forms of arsenic can be released to new areas through air (222), captured during bioremediation (7), or accumulate in crops such as rice (90). Microbial transformations of arsenic can have consequences for arsenic speciation and methylation; therefore, they impact arsenic ecotoxicity and the fate of arsenic in the environment.

Arsenic biogeochemical cycling by microbial communities is both an ancient (15, 17) and a contemporary (14, 63) phenomenon. Changes to the methylation or oxidation state of arsenic alter biogeochemical cycling of arsenic, and microbes have evolved a variety of mechanisms to carry out these functions. Arsenic related genes are generally separated into two categories:

resistance and metabolism (21). Arsenic resistance, or detoxification, is encoded by the *ars* operon (8). The *ars* operon protects the cell from arsenic but does not detoxify arsenic itself in the environment. This operon includes arsenite efflux (ArsB, Acr3) which is potentially precluded by cytoplasmic arsenate reduction with either glutaredoxin (ArsC (grx)) or thioredoxin (ArsC (trx)) (8). Arsenic metabolisms include methylation (ArsM), oxidation (AioAB, ArxAB), and dissimilatory reduction (ArrAB) (21). While these genetic determinants of arsenic detoxification and metabolism are well-characterized, the full scope of arsenic detoxification and metabolism gene distribution, diversity, and interspecies transfer is unknown (223–225).

Microbial arsenic resistance is reportedly widespread in the environment. Arsenic resistant organisms have been found in sites with low arsenic concentrations (< 7 ppm) (6, 226), and it has been speculated that nearly all organisms have arsenic resistance genes (1). While the number of identified microorganisms with arsenic resistance genes continues to grow (21), the number of microorganisms without arsenic resistance genes is unclear. Furthermore, though the complete arsenic biogeochemical cycle has been detected in the environment (17), the relative contributions of genes encoding detoxification and metabolism remain unknown (15). A global, biogeographic perspective of environmental arsenic related genes would improve understanding of their ecology. This information would expand foundational knowledge of arsenic detoxification and metabolism, including local and global abundances, gene diversity, dispersal across different environments, and representations over the microbial tree of life.

Knowledge gaps concerning the diversity of microbial arsenic related genes are driven, in part, by numerous inconsistencies in nomenclature and detection methods. Though public microbial metagenome and genome data continue to surge, there are several practical hurdles to

achieving a robust, global assessment of microbial arsenic related genes from this wealth of data. First, tools to detect these genes rely on imperfect annotation (223) and widely vary in nomenclature (31). Next, the use of different reference databases (58, 63, 227–229) and normalization techniques (229, 230) complicates comparisons between studies. To overcome these hurdles, we developed an open-access toolkit to examine arsenic resistance and metabolism genes in microbial sequence datasets. This toolkit allowed us probe genomic and metagenomic datasets simultaneously to investigate arsenic related genes in soil microbiomes. We first asked whether arsenic related genes are universal in soil-associated microorganisms. Next we tested the hypothesis that genes encoding arsenic detoxification are more abundant than those encoding arsenic metabolism. We also tested the hypothesis that arsenic resistance genes with redundant function (i.e. *acr3* and *arsB*; *arsC* (grx) and *arsC* (trx)) would have complementary environmental abundances. Third, we asked whether estimations of arsenic related gene abundance are biased by cultivation efforts, as cultivation is often a research emphasis because cultivable, arsenic resistant microorganisms can be used in bioremediation (225). Finally, we tested the hypothesis that sequence variants of arsenic related genes are endemic, not cosmopolitan.

## Materials and Methods

### Gene Selection and Functional Gene (FunGene) Database Construction

Marker genes can be used to estimate their potential to influence the arsenic biogeochemical cycle (31, 229), so we selected nine well-characterized genes: *acr3*, *aioA*, *arsB*, *arsC* (grx), *arsC* (trx), *arsD*, *arsM*, *arrA*, and *arxA*. FunGene databases (164) were constructed for the following arsenic related genes: *arsB*, *arsC* (grx), *arsC* (trx), *acr3*, *aioA*, *arrA*, and *arxA*. The *arxA* database was constructed with seed sequences from (63). For all other genes, UniProt (231) was used to obtain full length, reviewed sequences when possible. NCBI clusters of orthologous groups (COG) (27) for each gene were examined for evidence of function in the literature. All COG and UniProt sequences were aligned using MUSCLE (232). Aligned sequences were included in a maximum likelihood tree with 50 bootstrap replications made with MEGA (v7.0,(137)). Sequences that did not cluster with known sequences and had no evidence of function were removed. A final FASTA file for each gene was submitted to the Ribosomal Database Project (RDP) to construct a FunGene database (164). All arsenic related gene databases are freely available on FunGene (<http://fungene.cme.msu.edu/>).

### Arsenic related genes in cultivable soil microorganisms

The RefSoil+ database (233) was used to obtain high-quality genomes (chromosomes and plasmids) from soil microorganisms in the Genomes OnLine (GOLD) database (234). RefSoil+ chromosomes and plasmids were searched with hmmsearch (198) using hidden Markov models (HMMs) from FunGene with an e-value cutoff of  $10^{-10}$ . The top hits were analyzed in R (195). For each gene, scores and percent alignments were plotted to determine quality cutoffs. Stringent percent alignment scores were included since this search was against completed genome sequences: only hits with scores  $> 100$  and percent alignment  $> 90\%$  were included. Hits with the

lowest scores were manually examined to test for false positives. Due to false positives, hits against *aioA*, *arrA*, and *arxA* were further quality filtered to have scores > 1,000. When one open reading frame (ORF) contained multiple hits, the hit with a lower score was removed. Taxonomy was assigned using the RefSoil database (193), and the relative abundance of arsenic related genes within phyla were examined. A 16S rRNA gene maximum likelihood tree of RefSoil+ bacterial strains was with RAXML (v.8.0.6 (235)) based on the Whelan and Goldman (WAG) model with 100 bootstrap replicates (“-m PROTGAMMAWAG -p 12345 -f a -k -x 12345 -# 100”). Based on accession numbers, gene hits were extracted from RefSoil+ sequences and used to construct maximum likelihood trees for each gene.

#### Reference Database Construction

Reference gene databases of diverse, near full length sequences were constructed using limited sequences from FunGene databases (164) for the following genes: *acr3*, *aioA*, *arrA*, *arsB*, *arsC* (grx), *arsC* (trx), *arsD*, *arsM*, and *arxA*. Seed sequences and HMMs for each gene were downloaded from FunGene, and diverse protein and corresponding nucleotide sequences were selected with gene-specific search parameters (**Appendix A Table 13**). Briefly, minimum amino acid length was set to 70% of the HMM length; minimum HMM coverage was set to 80% as is recommended by Xander software for targeted gene assembly; and a score cutoff was manually selected based on a drop off point. Sequences were de-replicated before being used in subsequent analysis, and final sequence counts are included in **Appendix A Table 13**. Reference databases were converted to publicly available BLAST databases using BLAST+ (167).

Reference and BLAST databases are publicly available on GitHub

([https://github.com/ShadeLab/meta\\_arsenic](https://github.com/ShadeLab/meta_arsenic)).



### Sample collection and preparation

A soil surface core (20 cm depth and 5.1 cm diameter) was collected in October 2014 from Centralia, Pennsylvania (GPS coordinates: 40 48.070, 076 20.574). For cultivation-dependent work, a soil slurry was made by vortexing 5 g soil with 25 mL phosphate-buffered saline (PBS) for 1 min. Remaining soil was stored at -80°C until DNA extractions. The soil slurry was allowed to settle for 2 min. 100 µL of the slurry was then removed and serially diluted using PBS to a 10<sup>-2</sup> dilution. 100 µL of the solution was added to 50% trypticase soy agar (TSA50) with 200 µg/ml cycloheximide to prevent fungal growth. Plates were incubated at 60°C for 72 h. Lawns of growth were extracted by adding 600 µL trypticase soy broth with 25% glycerol to plates. The plate scrapings were stored at -80°C until DNA extraction.

### DNA extraction and metagenome sequencing

DNA for cultivation-independent analysis was manually extracted from soil using a phenol chloroform extraction (236) and the MoBio DNEasy PowerSoil Kit (MoBio, Solana Beach, CA, USA) according to the manufacturer's instructions. DNA extraction for cultivation-dependent analysis was performed in triplicate from 200 µL of plate scrapings using the E.Z.N.A. Bacterial DNA Kit according to the manufacturer's instructions. All DNA was quantified using a Qubit dsDNA BR Assay Kit (Life Technologies, NY, USA) and was submitted for NGS library prep and sequencing at the Michigan State University Genomics Core sequencing facility (East Lansing, MI, USA). Libraries were prepared using the Illumina TruSeq Nano DNA Library Preparation Kit. After QC and quantitation, the libraries were pooled and loaded on one lane of an Illumina HiSeq 2500 Rapid Run flow cell (v1). Sequencing was performed in a 2 x 150 bp paired end format using Rapid SBS reagents. Base calling was

performed by Illumina Real Time Analysis (RTA) v1.18.61 and output of RTA was demultiplexed and converted to FastQ format with Illumina Bcl2Fastq v1.8.4.

#### Public soil metagenome acquisition

In total, 38 soil metagenomes were obtained for this work (**Appendix A Table 14**). Datasets from Centralia, PA were generated in our research group. All other metagenome data sets were obtained from MG-RAST (<http://metagenomics.anl.gov/>). The MG-RAST database was searched on May 15, 2017, with the following criteria: material = soil, sequence type = shotgun, public = true. The resulting list of metagenome data sets was ordered by number of base pairs (bp). Metagenomic data sets with the most bp were only included if they were sequenced using Illumina to standardize sequencing errors, had an available FASTQ file for internal quality control, and contained < 30% low quality as determined by MG-RAST. Within high quality Illumina samples, priority for inclusion was given to projects with multiple samples so that comparisons could be made both within and between soil sites. When a project had multiple samples, data sets with the greatest bp were selected. While we prioritized samples with multiple datasets, several replicate samples were omitted early on due to > 30% of data removed during quality filtering, and samples Illinois soil, Russian permafrost, and Wyoming soil have just one sample. This search ultimately yielded 26 data sets from 12 locations and five countries.

#### Soil metagenome processing and gene targeted assembly

Sequences from MG-RAST data sets as well as Centralia sample Cen13 were quality controlled using the FASTX toolkit (fastq\_quality\_filter, "-Q33 -q 30 -p 50"). Twelve datasets from Centralia, PA, were obtained from the Joint Genome Institute and quality filtered as described previously (207). Quality filtered sequences were used in all downstream analyses. For

each data set, a gene targeted metagenome assembler (237) was used to assemble each gene of interest. For each gene of interest, seed sequences, HMMs, and reference gene databases described above were included. For *rplB*, reference gene database, seed sequences, and HMMs from the Xander package were used. In most instances, default assembly parameters were used except to incorporate differences in protein length (i.e. protein is shorter than default 150 amino acids) or to improve quality (i.e. maximum length is increased to improve specificity) (**Appendix A Table 13**). While the assembler includes chimera removal, additional quality control steps were added. Final assembled sequences (operational taxonomic units, OTUs) were searched against the reference gene database as well as the non-redundant database (nr) from NCBI (August 28, 2017) using BLAST (167). Genes were re-examined if the top hit had an e-value  $> 10^{-5}$  or if top hit descriptors were not the target gene. Genes with low quality results were re-assembled with adjusted parameters.

#### Soil metagenome comparison

To compare assembled sequences between samples, gene-based OTU tables were constructed. Aligned sequences from each sample were dereplicated and clustered at 90 amino acid identity using the RDP Classifier (168). Dereplicated, clustered sequences were converted into OTU tables with coverage-adjusted abundance. These tables were subsequently analyzed in R (195). *rplB* OTUs were used to compare community structure. The six most abundant phyla were extracted to include at least 75% of each community; the full community structure is available. To compare the abundance of arsenic related genes among data sets, total counts of *rplB* were used to normalize the abundance of each OTU. Relative abundance of arsenic related genes was also calculated for each sample.

## Results

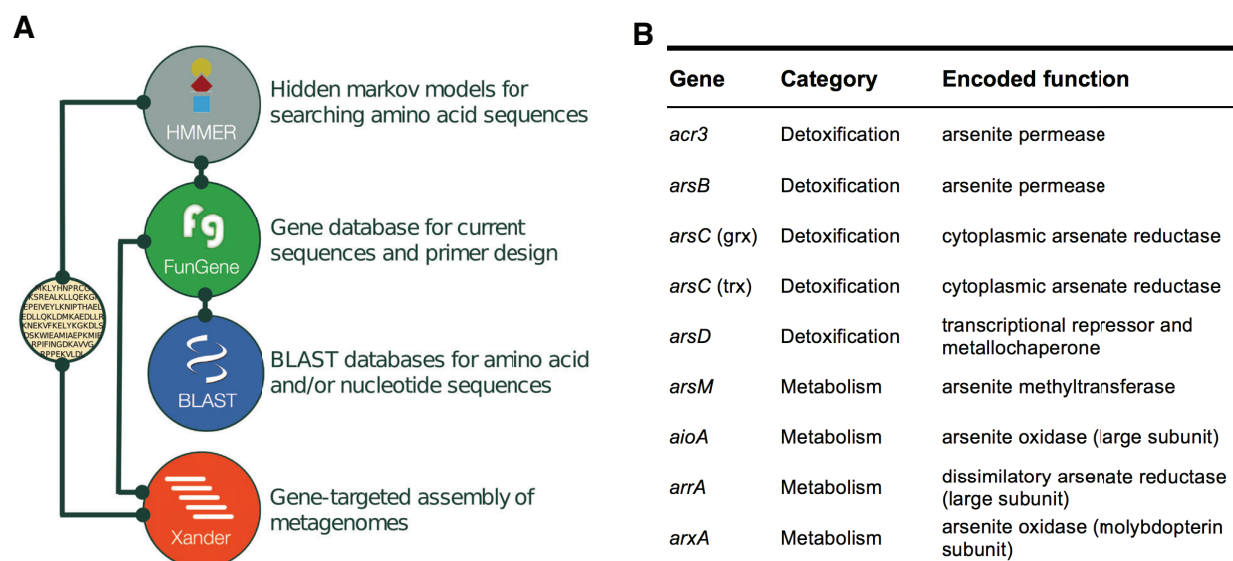
### *A bioinformatic toolkit for detecting and quantifying arsenic related genes*

We developed a toolkit to improve investigations of microbial arsenic related genes (**Figure 5.1**)(8, 25, 46, 238–240). We selected these nine genes because they are markers of arsenic detoxification and metabolism (31, 229) and because their genetic underpinnings are well established. Seed sequences (high quality and full length sequences) for each gene of interest were collected and used to construct BLAST databases (241), functional gene (FunGene) databases (164), hidden Markov models (HMMs (242)), and gene resources for gene-targeted assembly (Xander (237)) (**Figure 5.1A**). Altogether, this toolkit relies on consistent references and nomenclature and can search both amino acid and nucleotide sequence data.

To demonstrate the utility of our toolkit, we performed an analysis of arsenic related genes in soil-associated genomes and metagenomes. We used HMMs for marker genes for arsenic detoxification and metabolism to search RefSoil+ genomes, a set of complete chromosomes and plasmids from cultivable soil microorganisms (233). Additionally, we used a gene-targeted assembler (237) to test 38 public soil metagenomes from Brazil, Canada, Malaysia, Russia, and the United States for arsenic resistance and metabolism genes (**Appendix A Table 14**). Ultimately, these data serve as a broad baseline of arsenic detoxification and metabolism genes in soil.

### *Phylogenetic distributions and genomic locations of arsenic related genes*

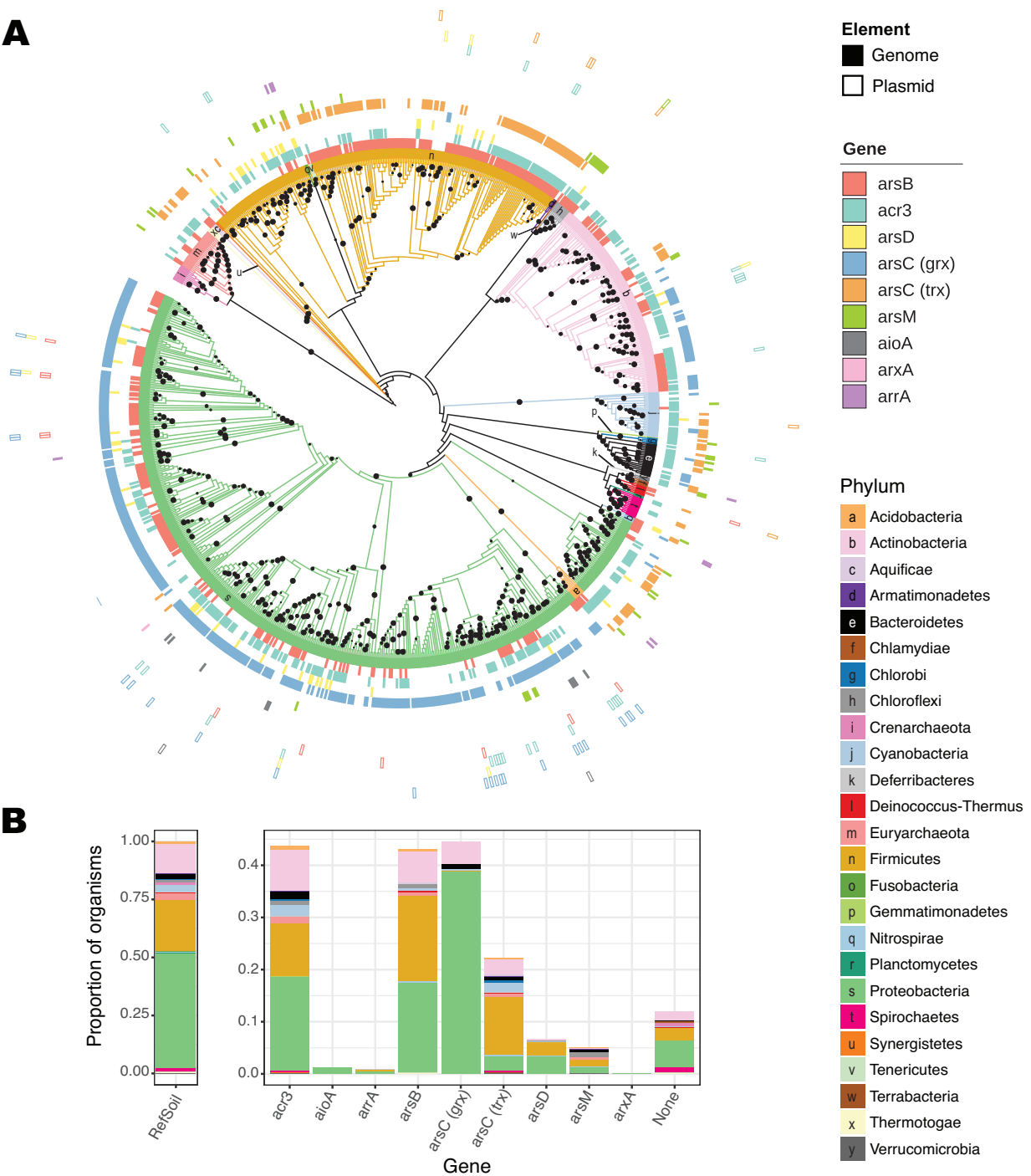
We asked whether arsenic resistance and metabolism genes were universal in RefSoil+ organisms (233). Of the 922 RefSoil+ genomes spanning 25 phyla (**Figure 5.2; Appendix A Table 15**), 14.3% (132 genomes) did not contain any tested arsenic related genes. Of the 25



**Figure 5.1. Arsenic resistance and metabolism gene toolkit schematic.**

**A)** Seed sequences for nine arsenic resistance genes were used to construct an arsenic resistance gene database with existing tools (164, 167, 198, 237). Lines indicate interdependence between modules. **B)** Table of arsenic resistance and metabolism genes included in the toolkit. The toolkit is freely available on GitHub: [github.com/ShadeLab/meta\\_arsenic](https://github.com/ShadeLab/meta_arsenic)

**A**



**Figure 5.2. Arsenic resistance and metabolism genes in RefSoil+ organisms.**

A) Maximum likelihood tree of 16S rRNA genes in RefSoil+ organisms. Bootstrap support > 50 is shown with black circles. Tree branches and the first ring are colored by organism

**Figure 5.2 (cont'd)**

taxonomy. Each node node is annotated with arsenic resistance genotype where color indicates the gene. Filled boxes indicate gene presence on chromosome, and open boxes indicate gene presence on plasmid. **B)** Proportion of RefSoil+ organisms and organisms containing arsenic resistance genes are colored by the taxonomy of the organism containing the gene. “None” refers to the number of genomes that do not test positive for any of the nine arsenic resistance genes analyzed. Note the difference between y-axes.

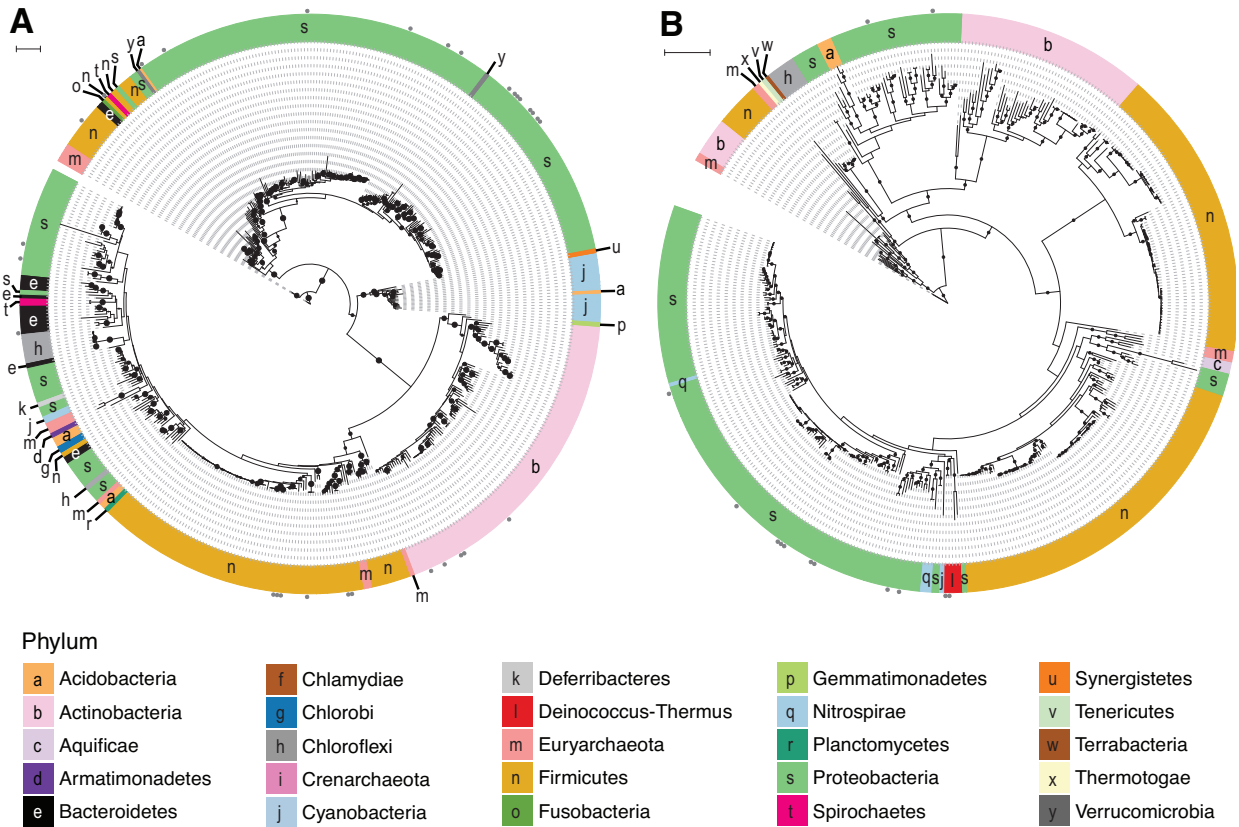
phyla in RefSoil+, two phyla (Chlamydiae and Crenarchaeota) did not have any of these genes. These phyla, however, had few RefSoil+ representatives (three and nine, respectively), so other members of these phyla may have arsenic detoxification and metabolism genes. Supporting this suggestion, a Crenarchaeota isolate was previously reported to oxidize arsenic (243). Nonetheless, these data suggest that arsenic related genes are widespread but not universal, even among cultivable soil organisms (**Figure 5.2**).

We next asked whether 16S rRNA gene phylogeny was predictive of arsenic genotypes using a test for phylogenetic signal (Bloomberg's K (244)). No phylogenetic signal was observed for plasmid-borne sequences or genes encoding arsenic metabolisms (*aioA*, *arrA*, *arxA*); however, relatively few RefSoil+ microorganisms tested positive for these genes. Despite their phylogenetic breadth (**Appendix B Figure 14 – Appendix B Figure 18**), chromosomally-encoded *acr3*, *arsB*, *arsC* (*grx*), *arsC* (*trx*), and *arsM* were similar between phylogenetically related organisms (false discovery rate adjusted  $p < 0.01$ ; **Figure 5.2A**).

#### *Phylogenetic diversity of arsenic related genes: insights into vertical and horizontal transfer*

We examined the phylogenetic diversity of distinct genes encoding arsenite efflux pumps, *acr3* and *arsB*, for soil-associated microorganisms (**Figure 5.3, Appendix B Figure 14 – Appendix B Figure 15**). Gene *acr3* is separated into two clades: *acr3*(1) and *acr3*(2) (30). Clade *acr3*(1) is typically composed of Proteobacterial sequences while *acr3*(2) is typically composed of Firmicutes and Actinobacterial sequences (30, 31, 86). Though RefSoil+ genomes were mostly composed of *acr3*(2) sequences from Proteobacteria (**Figure 5.3A; Appendix B Figure 14**), we





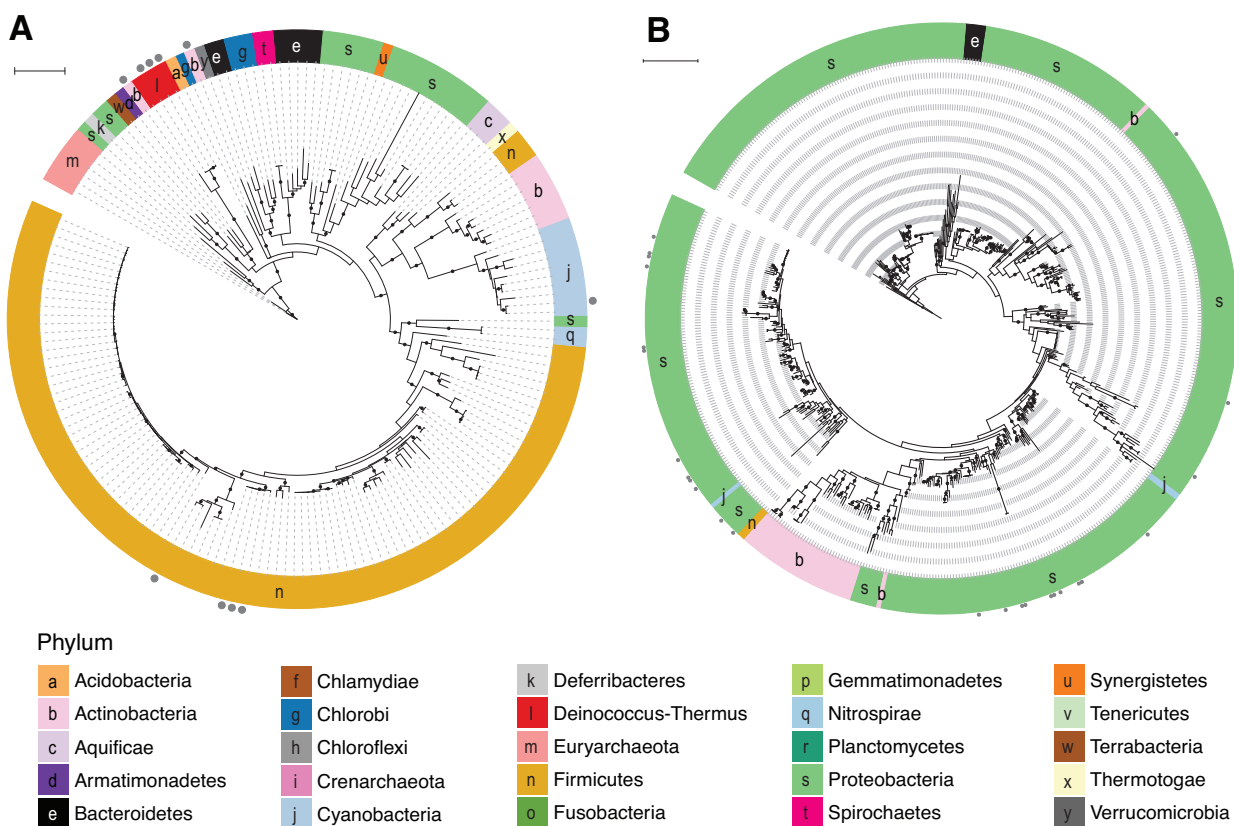
**Figure 5.3. Phylogeny of arsenite efflux pumps in RefSoil+ organisms.**

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

observed greater taxonomic diversity observed than previously reported for this clade (30, 31, 86). Surprisingly, there were deep branches in *acr3*(2) that belonged to Bacteroidetes, Euryarchaeota, Firmicutes, Fusobacteria, and Verrucomicrobia. Similarly, *acr3*(1) contained closely related *acr3* sequences present in a diverse array of phyla (10 out of 25). Both clades had sequences present on plasmids (6.1%). Plasmid-borne *arsB* sequences were only present in Proteobacteria and Deinococcus-Thermus strains (**Figure 5.3B; Appendix B Figure 15**). Sequences from Actinobacteria, Proteobacteria, and Firmicutes were each present in two distinct phylogenetic groups, and previous studies also observed separation of *arsB* sequences based on phylum (30, 86). Interestingly, our genome-centric analysis revealed that microorganisms with multiple copies of *arsB* did not harbor identical copies. For example, seven *Bacillus subtilis* subsp. *subtilis* strains had two copies of *arsB*, with one from each of the two clades (**Appendix B Figure 15**).

Cytoplasmic arsenate reductase (*ArsC* (trx)) was phylogenetically widespread in RefSoil+ microorganisms (**Figure 5.4A; Appendix B Figure 16**). While some *arsC* (trx) sequences were plasmid-borne, the majority were chromosomally encoded. Similarly, plasmid encoded *arsC* (grx) made up 4.6% of RefSoil+ hits (**Figure 5.4B; Appendix B Figure 17**). Notably, several Proteobacteria strains have multiple copies of *arsC* (grx) with distinct sequences. It is possible that this is the result of an early gene duplication event or HGT of a second *arsC* (grx).

*arsM* was relatively uncommon in RefSoil+ microorganisms (5.2%) (**Figure 5.2**). In the RefSoil+ database, *arsM* was observed in Euryarchaeota as well as several bacterial phyla



**Figure 5.4. Phylogeny of cytoplasmic arsenate reductases in RefSoil+ organisms.**

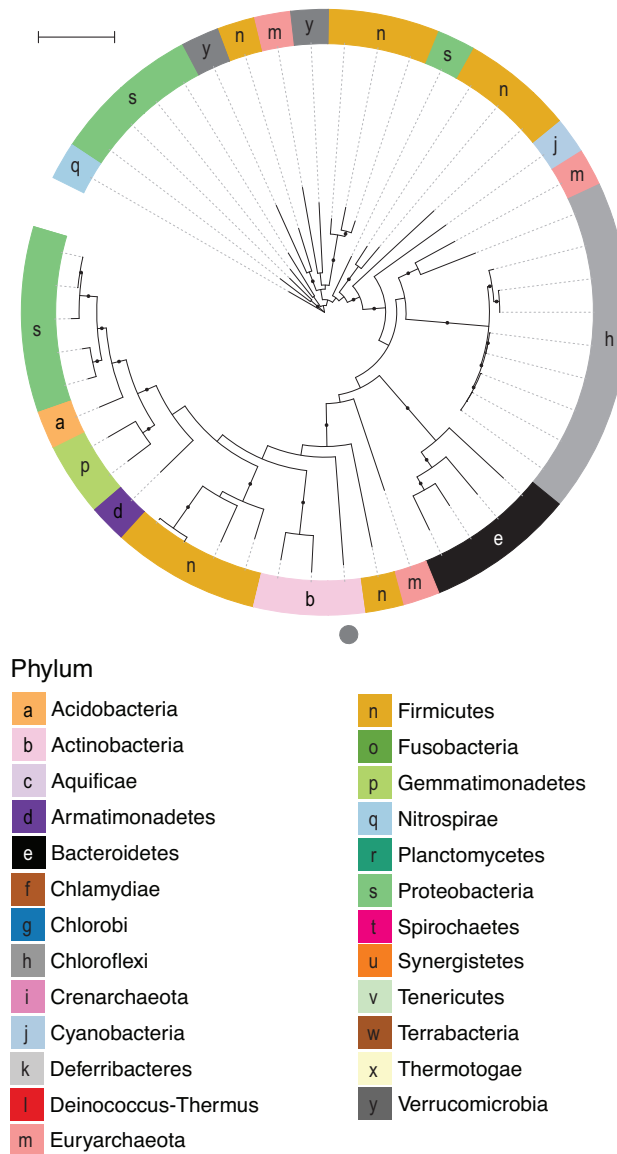
Maximum likelihood tree with 100 bootstrap replications of **A)** ArsC (trx) and **B)** ArsC (grx) sequences predicted from RefSoil+ genomes. Tree scale = 1. Leaf tip color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the tree. Grey circles on the exterior of the tree indicate that a hit was detected on a plasmid and not a chromosome.

Acidobacteria, Actinobacteria, Armatimonadetes, Bacteroidetes, Chloroflexi, Cyanobacteria, Firmicutes, Gemmatimonadetes, Nitrospirae, Proteobacteria, Verrucomicrobia (**Figure 5.5; Appendix B Figure 18**). Notably, only one RefSoil+ microorganism, *Rubrobacter radiotolerans* (NZ\_CP007516.1), had a plasmid-borne *arsM*.

Arsenic metabolism genes *aioA*, *arrA*, and *arxA* were phylogenetically conserved (**Figure 5.6**). Genes encoding arsenite oxidases *aioA* and *arxA* were restricted to Proteobacteria. *aioA* sequences clustered into two clades based on class-level taxonomy: all Alphaproteobacteria sequences cluster separately from Gamma- and Betaproteobacteria sequences. The gene encoding dissimilatory arsenate reduction *arrA* was also phylogenetically conserved in RefSoil+ strains, with strains from Proteobacteria clustering separate from Firmicutes (**Figure 5.6**).

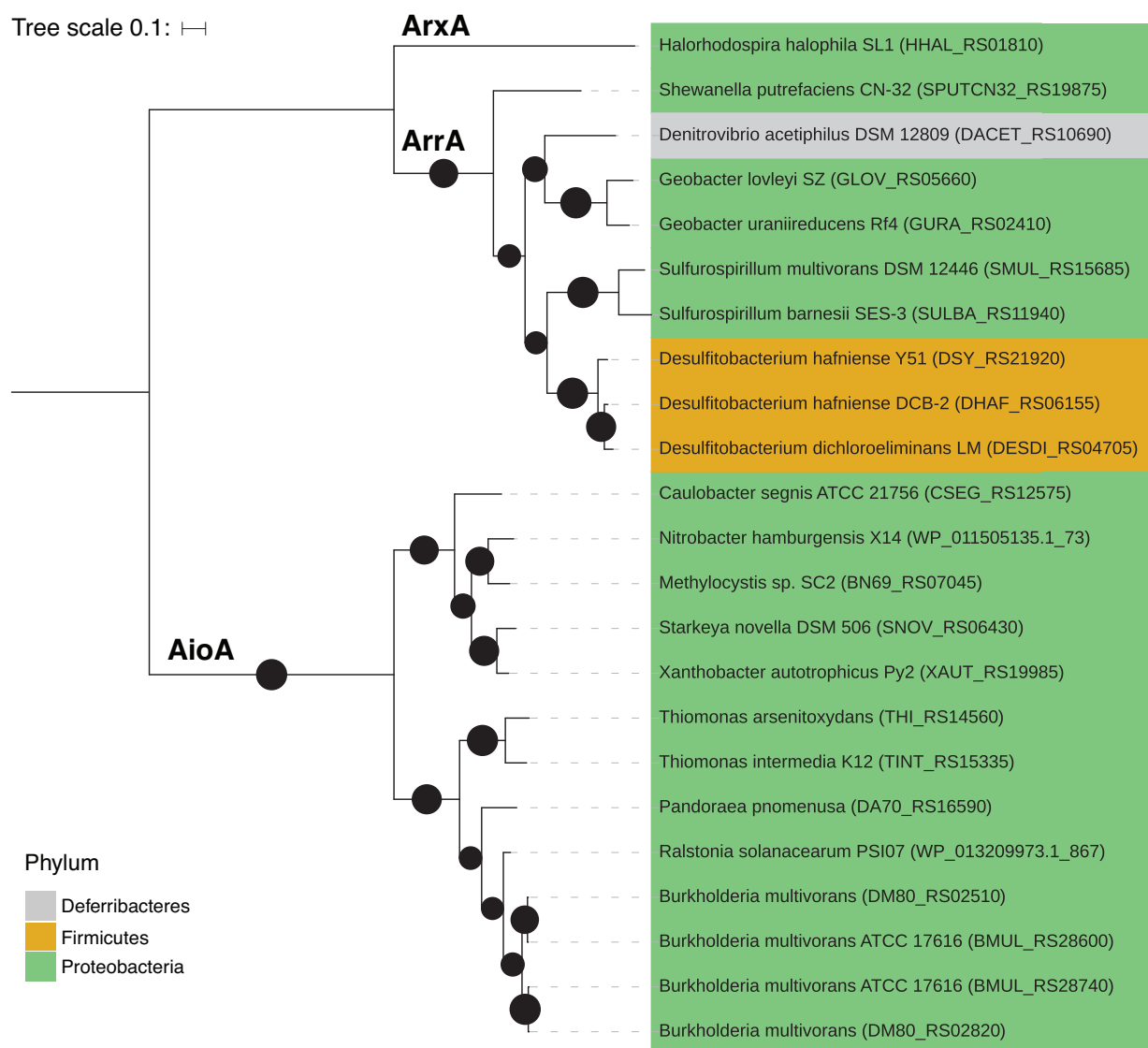
#### Cultivation bias and environmental distributions of arsenic related genes

To gain a cultivation-dependent perspective of the abundances of arsenic related genes in soils, we used inferred environmental abundances of RefSoil microorganisms (61, 193). The environmental abundance of RefSoil microorganisms, which are cultivable, soil-associated microorganisms, was previously estimated by comparing 16S rRNA gene sequences in RefSoil with those in soil metagenomes (193). We used this estimated abundance of cultivable microorganisms along with arsenic related gene information from this study (**Figure 5.2**) to estimate the environmental abundances of arsenic related genes from the cultivated bacteria. Arsenic metabolism genes (*aioA*, *arrA*, *arsM*, *arxA*) were predicted to be less common in the environment compared with arsenic detoxification genes (*acr3*, *arsB*, *arsC* (grx), *arsC* (trx), and *arsD*) (**Figure 5.7A**; Mann Whitney U test  $p < 0.01$ ). Despite similar distributions of *acr3* and *arsB* in RefSoil+ (**Figure 5.2B**), *acr3* was more abundant in most soil orders (**Figure 5.7A**;



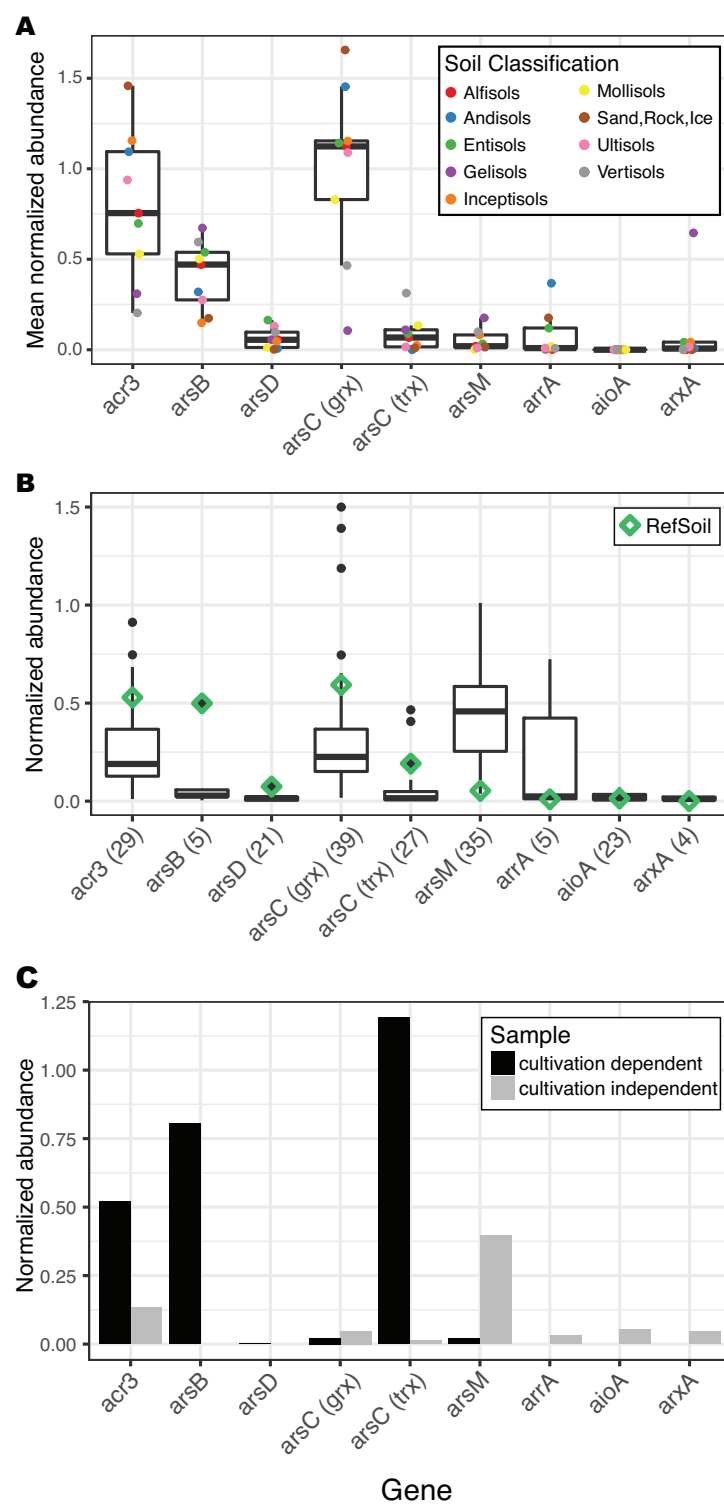
**Figure 5.5. Phylogeny of ArsM in RefSoil+ organisms.**

Maximum likelihood tree with 100 bootstrap replications of ArsM sequences predicted from RefSoil+ genomes. Tree scale = 1. Leaf tip color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the tree. Grey circles on the exterior of the tree indicate that a hit was detected on a plasmid and not a chromosome.



**Figure 5.6. Phylogeny of AioA, ArrA, and ArxA in RefSoil+ organisms.**

Maximum likelihood tree with 100 bootstrap replications of dissimilatory arsenic resistance proteins predicted from RefSoil+ genomes. Tree scale = 0.1. Leaf tips show the name of the RefSoil+ organisms and background color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the tree.



**Figure 5.7. Comparison of arsenic resistance and metabolism gene abundance between cultivation dependent and cultivation independent methods.**

**Figure 5.7 (cont'd)**

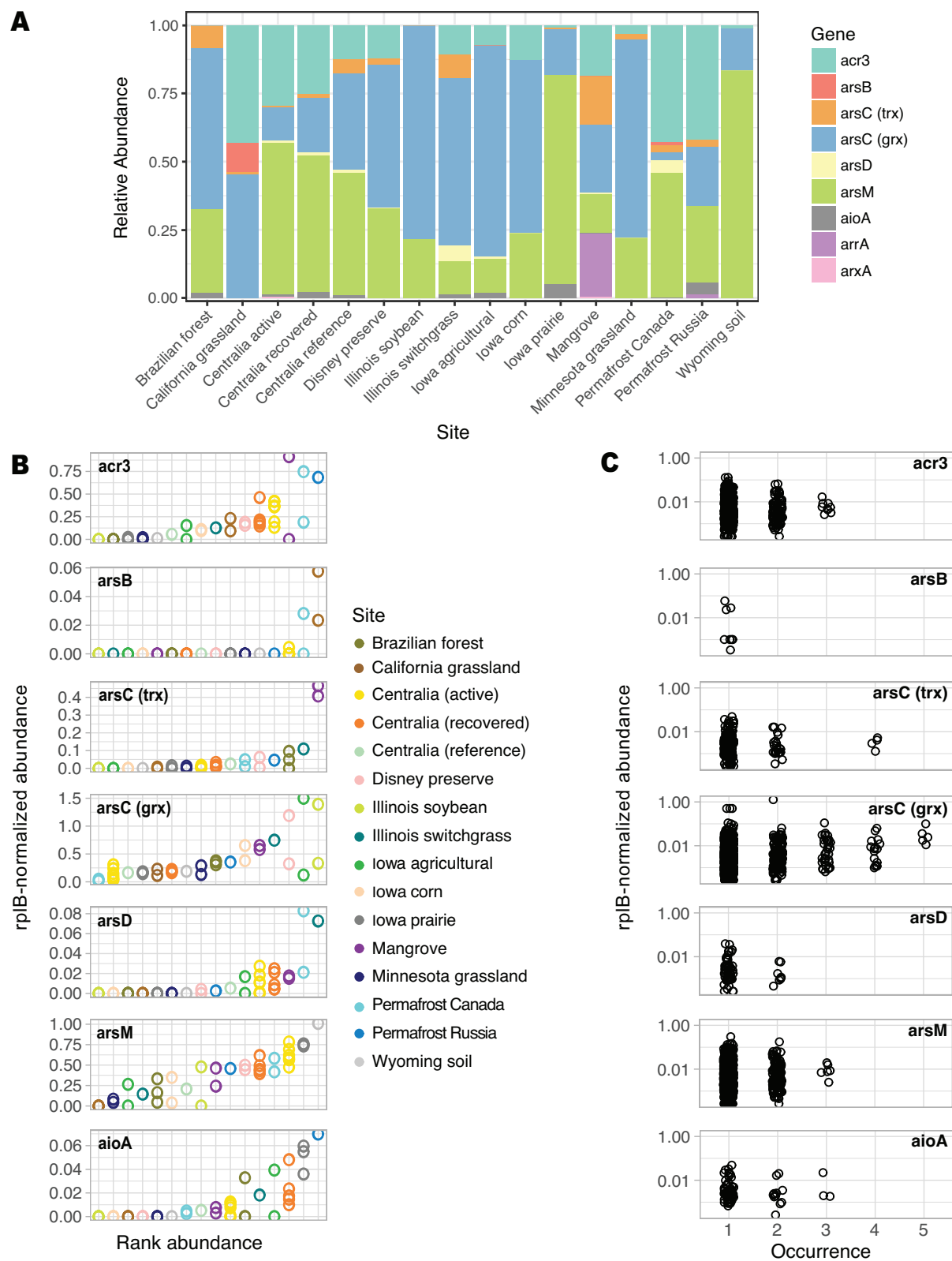
**A)** Mean normalized abundance of arsenic related genes based on RefSoil microorganisms abundance estimated from corresponding 16S rRNA gene abundance in Earth Microbiome Project datasets. Points are colored by soil order. **B)** Normalized abundance of arsenic resistance genes in RefSoil+ and 38 metagenomes. Metagenome abundance was normalized to *rplB*, and RefSoil+ normalized abundance was calculated using the number of RefSoil+ genomes. Only metagenomes with an arsenic resistance gene detected are shown, and the total number of datasets (including RefSoil+) is shown in parentheses. **C)** *rplB*-normalized abundance of arsenic resistance genes in cultivation dependent and independent metagenomes from the same soil sample.



Mann Whitney U test  $p < 0.05$ ). For genes encoding cytoplasmic arsenate reductases, *arsC* (grx) was more abundant than *arsC* (trx) (Mann Whitney U test  $p < 0.01$ ).

To gain a cultivation-independent perspective of the abundances of arsenic related genes, we examined their normalized abundance from soil metagenomes (**Figure 5.7B**). An undetected gene does not confirm absence, so we present a conservative estimate that only includes metagenomes testing positive for a gene. Arsenic detoxification genes (*acr3*, *arsB*, *arsC* (grx), *arsC* (trx), and *arsD*) were more abundant than arsenic metabolism genes (*aioA*, *arrA*, *arsM*, and *arxA*) (Mann Whitney-U test  $p < 0.01$ ; **Figure 5.7B**). Genes encoding arsenite efflux pumps differed in their abundance with *acr3* being more abundant than *arsB* (Mann Whitney U test  $p < 0.01$ ). We also observed differences in cytoplasmic arsenate reductases: *arsC* (grx) was more abundant than *arsC* (trx) (Mann Whitney U test  $p < 0.01$ ).

We explored cultivation bias of arsenic related genes with a case study comparing cultivation-dependent (lawn growth on the standard medium TSA50) and -independent communities from the same soil. Genes in the *ars* operon (*acr3*, *arsB*, *arsD*, and *arsC* (trx)) were elevated in the cultivation-dependent metagenome (**Figure 5.7C**). Additionally, arsenic metabolism genes were not detected (*aioA*, *arrA*, *arxA*) or in low abundance (*arsM*) in the cultivation-dependent sample; however, all four of these arsenic metabolism genes were detected in the cultivation-independent sample. Though this is a single case study of cultivation-dependent and independent methods, these results recapitulate the general discrepancies between RefSoil+ genomes and soil metagenomes (**Figure 5.7B**). This bias has important implications for studies focusing on arsenic bioremediation because cultivation-dependent studies could misestimate the potential of microbiomes for arsenic detoxification and metabolism *in situ*.



**Figure 5.8. Arsenic resistance and metabolism gene biogeography.**

**Figure 5.8 (cont'd)**

**A)** Relative abundance of arsenic resistance genes in soil metagenomes. **B)** Rank *rplB*-normalized abundance of arsenic related genes in soil metagenomes. Sites are ordered by rank mean abundance. Note the differences in y axes. **C)** Abundance-occurrence plots of arsenic related gene sequences clustered at 90% amino acid identity. Number of samples included are as follows Brazilian forest n = 3, California grassland n = 2, Centralia active n = 7, Centralia recovered n = 5, Centralia reference n = 1, Disney preserve n = 2, Illinois soybean n = 2, Illinois switchgrass n = 1, Iowa agricultural n = 2, Iowa corn n = 2, Iowa prairie n = 3, Mangrove n = 2, Minnesota grassland n = 2, Permafrost Canada n = 2, Permafrost Russia n = 1, and Wyoming soil n = 1.

### Arsenic related gene endemism

Arsenic related genes are globally distributed, but their biogeography is poorly understood. Broadly, arsenic related genes had comparable abundance among different soils (**Figure 5.7AB**). The relative distributions of distinct arsenic detoxification and metabolism mechanisms in one site, however, are relevant for predicting the impact of microbial communities on the fate of arsenic. To understand site-specific distributions, we explored soil metagenomes from Brazil, Canada, Malaysia, Russia, and the United States (**Appendix A Table 14**). These 16 sites had differences in community membership (**Appendix B Figure 20**) and arsenic related gene content (**Figure 5.8A**). Geographic location was not predictive of arsenic related gene content (Mantel's  $r = 0.03493$ ;  $p > 0.05$ ). Soils had different distributions of arsenic related genes and therefore differed in their potential impact on the biogeochemical cycling of arsenic. While *arsC* (grx) and *arsM* dominated most samples, their relative proportions varied greatly (**Figure 5.8A**). RefSoil+ data suggests that *arsM* can be found in Verrucomicrobia (100%,  $n = 2$ ), which is of particular importance for soil metagenomes since Verrucomicrobia are often underestimated with cultivation-dependent methods (245). The mangrove sample had the most even proportions of arsenic related genes (**Figure 5.8A**). This distribution was driven by a high abundance of *arsC* (trx) and *arrA*.

We further examined the arsenic resistance gene abundance at individual sites. We did not include *arr* and *arx* in this analysis due to limited available data. For each gene, the abundance varied greatly, but replicates within one site had similar abundances (**Figure 5.8B**). The majority of arsenic related gene sequences (99.3%) were endemic and only found in one to two sites, but 24 sequences were detected in three or more sites (**Figure 5.8C; Appendix A Table 16**). The majority (70.8%) of cosmopolitan sequences belonged to *arsC* (grx). This

analysis suggests that arsenic related genes *acr3*, *arsB*, *arsC* (trx), *arsD*, *arsM*, and *aioA* are generally endemic.

## Discussion

### *A bioinformatic toolkit for detecting and quantifying arsenic related genes*

We developed a toolkit for detecting arsenic related genes from sequence data that supports a variety of applications (**Figure 5.1A**): arsenic related genes can be detected in amino acid sequences from completed genomes (HMMs (198), BLAST (241)), nucleotide sequences in draft genomes (BLAST), as well as metagenomes and metatranscriptomes (Xander (237)). Because each tool relies on the same seed sequences, there is consistency and opportunity for comparison between sequence datasets that were generated from different sources. While primers already exist for arsenic related genes: *aioA* (130, 246), *acr3* (29), *arsB* (29), *arsC* (grx) (128), *arsC* (trx) (34), *arsM* (90), and *arrA* (129, 247, 248), these FunGene (164) databases can be used for testing primer breadth, designing new primers, and browsing sequences.

The toolkit is scalable for additional mechanisms for arsenic resistance and other functional genes of interest (e.g., methylarsenite oxidase (ArsH), C-As lyase (ArsI), trivalent organoarsenical efflux permease (ArsP), organoarsenical efflux permease (ArsJ) (1)), or redox transformations of elements involved in arsenic biogeochemical cycling (e.g., nitrate reductase (NarG) and sulfate reductase (DsrAB) (1, 14)). This toolkit serves as both a resource and an example workflow for developing similar toolkits to examine functional genes, beyond arsenic related genes, in microbial sequence datasets.

### Phylogenetic diversity and distribution of arsenic related genes

It has been conjectured that nearly all organisms have arsenic resistance genes (1), and though this assumption has propagated in the literature, it had never been explicitly quantified. Our data suggest that arsenic detoxification and metabolism genes are ubiquitous but not universal in RefSoil+ microorganisms (**Figure 5.2**). It is possible for these 132 organisms to have untested or novel arsenic related genes; nonetheless, these nine well-characterized genes were not universally detected. Additionally, phylogeny was predictive of the presence of *acr3*, *arsB*, *arsC* (grc), *arsC* (trx), and *arsM*. This correlation suggests that taxonomy is predictive of arsenic genotype despite documented potential for HGT (19, 30, 40, 226, 249). This result could be explained by ancient rather than contemporary HGT, as seen with *arsM* (19) and *arsC* (grx) (249). Therefore, we next assessed evidence for HGT by examining the phylogenetic congruence and genomic location (e.g. chromosome or plasmid) of arsenic related gene sequences.

Horizontal transfer of arsenic related genes has been well documented (19, 30, 40, 226, 249–251) and is an important consideration for understanding the propagation and taxonomic identity of arsenic related genes. We examined the phylogenetic diversity of arsenic related genes in RefSoil+ microorganisms, including plasmids and chromosomes, and compared them with the 16S rRNA gene taxonomy.

While known *acr3* sequences separates into two clades (30, 31, 86), plasmid-borne *acr3* sequences were present across clades, suggesting a potential for transfer across unrelated taxa. Therefore, studies assigning taxonomy to *acr3* in the absence of host information should consider the clade precisely and proceed with caution. Despite their functional redundancy as arsenite efflux pumps, *acr3* and *arsB* have very distinctive diversity. As compared with *acr3*,

*arsB* was less diverse and more phylogenetically conserved (**Figure 5.3B**; **Appendix B Figure 15**). This observation is in agreement with previous reports comparing the diversity of *arsB* to *acr3* (30, 86). Multiple, phylogenetically distinct copies of *arsB* were present in some RefSoil+ organisms, which could be due to an early gene duplication and subsequent diversification or to an early transfer event. Therefore, despite relatively lower sequence variation, this *arsB* phylogeny suggests an interesting evolutionary history that could be investigated further.

*arsC* (trx) was predominantly found on RefSoil+ chromosomes, not plasmids, suggesting vertical transfer of *arsC* (trx) is common. *arsC* (trx) was present in both Bacteria and Archaea, and sequences from the two domains formed two distinct clades. *arsC* (trx) sequences that cluster separately from Bacterial-*arsC* (trx) sequences have been documented in Thermococci, Archaeoglobi, Thermoplasmata, and Halobacteria (16). Together, this distribution supports an early evolutionary origin for *arsC* (trx). Thus, *arsC* (trx) appears to be an evolutionarily old enzyme that is phylogenetically conserved despite its presence on plasmids and potential for HGT. Plasmid-encoded *arsC* (grx) were also observed in RefSoil+ microorganisms, highlighting a contemporary potential for HGT that has been documented in soil (249). Thus, both genes encoding cytoplasmic arsenate reductases were more common on chromosomes.

The evolutionary history of the gene encoding arsenite S-adenosylmethionine methyltransferase, *arsM*, was recently investigated (19, 40). Both studies independently determined that *arsM* evolved billions of years ago and was subject to HGT (19, 40). In this work, *arsM* sequences from Euryarchaeota were dispersed throughout the *arsM* phylogeny, supporting the potential for inter-kingdom transfer events that were recently suggested (19, 40). Very few RefSoil+ organisms had arsenic metabolism genes *aioA*, *arrA*, or *arxA*, which limits

phylogenetic analysis. Nonetheless, they were mostly found in Proteobacteria, which is in agreement with previous work (21).

#### Cultivation bias and environmental distributions of arsenic related genes

Cultivation-based assessments of arsenic related gene content are important since cultivable strains are often favored for bioremediation (100). We estimated distributions of arsenic related genes in cultivable microorganisms from soils and found a greater abundance of arsenic detoxification genes *acr3*, *arsB*, and *arsC* (trx) (**Figure 5.7A**). A previous study also reported an abundance of *acr3* over *arsB* in cultivable microorganisms from forest soils and attributed this to the greater phylogenetic distribution of *acr3* compared with *arsB* (86). Additionally, they found that *arsC* (grx) was more abundant than *arsC* (trx) in cultivated microorganisms from these soils. It has been posited in cultivation-independent studies that *arsC* (trx) is more efficient than *arsC* (grx) and that high local arsenic concentrations result in a relatively greater abundance of *arsC* (trx) (31, 37). Our cultivation-dependent abundances suggest that *acr3* and *arsC* (grx), rather than *arsB* and *arsC* (trx), predominantly comprise the arsenic detoxification pathway in soils.

To assess arsenic related gene content without cultivation bias, we examined arsenic related genes in soil metagenomes. As predicted by cultivable organisms, arsenic metabolism genes (*aioA*, *arrA*, *arxA*) were generally in low abundance while *acr3* and *arsC* (grx) were in high abundance. Estimates of genes encoding arsenic detoxification (*acr3*, *arsB*, *arsD*, *arsC* (grx), *arsC* (trx)) were considerably lower in these cultivation-independent samples. This result could be due, in part, to the large number of RefSoil+ microorganisms with multiple copies of these genes (**Appendix B Figure 19**). Cultivation-independent genomes (e.g., single-cell



amplified genomes and metagenome-assembled genomes) could provide greater context about the environmental distributions of copy numbers of arsenic related genes.

Notably, *arsM* was abundant in soil (median 48%), which greatly exceeds cultivation-dependent estimations, and in a case-study of cultivation dependent and independent techniques, *arsM* was more abundant in the cultivation independent sample (**Figure 5.7C**). Due to the early phylogenetic origins of *arsM* and its independent functionality (19), this abundance of *arsM* in soil metagenomes is not unexpected. *arsM* is typically studied in paddy soils (221, 252, 253), but metagenomes in this study suggest it is an important component of the arsenic biogeochemical cycle in a variety of soils.

#### Arsenic related gene endemism

We examined the relative abundance of arsenic related genes in soil metagenomes and observed differences in genetic potential for arsenic transformation that could impact biogeochemical cycling (**Figure 5.8A**). Notably, the mangrove sample had the most even proportions of arsenic related genes. While the arsenic concentrations in this sample are unknown, mangroves are considered sources and sinks for arsenic (254–256). This could explain the greater abundance of *arsC* (*trx*), which is hypothesized to be more abundant in high arsenic sites (31, 37). Additionally, *arrA* encodes a dissimilatory arsenate reductase that functions in an anaerobic environment (240), so its greater abundance in sediment is expected. Soil geochemical data was not available for all metagenomes examined in this work, so direct comparisons of arsenic related gene content and soil geochemistry were not possible. This highlights the importance and utility of depositing geochemical data with DNA sequences. Future work,

however, could further examine relationships between arsenic resistance genes and soil geochemical data, including arsenic concentration and redox potential.

We also measured whether arsenic related gene sequence variants were endemic or cosmopolitan in soil metagenomes (**Figure 5.8C**). We found that genes *acr3*, *arsB*, *arsC* (*trx*), *arsD*, *arsM*, and *aioA* were generally endemic, suggesting regional dispersal limitation. Only one *aioA* and three *acr3* sequences were detected in multiple sites. This supports a previous finding that *acr3* and *aioA* from the acid mine drainage in Carnoulès were endemic (257). Conversely, *arsC* (*grx*) was cosmopolitan which could suggest genetic migration via HGT or vertical transfer and a limited gene diversification. Both are plausible since *arsC* (*grx*) was common in RefSoil+ plasmids and had low phylogenetic diversity (**Figure 5.4B**; **Appendix B Figure 17**).

### Conclusions

We developed a bioinformatic toolkit for detecting arsenic detoxification and metabolism genes in microbial sequence data and applied it to analyze the genomes and metagenomes from soil microorganisms. This toolkit informs hypotheses about the evolutionary histories of these genes (including potential for vertical and horizontal transfers) and how community ecology *in situ* may influence their prevalence and distribution. This study reports the phylogenetic diversity, genomic locations, and biogeography of arsenic related genes in soils, integrating information from different ‘omics datasets and resources to provide a broad synthesis. The toolkit and the synthesis presented here can catalyze future work to understand the ecology and evolution of microbial arsenic biogeochemistry. Furthermore, the toolkit acts as a framework for similar studies of other functional genes of interest.

## **CHAPTER 6 : Conclusions and future directions**

## Summary

This dissertation examined the impact of disturbance on the soil microbiome using the Centralia, PA, coalmine fire as a model disturbance (Chapters 2 – 4). Chapter 2 showed that Arsenic resistant strains can be isolated from the low arsenic (1.5 ppm) surface soils in Centralia, PA. We found that arsenic resistant isolates are in low abundance (rare) in this soil system. This finding is important for bioremediation as it suggests that the rare biosphere could contribute to microbiome dynamics during arsenic stress. Chapter 3 showed that antibiotic resistance genes related to clinical treatments decreased with disturbance in Centralia. This change was related to concomitant changes in the soil microbiome composition, suggesting that reduced microbiome diversity decreases the abundance clinically-relevant ARGs in the environment. This body of work exemplifies the utility of examining functional genes along disturbance gradients.

This dissertation also examined antibiotic and arsenic related genes in soil microbiomes, broadening the scope of this work beyond Centralia. Chapter 4 detailed RefSoil+, a new database of soil-associated chromosomes and plasmids. This database showed that antibiotic resistance genes were less common on soil-associated plasmids compared with plasmids in general, suggesting reduced potential for plasmid-mediated HGT in soil microbiomes. This observation broadens understanding of plasmids in soil and provides a community resource for investigating clinic-environment dynamics of important plasmid-associated genes. Chapter 5 described a new bioinformatic toolkit to detect arsenic related genes in genomes and metagenomes. This toolkit showed the phylogenetic diversity, genomic locations, and biogeography of arsenic related genes in soils, integrating information from different ‘omics datasets and resources to provide a broad synthesis. Arsenic related genes were widespread in soil, suggesting that endemic microbiomes could be harnessed for bioremediation.

Broadly, this dissertation highlights the importance of ecological phenomena on dynamics and distributions of functional genes in soil microbiomes. It showed how environmental disturbances can impact functional genes through changes in community structure. Furthermore, data from this work suggest that soils offer limited opportunity for plasmid-mediated transfer of ARGs. This finding has important implications for ARGs whose environmental proportions are relevant to public health. The tools created in this work (RefSoil+ and arsenic toolkit) can be applied to future research questions concerning these important functional genes. Furthermore, ecological insights from this work can be used to model dynamics of these functional genes in the environment. Ultimately, this set of findings can allow for manipulation of endemic microbiomes for public health outcomes such as arsenic bioremediation or reduction of antibiotic resistance.

## **Future Directions**

An ambitious goal is to be able to predict and manage the outcomes of arsenic resistance and metabolism as it relates to arsenic bioremediation or biogeochemistry. Ecological aspects of microbial communities, including rarity, dormancy, and dispersal, influence microbial community structure and can impact local arsenic resistance gene content and activity. These factors are expected to exhibit considerable variation over time and are sensitive to environmental changes (258). Longitudinal studies that examine the contributions of these factors to arsenic resistance and metabolism will inform prediction and, potentially, management of arsenic outcomes in the environment.

Eco-evolutionary processes, such as HGT, play an important role in the history and dissemination of arsenic related genes (19, 40). While studies frequently cite evidence of HGT as

inferred from environmental samples (34), it is unclear whether these transfers are due to historical evolutionary outcomes (19, 40, 210, 250) or recent endemic transfers. Efforts should be made to determine contemporary rates of arsenic related gene HGT. Furthermore, metagenomic studies should exert care when assigning taxonomy to arsenic related genes, particularly for sequence variants that are known to be phylogenetically widespread. Relatedly, many studies highlight the co-selection of arsenic (and heavy metal) related genes with ARGs (65, 66, 259), and increases in antibiotic resistance genes have been observed in metal(loid) contaminated environments (68–70). It is unclear, however, whether the reverse phenomenon, where antibiotics increase metal(loid) resistance genes, occurs. This is another direction ripe for interrogation.

Determining the distribution of ARGs on plasmids in the environment is another potential future direction. This dissertation showed that ARGs are more common on soil associated plasmids than chromosomes in cultivable microorganisms, but it is unclear whether this localization is true for soil microbiomes which may differ in plasmid content due to variations in community membership, presence of uncultivable microorganisms, and variations in plasmid content and copy number. As methodologies such as pre-sequencing proximity linkage (e.g. Hi-C) (192), long-read technology (215), single cell sequencing (216), and other technologies advance, it will be possible to estimate the genomic distribution of ARGs in the environment and ultimately determine the influence of the soil microbiome on clinical antibiotic resistance.

## **APPENDICES**

## **APPENDIX A: Supplementary Tables**



**Appendix A Table 1. Soil geochemical data.**

<b>Air temperature (°C)</b>	<b>Soil temperature (°C)</b>	<b>Organic matter (360°C)</b>	<b>Organic matter (500°C)</b>	<b>NO<sub>3</sub><sup>-</sup> (ppm)</b>	<b>NO<sub>4</sub><sup>+</sup> (ppm)</b>	<b>pH</b>	<b>S (ppm)</b>	<b>K (ppm)</b>	<b>Ca (ppm)</b>	<b>Mg (ppm)</b>	<b>Fe (ppm)</b>	<b>As (ppm)</b>
13.3	57.4	3.1	7.1	4.6	1.7	8	28	37	2545	114	67.1	2.58


























**Appendix A Table 2. Degenerate primers used for end point PCR.**

Gene	Primer Sequence (5'-3')	Name	Source
16S	AGAGTTTGATCCTGGCTCAG	Uni-27F	Weisburg <i>et al.</i> , 1991
16S	GGTACCTTGTTACGACTT	Uni-1492R	Weisburg <i>et al.</i> , 1991
16S	GTGCCAGCMGCCGCGGTAA	U515F	Baker <i>et al.</i> , 2003
<i>arsB</i>	GGTGTGGAACATCGTCTGGAAYGCNAC	darsB1F	Achour <i>et al.</i> , 2007
<i>arsB</i>	CAGGCCGTACACCACCAGRTACATNCC	darsB1R	Achour <i>et al.</i> , 2007
<i>ACR3(1)</i>	GCCATCGGCCTGATCGTNATGATGTAYCC	dacr1F	Achour <i>et al.</i> , 2007
<i>ACR3(1)</i>	CGGCG ATGGCCAGCTCYAAYTTYTT	dacr1R	Achour <i>et al.</i> , 2007
<i>ACR3(2)</i>	TGA TCTGGGTCATGATCTTCCC VATGMTGVT	dacr5F	Achour <i>et al.</i> , 2007
<i>ACR3(2)</i>	CGGCCACG GCCAGYTCRAARAARTT	dacr4R	Achour <i>et al.</i> , 2007
<i>arsC</i>	TCGCGTAATACGCTGGAGAT	amlt-42-f	Sun <i>et al.</i> , 2004
<i>arsC</i>	ACTTTCTCGCCGTCTTCCTT	amlt-376-r	Sun <i>et al.</i> , 2004
<i>arsC</i>	TCACGCAATACCCTTGAAATGATC	smrc-42-f	Sun <i>et al.</i> , 2004
<i>arsC</i>	ACCTTTTCACCGTCCTCTTTCGT	smrc-376-r	Sun <i>et al.</i> , 2004
<i>arsC</i>	AGCCAAATGGCAGAAGC	P52F	Cavalca, <i>et al.</i> , 2010
<i>arsC</i>	GCTGGRTCRTCAAATCCCCA	P323R	Cavalca, <i>et al.</i> , 2010
<i>arrA</i>	CGAAGTTCGTCCCGATHACNTGG	AS1F	Song <i>et al.</i> , 2009
<i>arrA</i>	GGGGTGCGGTCYTTNARYTC	AS1R	Song <i>et al.</i> , 2009
<i>arrA</i>	GTCCCNATBASNTGGGANRARGCNMT	AS2F	Song <i>et al.</i> , 2009
<i>arrA</i>	ATANGCCARTGNCCYTGN	AS2R	Song <i>et al.</i> , 2009
<i>aioA</i>	CCACTTCTGCATCGTGGGNTGYGGNTA	aoxBM1-2F	Quemeneur <i>et al.</i> , 2008
<i>aioA</i>	TGTCGTTGCCCCAGATGADNCCYTTYTC	aoxBM3-2R	Quemeneur <i>et al.</i> , 2008
<i>arsM</i>	TCYCTCGGCTGCGGCAAYCCVAC	arsMF1	Jia <i>et al.</i> , 2013
<i>arsM</i>	GTGCTCGAYCTSGGCWCCGGC	arsMF2	Jia <i>et al.</i> , 2013
<i>arsM</i>	GGCATCGACGTGCTKCTBTCSGC	arsMF3	Jia <i>et al.</i> , 2013
<i>arsM</i>	AGGTTGATGACRCAGTTWGAGAT	arsMR1	Jia <i>et al.</i> , 2013
<i>arsM</i>	CGWCCGCCWGGCTTWAGYACCCG	arsMR2	Jia <i>et al.</i> , 2013
<i>arsM</i>	GCGCCGGCRAWGCAGCCWACCCA	arsMR3	Jia <i>et al.</i> , 2013

**Appendix A Table 3. Isolates with short *arsC* sequences (< 200 bp).**

Isolate	<i>arsC</i> sequence
A2707	cgatgctgatttagtctgttacgctttgtggccatgcggatgctgttgtccgtctactccgccgcatgtgaatcgagttcactggggatttgacgaccagcaa
A2723	gctggatttagtctgtnacncttgggtcacgcagatgctgtctgtccnncaacacctccgcacgtgaancgagttcactggggatttgacgaccagcaa
A2733	caatgaanatcnggatggtttgattccgttatgatcgtgtctacttcgttcattgctttaattgcnttcggattcactccgtgtgcctcnataccgcagaaantac
A2735	cgatgctgatttagtctgttacgctttgtggtcacgcagatgctgtctgtccnncaacacctcctcacgtgaancgagttcactggggatttgacgaccagcaaa

**Appendix A Table 4. Phenotypes of arsenic resistant isolates.**

Isolate	Closest 16S rRNA gene sequence described (% similarity)	Colony Morphology	Temperature Maximum (°C)	Length (µm)	Width (µm)
I2706	<i>Enterobacter absuriae</i> JM 6051 (99.43%)		44.3	1.43	1.21
I2707	<i>Enterobacter absuriae</i> JM 6051 (99.35%)		44.3	1.34	1.16
I2716	<i>Bacillus nealsonii</i> DSM 150-7577 (99.49%)		44.3	4.63	1.16
I2723	<i>Bacillus anthracis</i> ATCC 14578 (100%)		44.3	4.46	1.36
I2726	<i>Enterobacter absuriae</i> JM 6051 (99.5%)		44.3	2.37	1.45
I2727	<i>Enterobacter absuriae</i> JM 6051 (99.56%)		44.3	2.89	1.05
I2742	<i>Bacillus nealsonii</i> DSM 150-7577 (99.49%)		44.3	2.60	0.85
I2745	<i>Bacillus anthracis</i> ATCC 14578 (99.86%)		44.3	4.11	0.90
I2746	<i>Paenibacillus xylanilytius</i> XIL14 (98.63%)		44.3	1.68	1.46
I2747	<i>Paenibacillus xylanilytius</i> XIL14 (98.58%)		39.7	3.96	1.12
I2748	<i>Mirobacterium paraoxydans</i> F36 (99.85%)		47.7	1.19	1.14
I2749	<i>Olivibater oleidegrans</i> TBF2/20.2 (99.42%)		44.3	1.58	1.23
I2759	<i>Acinetobacter baumannii</i> ATCC 19606 (99.78%)		44.3	1.20	1.16
A2705	<i>Acinetobacter baumannii</i> ATCC 19606 (99.64%)		44.3	1.25	1.19
A2706	<i>Enterobacter absuriae</i> JM 6051 (99.50%)		44.3	2.07	1.10
A2707	<i>Bacillus anthracis</i> ATCC 14578 (100%)		44.3	3.83	0.91
A2708	<i>Bacillus subtilis subsp. inoquosorum</i> KT 13429 (99.93%)		52.0	3.26	0.90
A2712	<i>Pseudomonas hibisicola</i> ATCC 19867 (99.36%)		39.7	2.01	1.01
A2716	<i>Acinetobacter baumannii</i> ATCC 19606 (99.78%)		44.3	1.12	0.96
A2723	<i>Bacillus anthracis</i> ATCC 14578 (99.73%)		44.3	3.20	1.64
A2724	<i>Enterobacter absuriae</i> JM 6051 (99.57%)		44.3	1.19	1.09
A2727	<i>Pseudomonas geniculata</i> ATCC 19374 (99.78%)		39.7	1.23	0.79
A2731	<i>Enterobacter absuriae</i> JM 6051 (99.49%)		44.3	1.18	1.15
A2733	<i>Bacillus subtilis subsp. inoquosorum</i> KT 13429 (99.93%)		52.0	3.61	0.91
A2735	<i>Bacillus anthracis</i> ATCC 14578 (99.85%)		44.3	4.30	2.04

**Appendix A Table 5. Comparison of arsenic resistance gene sequences with NCBI references.**

Isolate	Genus	Gene	Closest NCBI match (% similarity)	%GC gene	%GC reference genomes
I2706	<i>Enterobacter</i>	<i>arsB</i>	<i>Enterobacter cloacae</i> isolate MBRL1077 (97%)	59.62	55 ± 0.37
I2707	<i>Enterobacter</i>	<i>arsB</i>	<i>Enterobacter cloacae</i> isolate MBRL1077 (97%)	59.42	55 ± 0.37
I2726	<i>Enterobacter</i>	<i>arsB</i>	<i>Enterobacter cloacae</i> isolate MBRL1077 (97%)	59.51	55 ± 0.37
I2727	<i>Enterobacter</i>	<i>arsB</i>	<i>Enterobacter cloacae</i> isolate MBRL1077 (97%)	59.83	55 ± 0.37
I2759	<i>Acinetobacter</i>	<i>arsB</i>	<i>Enterobacter cloacae</i> isolate MBRL1077 (97%)	59.69	39.03 ± 0.11
A2706	<i>Enterobacter</i>	<i>arsB</i>	<i>Enterobacter cloacae</i> isolate MBRL1077 (97%)	59.14	55 ± 0.37
A2724	<i>Enterobacter</i>	<i>arsB</i>	<i>Enterobacter cloacae</i> isolate MBRL1077 (97%)	59.78	55 ± 0.37
A2731	<i>Enterobacter</i>	<i>arsB</i>	<i>Enterobacter cloacae</i> isolate MBRL1077 (97%)	60	55 ± 0.37
A2712	<i>Pseudomonas</i>	<i>ACR3(2)</i>	<i>Stenotrophomonas maltophilia</i> strain ISMMS2 (84%)	62.96	66.54 ± 0.25
A2727	<i>Pseudomonas</i>	<i>ACR3(2)</i>	<i>Stenotrophomonas maltophilia</i> strain ISMMS2 (95%)	64.69	66.54 ± 0.25
A2733	<i>Bacillus cereus</i>	<i>ACR3(2)</i>	<i>Stenotrophomonas maltophilia</i> D457 (83%)	64.83	43.89 ± 0.71
I2716	<i>Bacillus nealsonii</i>	<i>arsC</i>	<i>Bacillus cereus</i> ATCC 14579 (95%)	39.5	35.1
I2723	<i>Bacillus cereus</i>	<i>arsC</i>	<i>Bacillus cereus</i> ATCC 10987 (96%)	41.2	35.26 ± 0.18
I2726	<i>Enterobacter</i>	<i>arsC</i>	<i>Bacillus cereus</i> F837/76 (99%)	40.97	55 ± 0.37
I2727	<i>Enterobacter</i>	<i>arsC</i>	<i>Bacillus cereus</i> ATCC 10987 (96%)	40.89	55 ± 0.37
I2745	<i>Bacillus cereus</i>	<i>arsC</i>	<i>Bacillus thuringiensis</i> strain KNU-07 (98%)	37.67	35.26 ± 0.18
I2746	<i>Paenibacillus</i>	<i>arsC</i>	<i>Bacillus cereus</i> strain A1 (98%)	37.8	50.9 ± 0.14
I2747	<i>Paenibacillus</i>	<i>arsC</i>	<i>Bacillus</i> sp. ABP14 (95%)	39.29	50.9 ± 0.14
A2707	<i>Bacillus cereus</i>	<i>arsC</i>	<i>Bacillus cereus</i> D17 (92%)	43.28	35.26 ± 0.18
A2708	<i>Bacillus subtilis</i>	<i>arsC</i>	<i>Bacillus anthracis</i> strain Tyrol 4675 (95%)	40.38	43.89 ± 0.71
A2723	<i>Bacillus cereus</i>	<i>arsC</i>	<i>Bacillus</i> sp. CH19 (86%)	42.79	35.26 ± 0.18
A2733	<i>Bacillus subtilis</i>	<i>arsC</i>	<i>Bacillus cereus</i> strain FORC_024 (88%)	40.1	43.89 ± 0.71
A2735	<i>Bacillus cereus</i>	<i>arsC</i>	<i>Bacillus</i> sp. CH19 (88%)	42.29	35.26 ± 0.18

**Appendix A Table 6. Parameters for reference gene (FunGene) database construction and gene-targeted assembly for each protein of interest.**

FunGene Database	Protein	Minimum HMM score	Minimum length (aa)	Minimum HMM coverage (%)	Number of FunGene sequences	Number of dereplicated sequences	Minimum length (aa)
Chloramphenicol efflux pump	CmlA	298	390	80	3747	491	150
Dfra1	Dfra1	100	135	80	4659	211	50
Dfra12	Dfra12	90	130	80	26637	1252	50
IntI	IntI	90	315	80	9418	2562	150
RepA	RepA	400	220	80	387	31	150
Resfam_AAC6-Ia	AAC6-Ia	100	170	80	757	112	100
Resfam_AdeB	AdeB	1400	1000	80	53493	10025	150
Resfam_ANT3	ANT3	310	245	80	7806	790	150
Resfam_ANT6	ANT6	130	260	80	4097	1066	150
Resfam_ANT9	ANT9	400	245	80	4044	41	150
Resfam_Chloramphenicol Acetyltransferase CAT	CAT	195	200	80	9996	1299	150
Resfam_ClassA	ClassA	179	275	80	34258	5713	150
Resfam_ClassB	ClassB	76	255	80	9853	2087	150
Resfam_ClassC	ClassC	400	370	80	12916	3641	150
Resfam_ermB	Resfam_ermB	400	200	80	2090	182	100
Resfam_ermC	Resfam_ermC	265	200	80	7173	246	100
Resfam_MexC	MexC	300	340	80	2569	720	150
Resfam_MexE	MexE	400	390	80	1567	665	150
Resfam_Quinolone Resistance Protein Qnr	Qnr	230	200	80	2562	558	100
Resfam_tetA	TetA	680	390	80	2060	70	150
Resfam_tetD	TetD	795	350	80	261	9	150
Resfam_tetX	TetX	300	360	80	227	112	150
Resfam_TolC	TolC	350	430	80	19431	3189	150
Resfam_vanA	VanA	700	300	80	250	28	150
Resfam_vanC	VanC	730	300	80	35	29	150
Resfam_vanH	VanH	500	280	80	438	61	150
Resfam_vanT	VanT	600	650	80	304	97	150
Resfam_vanW	VanW	130	220	80	1311	423	150
Resfam_vanX	VanX	100	150	80	16689	2340	100
Resfam_vanY	VanY	220	300	80	250	35	150
Resfam_vanZ	VanZ	80	120	80	1042	189	100
StrA	StrA	400	230	80	4286	154	150
StrB	StrB	159	230	80	4695	222	150
tet_sul2	Sul2	200	245	80	9031	298	150
TetM	TetM	1175	600	80	5531	543	150
TetQ	TetQ	650	600	80	242	70	150
TetW	TetW	1260	600	80	345	169	150

**Appendix A Table 7. Sequencing depth and Nonpareil-estimated coverage of *Centralia* metagenomes.**

<b>Site name</b>	<b>Fire History</b>	<b>Sequencing depth (Gbases)</b>	<b>Coverage (%)</b>
Cen01	Recovered	23	58.96
Cen03	Recovered	26	49.49
Cen04	Recovered	25	38.32
Cen05	Recovered	25	45.97
Cen06	Fire-affected	22	54.23
Cen07	Recovered	21	53.26
Cen10	Fire-affected	36	89.96
Cen12	Fire-affected	24	88.63
Cen14	Fire-affected	24	82.79
Cen15	Fire-affected	20	76.48
Cen16	Fire-affected	51	76.30
Cen17	Reference	24	29.12

**Appendix A Table 8. Sample site characteristics and measured soil geochemical data.**

Sample	latitude	longitude	Soil temperature (°C)	Classification	Date since fire (Elick 2011)	Organic matter (500°C)	NO3- (ppm)	NH4- (ppm)	pH	S (ppm)	K (ppm)	Ca (ppm)	Mg (ppm)	Fe (ppm)	As (ppm)	Soil Mois (%)
Cen01	40 47.926	076 20.357	14.1	Recovered	1982	3.9	0.7	2.2	4.7	1	35	194	61	48.6	2.63	7
Cen03	40 47.881	076 20.468	14.7	Recovered	2002	48.9	0.3	3.2	4.5	4	31	1416	241	54.7	7.1	5
Cen04	40 47.870	076 20.489	13.3	Recovered	1999	12.8	0.8	5	4.6	23	34	103	46	167.2	3.6	4
Cen05	40 47.831	076 20.572	14.0	Recovered	2009	25.4	5.7	5	4.1	6	43	63	43	164.5	1.75	27
Cen06	40 47.849	076 20.506	24.1	Fire affected	2014	11.9	0.8	3.2	4.7	4	46	111	52	75.6	2.05	17
Cen07	40 48.086	076 20.736	13.5	Recovered	2005	6	0.7	4	4.6	14	40	78	37	108.9	5.56	242
Cen10	40 48.062	076 20.582	54.2	Fire affected	2007	24.5	98.4	120.6	4	21	57	245	70	508	3.79	111
Cen12	40 48.078	076 20.589	32.0	Fire affected	2009	6	0.2	2	4.8	7	24	51	30	150.3	3.9	70
Cen14	40 48.040	076 20.469	34.1	Fire affected	2002	21.9	0.9	2.7	5	5	58	394	64	102.7	2.97	64
Cen15	40 48.045	076 20.489	38.9	Fire affected	2002	9.6	1.1	4	5.2	13	44	224	50	93.3	2.25	119
Cen16	40 48.048	076 20.487	21.7	Fire affected	2002	10.6	0.5	1.2	5.6	8	33	497	56	80.8	3.57	78
Cen17	40 47.998	076 20.416	12.1	Reference	NA	6.1	0.1	3.3	5.7	6	99	652	73	48.6	1.99	27



**Appendix A Table 9. Spearman's rank correlation between relative abundance of ARGs and soil temperature.**

Significant differences are bolded.

Gene	Spearman's rho	p value
AAC6-Ia	0.119326331	0.711849439
<b>ClassA</b>	<b>-0.594405594</b>	<b>0.0457531</b>
<b>ClassB</b>	<b>-0.741258741</b>	<b>0.00817064</b>
ClassC	-0.035212141	0.91348822
CEP	0.204271555	0.524238954
intI	-0.013986014	0.973693904
adeB	-0.045765054	0.887688984
<b>tolC</b>	<b>-0.697619343</b>	<b>0.011658884</b>
sul2	0.542908013	0.068150085
<b>dfra12</b>	<b>-0.706293706</b>	<b>0.013286114</b>
vanA	-0.108581603	0.736943708
vanH	-0.169018276	0.59949787
vanX	-0.125874126	0.699712221
vanZ	-0.270429867	0.39525819

**Appendix A Table 10. Correlations between ARG phylum and Proteobacteria class normalized abundances and soil temperature.**

Spearman's rank correlations with *rplB*-normalized abundance and soil temperature.

Significant correlations ( $p < 0.05$ ) are bolded.

Gene	Phylum	Spearman's rho	p value
ClassA	Acidobacteria	-0.045765054	0.887688984
<b>ClassA</b>	<b>Actinobacteria</b>	<b>0.66567352</b>	<b>0.018134057</b>
ClassA	Alphaproteobacteria	-0.293706294	0.354332534
<b>ClassA</b>	<b>Betaproteobacteria</b>	<b>-0.8048763</b>	<b>0.001588848</b>
ClassA	Deltaproteobacteria	-0.425485705	0.167901396
ClassA	Gammaproteobacteria	-0.448469578	0.143665129
ClassB	Acidobacteria	-0.440559441	0.15421575
ClassB	Alphaproteobacteria	-0.004160459	0.989761605
ClassB	Bacteroidetes	-0.086009076	0.790410909
ClassB	Betaproteobacteria	0.043671315	0.892800448
ClassB	Deltaproteobacteria	0.550736912	0.063497906
ClassB	Gammaproteobacteria	-0.384615385	0.218387427
ClassB	Gemmatimonadetes	0.393041832	0.206255757
ClassB	Verrucomicrobia	0.263402795	0.408125175
dfra12	Acidithiobacillia	-0.354787438	0.257796301
dfra12	Actinobacteria	0.070727811	0.827099003
<b>dfra12</b>	<b>Alphaproteobacteria</b>	<b>-0.734265734</b>	<b>0.009052097</b>
<b>dfra12</b>	<b>Bacteroidetes</b>	<b>-0.683115531</b>	<b>0.014338998</b>
<b>dfra12</b>	<b>Betaproteobacteria</b>	<b>-0.690018571</b>	<b>0.013012243</b>
dfra12	C. Peregrinibacteria	-0.305699203	0.333893136
dfra12	Deinococcus-Thermus	-0.082610537	0.798539196
dfra12	Deltaproteobacteria	-0.514830787	0.08675862
dfra12	Firmicutes	-0.608391608	0.040002049
dfra12	Gammaproteobacteria	0.202797203	0.528100237
dfra12	Microgenomates	0.108237592	0.737751195
dfra12	Parcubacteria	-0.330442147	0.294152738
dfra12	Verrucomicrobia	-0.156042125	0.628187113
intl	Acidithiobacillia	0.043671315	0.892800448
intl	Betaproteobacteria	-0.475524476	0.121319356
intl	Chlorobi	0.241900526	0.448765409
intl	Chloroflexi	0.06425264	0.842745428
intl	Cyanobacteria	0.366606083	0.241141519
intl	Deltaproteobacteria	-0.364273764	0.244376487
intl	Gammaproteobacteria	-0.496503497	0.104092833
intl	Gemmatimonadetes	0.319717875	0.311035387
intl	Nitrospirae	0.517482517	0.088650879
intl	Planctomycetes	0.458947427	0.133408824
intl	Verrucomicrobia	0.104011487	0.747690649

**Appendix A Table 11. Correlations between ARG phylum and Proteobacteria class relative abundances and soil temperature.**

Spearman's rank correlations with relative abundance and soil temperature. Significant correlations ( $p < 0.05$ ) are bolded.

Gene	Phylum	Spearman's rho	p value
ClassA	Acidobacteria	-0.045765054	0.887688984
<b>ClassA</b>	<b>Actinobacteria</b>	<b>0.66567352</b>	<b>0.018134057</b>
ClassA	Alphaproteobacteria	-0.293706294	0.354332534
<b>ClassA</b>	<b>Betaproteobacteria</b>	<b>-0.8048763</b>	<b>0.001588848</b>
ClassA	Deltaproteobacteria	-0.425485705	0.167901396
ClassA	Gammaproteobacteria	-0.448469578	0.143665129
ClassB	Acidobacteria	-0.440559441	0.15421575
ClassB	Alphaproteobacteria	-0.004160459	0.989761605
ClassB	Bacteroidetes	-0.086009076	0.790410909
ClassB	Betaproteobacteria	0.043671315	0.892800448
ClassB	Deltaproteobacteria	0.550736912	0.063497906
ClassB	Gammaproteobacteria	-0.384615385	0.218387427
ClassB	Gemmatimonadetes	0.393041832	0.206255757
ClassB	Verrucomicrobia	0.263402795	0.408125175
dfra12	Acidithiobacillia	-0.354787438	0.257796301
dfra12	Actinobacteria	0.070727811	0.827099003
<b>dfra12</b>	<b>Alphaproteobacteria</b>	<b>-0.734265734</b>	<b>0.009052097</b>
<b>dfra12</b>	<b>Bacteroidetes</b>	<b>-0.683115531</b>	<b>0.014338998</b>
<b>dfra12</b>	<b>Betaproteobacteria</b>	<b>-0.690018571</b>	<b>0.013012243</b>
dfra12	C. Peregrinibacteria	-0.305699203	0.333893136
dfra12	Deinococcus-Thermus	-0.082610537	0.798539196
dfra12	Deltaproteobacteria	-0.514830787	0.08675862
dfra12	Firmicutes	-0.608391608	0.040002049
dfra12	Gammaproteobacteria	0.202797203	0.528100237
dfra12	Microgenomates	0.108237592	0.737751195
dfra12	Parcubacteria	-0.330442147	0.294152738
dfra12	Verrucomicrobia	-0.156042125	0.628187113
intl	Acidithiobacillia	0.043671315	0.892800448
intl	Betaproteobacteria	-0.475524476	0.121319356
intl	Chlorobi	0.241900526	0.448765409
intl	Chloroflexi	0.06425264	0.842745428
intl	Cyanobacteria	0.366606083	0.241141519
intl	Deltaproteobacteria	-0.364273764	0.244376487
intl	Gammaproteobacteria	-0.496503497	0.104092833
intl	Gemmatimonadetes	0.319717875	0.311035387
intl	Nitrospirae	0.517482517	0.088650879
intl	Planctomycetes	0.458947427	0.133408824
intl	Verrucomicrobia	0.104011487	0.747690649
rplB	Acidobacteria	-0.412587413	0.184480685
rplB	Actinobacteria	0.195804196	0.542873521
rplB	Alphaproteobacteria	0.251748252	0.430115289
<b>rplB</b>	<b>Aquificae</b>	<b>-0.84354992</b>	<b>0.000564148</b>

## Appendix A Table 11 (cont'd)

rplB	Bacteroidetes	-0.53522454	0.072939241
<b>rplB</b>	<b>Betaproteobacteria</b>	<b>-0.664335664</b>	<b>0.022159207</b>
rplB	Caldiserica	0.043671315	0.892800448
rplB	C. Saccharibacteria	0.17935393	0.577012393
rplB	Chlamydiae	-0.128974244	0.689537493
rplB	Chlorobi	0.137295163	0.67047431
<b>rplB</b>	<b>Chloroflexi</b>	<b>0.727272727</b>	<b>0.010000917</b>
rplB	Cyanobacteria	0.082610537	0.798539196
rplB	Deltaproteobacteria	-0.485927542	0.109227185
rplB	Dictyoglomi	-0.480384461	0.113937412
rplB	Elusimicrobia	0.431410581	0.161423074
rplB	Epsilonproteobacteria	-0.530263793	0.076151068
rplB	Firmicutes	0.223776224	0.48491114
rplB	Gammaproteobacteria	0.251748252	0.430115289
rplB	Gemmatimonadetes	-0.433566434	0.161446426
rplB	Ignavibacteriae	0.004160459	0.989761605
rplB	metagenomes	-0.188811189	0.557827775
rplB	Nitrospirae	-0.119326331	0.711849439
rplB	Planctomycetes	0.083916084	0.800197518
rplB	Synergistetes	0.197187988	0.53902672
rplB	Thermobaculum	0.626433696	0.029292738
rplB	Thermodesulfobacteria	-0.470131923	0.122998942
rplB	Thermotogae	-0.218356573	0.49536676
rplB	Verrucomicrobia	-0.34965035	0.266004309
rplB	Viridiplantae	0.480384461	0.113937412

**Appendix A Table 12. ResFams HMMs and antibiotic classifications.**

<b>ResfamID</b>	<b>Gene</b>	<b>Description</b>	<b>Classification</b>
RF0001	16S_rRNA_methyltrans	16S ribosomal RNA methyltransferase	rRNA Methyltransferase
RF0002	AAC3	Aminoglycoside Acetyltransferase (AAC3)	Aminoglycoside
RF0003	AAC3-I	Aminoglycoside Acetyltransferase (AAC3-I)	Aminoglycoside
RF0004	AAC6-I	Aminoglycoside Acetyltransferase (AAC6-I)	Aminoglycoside
RF0005	AAC6-Ib	Aminoglycoside Acetyltransferase (AAC6-Ib)	Aminoglycoside
RF0006	AAC6-II	Aminoglycoside Acetyltransferase (AAC6-II)	Aminoglycoside
RF0007	ABC_efflux	ATP-binding cassette (ABC) antibiotic efflux pump	ABC Transporter
RF0008	ABC_tran	PF00005.22 ABC transporter	ABC Transporter
RF0009	ABC1	PF03109.11 ABC1 family	ABC Transporter
RF0010	ABC2_membrane	PF01061.19 ABC-2 type transporter	ABC Transporter
RF0011	Acetyltransf_1	PF00583.19 Acetyltransferase (GNAT) family	Aminoglycoside
RF0012	Acetyltransf_3	PF13302.1 Acetyltransferase (GNAT) domain	Aminoglycoside
RF0013	Acetyltransf_4	PF13420.1 Acetyltransferase (GNAT) domain	Aminoglycoside
RF0014	Acetyltransf_7	PF13508.1 Acetyltransferase (GNAT) domain	Aminoglycoside
RF0015	Acetyltransf_8	PF13523.1 Acetyltransferase (GNAT) domain	Aminoglycoside
RF0016	Acetyltransf_9	PF13527.1 Acetyltransferase (GNAT) domain	Aminoglycoside
RF0017	ACR_tran	PF00873.14 AcrB/AcrD/AcrF family	Other Efflux
RF0018	Acyltransferase	PF01553.16 Acyltransferase	Aminoglycoside RND Antibiotic Efflux
RF0019	adeA-adeI	adeA-adeI: membrane fusion protein of multidrug efflux complex	RND Antibiotic Efflux
RF0020	adeB	adeB: membrane fusion protein of multidrug efflux complex	RND Antibiotic Efflux
RF0021	adeC-adeK-oprM	adeC-adeK-oprM: outer membrane factor the multidrug efflux complex	RND Antibiotic Efflux
RF0022	adeR	adeR: positive regulator of AdeABC efflux system	Other Efflux
RF0023	adeS	adeS: gene modulating antibiotic efflux regulating AdeABC	Other Efflux
RF0024	Aminotran_1_2	PF00155.16 Aminotransferase class I and II	Other
RF0025	Aminotran_4	PF01063.14 Aminotransferase class IV	Other
RF0026	ANT2	Aminoglycoside Nucleotidyltransferase (ANT2)	Aminoglycoside
RF0027	ANT3	Aminoglycoside Nucleotidyltransferase (ANT3)	Aminoglycoside
RF0028	ANT4	Aminoglycoside Nucleotidyltransferase (ANT4)	Aminoglycoside
RF0029	ANT6	Aminoglycoside Nucleotidyltransferase (ANT6)	Aminoglycoside
RF0030	ANT9	Aminoglycoside Nucleotidyltransferase (ANT9)	Aminoglycoside
RF0031	Antibiotic_NAT	PF02522.9 Aminoglycoside 3-N-acetyltransferase	Aminoglycoside
RF0032	APH	PF01636.18 Phosphotransferase enzyme family	Aminoglycoside
RF0033	APH3	aminoglycoside phosphotransferase (APH3)	Aminoglycoside
RF0034	APH6	aminoglycoside phosphotransferase (APH3)	Aminoglycoside Gene Modulating Resistance
RF0035	baeR	baeR: subunit of gene modulating antibiotic efflux	Gene Modulating Resistance
RF0036	baeS	baeS: subunit of gene modulating antibiotic efflux Bacillus cereus beta-lactamase II (subclass B1 (metallo-) beta-lactamase)	Beta-Lactam
RF0037	BCII		Beta-Lactam
RF0038	Beta-lactamase	PF00144.19 Beta-lactamase	Beta-Lactam
RF0039	Beta-lactamase2	PF13354.1 Beta-lactamase enzyme family	Beta-Lactam

## Appendix A Table 12 (cont'd)

RF0040	BJP	BJP beta-lactamase (subclass B3 (metallo-) beta-lactamase)	Beta-Lactam
RF0041	BlaB	Beta-lactamase B (BlaB) (subclass B1 (metallo-) beta-lactamase)	Beta-Lactam
RF0042	blaI	blaI: gene modulating beta-lactam resistance, regulates PC1 beta-lactamase (blaZ)	Gene Modulating Resistance
RF0043	blaR1	blaR1: gene modulating beta-lactam resistance, regulates PC1 beta-lactamase (blaZ)	Gene Modulating Resistance
RF0044	CARB-PSE	CARB-PSE beta-lactamases (class a)	Beta-Lactam
RF0045	CAT	PF00302.13 Chloramphenicol acetyltransferase	Aminoglycoside
RF0046	CblA	CblA cephalosporin (class a)	Beta-Lactam
RF0047	CepA	CepA cephalosporin (class a)	Beta-Lactam rRNA
RF0048	Cfr23_rRNA_methyltrans	Cfr 23S ribosomal RNA methyltransferase	Methyltransferase
RF0049	CfxA	CfxA cephalosporin (class a)	Beta-Lactam
RF0050	Chlor_Acetyltrans_CAT	chloramphenicol acetyltransferase (CAT)	Chloramphenicol
RF0051	Chlor_Efflux_Pump	chloramphenicol efflux pump	Chloramphenicol
RF0052	Chlor_Phospho_CPT	chloramphenicol phosphotransferase (CPT)	Chloramphenicol
RF0053	ClassA	Class A beta-lactamase	Beta-Lactam
RF0054	ClassB	Class B beta-lactamase	Beta-Lactam
RF0055	ClassC-AmpC	Class C beta-lactamases	Beta-Lactam
RF0056	ClassD	Class D beta-lactamases	Beta-Lactam
RF0057	CMY-LAT-MOX-ACT-MIR-FOX	A grouping of the related CMY, LAT, MOX, ACT, MIR, and FOX beta-lactamases (class c)	Beta-Lactam
RF0058	CPT	PF07931.7 Chloramphenicol phosphotransferase-like protein	Chloramphenicol
RF0059	CTXM	CTX-M beta-lactamase (class a)	Beta-Lactam
RF0060	Dala_Dala_lig_C	PF07478.8 D-ala D-ala ligase C-terminus	Glycopeptide
	D_ala_D_ala		Glycopeptide
RF0061	Dala_Dala_lig_N	PF01820.16 D-ala D-ala ligase N-terminus	Glycopeptide
RF0062	DHA	DHA beta-lactamase (class c)	Beta-Lactam
RF0063	DHFR_1	PF00186.14 Dihydrofolate reductase	Trimethoprim
RF0064	DIM-GIM-SIM	A grouping of the related DIM, GIM, and SIM beta-lactamases (subclass B1 (metallo-) beta-lactamases)	Beta-Lactam
	emrB		MFS Transporter
RF0065	efflux_EmrB	emrB: subunit of efflux pump conferring antibiotic resistance	MFS Transporter
RF0066	emrE	emrE: small multidrug resistance (SMR) antibiotic efflux pump	Other Efflux
RF0067	Erm23S_rRNA_methyltrans	Emr 23S ribosomal RNA methyltransferase: rRNA methyltransferase conferring antibiotic resistance	rRNA Methyltransferase
RF0068	Erm38	Erm38: Erm 23S ribosomal RNA methyltransferase: rRNA methyltransferase conferring antibiotic resistance	rRNA Methyltransferase
RF0069	ErmA	ErmA: Erm 23S ribosomal RNA methyltransferase: rRNA methyltransferase conferring antibiotic resistance	rRNA Methyltransferase
RF0070	ErmB	ErmB: Erm 23S ribosomal RNA methyltransferase: rRNA methyltransferase conferring antibiotic resistance	rRNA Methyltransferase
RF0071	ErmC	ErmC: Erm 23S ribosomal RNA methyltransferase: rRNA methyltransferase conferring antibiotic resistance	rRNA Methyltransferase
RF0072	Exo	Exo beta-lactamase (class a)	Beta-Lactam
RF0073	FAD_binding_2	PF00890.19 FAD binding domain	Other
RF0074	Fluor_Res_DNA_Topo	Fluoroquinolone Resistant DNA Topoisomerase	Quinolone rRNA
RF0075	FmrO	PF07091.6 Ribosomal RNA methyltransferase (FmrO)	Methyltransferase
RF0076	GES	GES beta-lactamase (class a)	Beta-Lactam
RF0077	Glyoxalase	PF00903.20 Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily	Other
RF0078	GOB	GOB beta-lactamase (subclass B3 (metallo-) beta-lactamase)	Beta-Lactam

## Appendix A Table 12 (cont'd)

RF0079	HTH_AraC	PF00165.18 Bacterial regulatory helix-turn-helix proteins, AraC family	Gene Modulating Resistance
RF0080	IMP	Plasmid mediated IMP-type carbapenemases (subclass B1 (metallo-) beta-lactamase)	Beta-Lactam
RF0081	IND	IND beta-lactamases (subclass B1 (metallo-) beta-lactamase)	Beta-Lactam
RF0082	KHM	KHM beta-lactamases (subclass B1 (metallo-) beta-lactamase)	Beta-Lactam
RF0083	KPC	Klebsiella pneumoniae carbapenem resistant (KPC) beta-lactamases (class a)	Beta-Lactam
RF0084	L1	L1 beta-lactamase (subclass B3 (metallo-) beta-lactamase)	Beta-Lactam
RF0085	Lactamase_B	PF00753.22 Metallo-beta-lactamase superfamily	Beta-Lactam
RF0086	Lactamase_B_2	PF12706.2 Beta-lactamase superfamily domain	Beta-Lactam
RF0087	LRA	LRA beta-lactamase (subclass B3 (metallo-) beta-lactamase)	Beta-Lactam
RF0088	macA	macA: subunit of efflux pump conferring antibiotic resistance	Macrolide
RF0089	macB	macB: subunit of efflux pump conferring antibiotic resistance	Macrolide
	macrolide_glycosyl		Macrolide
RF0090	MacrolideGlycosyltransfer	macrolide glycosyltransferase: macrolide inactivation enzyme	Macrolide
RF0091	marA	marA: transcription factor induces MDR efflux pump AcrAB	Gene Modulating Resistance
RF0092	MarR	PF01047.17 MarR family	Gene Modulating Resistance
RF0093	MarR_2	PF12802.2 MarR family	Gene Modulating Resistance
RF0094	mecR1	mecR1: gene modulating beta-lactam resistance	Gene Modulating Resistance
RF0095	Methyltransf_18	PF12847.2 Methyltransferase domain	Other
RF0096	MexA	mexA: membrane fusion protein of the MexAB-OprM multidrug efflux complex	RND Antibiotic Efflux
RF0097	MexC	mexC: membrane fusion protein of the MexCD-OprJ multidrug efflux complex	RND Antibiotic Efflux
RF0098	MexE	mexE: membrane fusion protein of the MexEF-OprN multidrug efflux complex	RND Antibiotic Efflux
RF0099	MexH	mexH: membrane fusion protein of the efflux complex MexGHI-OpmD	RND Antibiotic Efflux
RF0100	MexW-MexI	A grouping of related mexW and mexI subunits of efflux pumps conferring antibiotic resistance	RND Antibiotic Efflux
RF0101	MexX	mexX: subunit of efflux pump conferring antibiotic resistance	RND Antibiotic Efflux
RF0102	MFS_1	PF07690.11 Major Facilitator Superfamily	MFS Transporter
RF0103	MFS_3	PF05977.8 Transmembrane secretion effector	MFS Transporter
RF0104	MFS_efflux	major facilitator superfamily (MFS) antibiotic efflux pump	MFS Transporter
RF0105	MoxA	MoxA beta-lactamase (class a)	Beta-Lactam
RF0106	mprF	mprF: peptide antibiotic resistance gene	Gene Modulating Resistance
RF0107	msbA	msbA: ATP-binding cassette (ABC) antibiotic efflux pump	ABC Transporter
RF0108	NDM-CcrA	A grouping of related NDM and CcrA beta-lactamases	Beta-Lactam
RF0109	norA	norA: major facilitator superfamily (MFS) antibiotic efflux pump	MFS Transporter
RF0110	Nuc_H_symport	PF03825.11 Nucleoside H <sup>+</sup> symporter	Other
RF0111	PC1	PC1: blaZ beta-lactamase (class a)	Beta-Lactam
RF0112	phoQ	phoQ: subunit of gene modulating antibiotic efflux	Gene Modulating Resistance
RF0113	Qnr	quinolone resistance protein (Qnr): antibiotic target protection protein	Quinolone
RF0114	ramA	ramA: gene modulating antibiotic efflux	Gene Modulating Resistance
RF0115	RND_efflux	resistance-nodulation-cell division (RND) antibiotic efflux pump	RND Antibiotic Efflux
RF0116	robA	robA: transcriptional activator of AcrAB antibiotic efflux pump	Gene Modulating Resistance

## Appendix A Table 12 (cont'd)

RF0117	romA	romA: transcription factor mediating antibiotic resistance	Gene Modulating Resistance
RF0118	Sfh	sfh beta-lactamases (subclass B2 (metallo-) beta-lactamase)	Beta-Lactam
RF0119	SHV-LEN	A grouping of the related SHV and LEN beta-lactamases (class a)	Beta-Lactam
RF0120	SME	SME beta-lactamase (class a)	Beta-Lactam
RF0121	soxR	soxR: mutant efflux regulatory protein conferring antibiotic resistance	Gene Modulating Resistance
RF0122	SPM	Sao Paulo metallo-beta-lactamase (SPM-1) (subclass B1 (metallo-) beta-lactamase)	Beta-Lactam
RF0123	SubclassB1	Subclass B1 (metallo-) beta-lactamase hydrolyze penicillins, cephalosporins and carbapenems	Beta-Lactam
RF0124	SubclassB2	Subclass B2 (metallo-) beta-lactamase selectively hydrolyze carbapenems	Beta-Lactam
RF0125	SubclassB3	Subclass B3 (metallo-) beta-lactamase hydrolyze penicillins, cephalosporins and carbapenems	Beta-Lactam
RF0126	TEM	TEM beta-lactamase (class a)	Beta-Lactam
RF0127	TetA	tetA: tetracycline resistance MFS efflux pump	Tetracycline
RF0128	TetA-B	tetA(B): tetracycline resistance MFS efflux pump	Tetracycline
RF0129	TetA-G	tetA(G): tetracycline resistance MFS efflux pump	Tetracycline
RF0130	TetD	tetD: tetracycline resistance MFS efflux pump	Tetracycline
RF0131	TetE	tetE: tetracycline resistance MFS efflux pump	Tetracycline
RF0132	TetH-TetJ	tetH and TetJ: tetracycline resistance MFS efflux pumps	Tetracycline
RF0133	TetM-TetW-TetO-TetS	tetM, tetW, tetO, and tetS: tetracycline resistance ribosomal protection protein	Tetracycline
RF0134	tet_MFS_efflux	tetracycline resistance MFS efflux pump: selectively pump out tetracycline or tetracycline derivatives	Tetracycline
RF0135	tet_ribosomal_protect	tetracycline resistance ribosomal protection protein: protect RNA-polymerase from tetracycline inhibition	Tetracycline
RF0136	TetX	tetX: tetracycline inactivation enzyme	Tetracycline
RF0137	TetY	tetY: tetracycline resistance MFS efflux pump	Tetracycline
RF0138	Tex_N	PF09371.5 Tex-like protein N-terminal domain	Gene Modulating Resistance
RF0139	thym_sym	PF00303.14 Thymidylate synthase	Other
RF0140	efflux_Bcr_CfIA	TIGR00710 efflux Bcr CfIA: drug resistance transporter, Bcr/CfIA subfamily	MFS Transporter
RF0141	EmrB	TIGR00711 efflux EmrB: drug resistance MFS transporter, drug:H <sup>+</sup> antiporter-2 (14 Spanner) (DHA2) family	MFS Transporter
RF0142	MATE_efflux	TIGR00797 matE: MATE efflux family protein	Other Efflux
RF0143	drdA	TIGR01188 drdA: daunorubicin resistance ABC transporter, ATP-binding protein	ABC Transporter
RF0144	DalaDala	TIGR01205 D ala D alaTIGR: D-alanine--D-alanine ligase	Glycopeptide Gene Modulating Resistance
RF0145	SoxR	TIGR01950 SoxR: redox-sensitive transcriptional activator SoxR	Resistance
RF0146	Thymidylat_synt	TIGR03284 thym sym: thymidylate synthase	Other
RF0147	tolC	tolC: subunit of efflux pump conferring antibiotic resistance	ABC Transporter
RF0148	Transpeptidase	PF00905.17 Penicillin binding protein transpeptidase domain	Beta-Lactam
RF0149	vanA	VanA: D-Ala-D-Ala ligase that can synthesize D-Ala-D-Lac	Glycopeptide
RF0150	vanB	VanB: D-Ala-D-Ala ligase that can synthesize D-Ala-D-Lac	Glycopeptide
RF0151	vanC	VanC: D-Ala-D-Ala ligase that can synthesize D-Ala-D-Ser	Glycopeptide
RF0152	vanD	VanD: D-Ala-D-Ala ligase that can synthesize D-Ala-D-Lac	Glycopeptide
RF0153	vanH	VanH: D-specific alpha-ketoacid dehydrogenase that synthesizes D-lactate	Glycopeptide
RF0154	vanR	VanR: transcriptional activator regulating VanA, VanH and VanX	Glycopeptide
RF0155	vanS	VanS: transcriptional regulator of van glycopeptide resistance genes	Glycopeptide
RF0156	vanT	VanT: membrane bound serine racemase, converting L-serine to D-serine	Glycopeptide



## Appendix A Table 12 (cont'd)

RF0157	vanW	VanW: glycopeptide resistance gene	Glycopeptide
RF0158	vanX	VanX: glycopeptide resistance gene	Glycopeptide
RF0159	vanY	VanY: glycopeptide resistance gene	Glycopeptide
RF0160	vanZ	VanZ: glycopeptide resistance gene	Glycopeptide
RF0161	VEB-PER	VEB and PER beta-lactamases (class a)	Beta-Lactam
RF0162	VIM	Verone integron-encoded (VIM) metallo-beta-lactamase (subclass B1 (metallo-) beta-lactamase)	Beta-Lactam Gene Modulating
RF0163	Whib	PF02467.11 Transcription factor WhiB	Resistance
RF0164	RND_mfp	TIGR01730: RND_mfp: efflux transporter, RND family, MFP subunit	RND Antibiotic Efflux
RF0165	Dihydropteroate	TIGR01496_DHPS: dihydropteroate synthase	Trimethoprim
RF0166	ANT	Aminoglycoside Nucleotidyltransferase	Aminoglycoside
RF0167	Aminoglyc_resit	PF10706.4_Aminoglycoside-2"-adenylyltransferase	Aminoglycoside
RF0168	TE_Inactivator	Leginoella_TE_Inactivator	Tetracycline
RF0169	Small_Multi_Drug_Res	PF00893.14_Small Multidrug Resistance protein	Other Efflux
RF0171	Usp	PF00582.21_Universal stress protein family	Other
RF0172	APH3"	Streptomycin phosphotransferase	Aminoglycoside
RF0173	APH3'	Broad-spectrum Aminoglycoside Phosphotransferase	Aminoglycoside
RF0174	ArmA_Rmt	16S rRNA methyltransferase providing aminoglycoside resistance	Aminoglycoside

**Appendix A Table 13. Summary of reference arsenic resistance and metabolism protein sequences from FunGene databases.**

<b>FunGene Database/ Protein</b>	<b>Minimum HMM score</b>	<b>Minimum length (aa)</b>	<b>Minimum HMM coverage (%)</b>	<b>Number of FunGene sequences</b>	<b>Number of dereplicated sequences</b>	<b>Minimum assembled length (aa)</b>
<b>ArsB</b>	150	400	80	23680	5250	150
<b>ACR3</b>	140	300	80	19812	8002	150
<b>ArsC_glut</b>	80	120	85	18082	9635	50
<b>ArsC_thio</b>	172	100	80	7180	7180	50
<b>ArrA</b>	175	75	5	1621	1487	150
<b>AioA</b>	800	800	80	382	293	150
<b>ArsM</b>	200	100	30	3446	2948	160
<b>ArsD</b>	80	100	80	5404	876	150
<b>ArxA</b>	600	800	80	67	54	150

**Appendix A Table 14. Available metadata and accession numbers for soil metagenomes used in this study.**

Project Name	Sample Location	Country	Sample Shortname	Accession location	Project ID	Sample Name	Gbp
ARMO	Rondonia	Brazil	Brazilian_forest	MG-RAST	mgp3731	mgm4546395.3	13.27
ARMO	Rondonia	Brazil	Brazilian_forest	MG-RAST	mgp3731	mgm4536139.3	9.04
ARMO	Rondonia	Brazil	Brazilian_forest	MG-RAST	mgp3731	mgm4535554.3	9.69
Axel Heiberg Permafrost: Part 4A	Central Axel Heiberg Island	Canada	Permafrost_Canada	MG-RAST	mgp252	mgm4523023.3	6.52
Axel Heiberg Permafrost: Part 4A	Central Axel Heiberg Island	Canada	Permafrost_Canada	MG-RAST	mgp252	mgm4523145.3	5.52
CedarCreek_minsoil_June2013	Bethel, MN	USA	Minnesota_grassland	MG-RAST	mgp5588	mgm4541646.3	10.65
CedarCreek_minsoil_June2013	Bethel, MN	USA	Minnesota_grassland	MG-RAST	mgp5588	mgm4541645.3	9.77
Fermi-syntheticlongreads	Fermi National Accelerator Laboratory	USA	Illinois_switchgrass	MG-RAST	mgp14596	mgm4653791.3	7.95
GED prairie unassembled	Iowa	USA	Iowa_prairie	MG-RAST	mgp6377	mgm4539575.3	18.79
GED prairie unassembled	Iowa	USA	Iowa_prairie	MG-RAST	mgp6377	mgm4539572.3	17.58
GED prairie unassembled	Iowa	USA	Iowa_prairie	MG-RAST	mgp6377	mgm4539576.3	17.43
GP corn unassembled	Iowa	USA	Iowa_corn	MG-RAST	mgp6368	mgm4539522.3	8.19
GP corn unassembled	Iowa	USA	Iowa_corn	MG-RAST	mgp6368	mgm4539523.3	8.12
Hofmockel Soil Aggregate COB KBASE	Boone County, IA	USA	Iowa_agricultural	MG-RAST	mgp2592	mgm4509400.3	24.98
Hofmockel Soil Aggregate COB KBASE	Boone County, IA	USA	Iowa_agricultural	MG-RAST	mgp2592	mgm4509401.3	7.86
ISA-SMC-2011	Auburn, IL	USA	Illinois_soybean	MG-RAST	mgp2076	mgm4502542.3	12.54
ISA-SMC-2011	Auburn, IL	USA	Illinois_soybean	MG-RAST	mgp2076	mgm4502540.3	10.60
Loma_Ridge_grassland	Loma Ridge, CA	USA	California_grassland	MG-RAST	mgp1992	mgm4511115.3	6.50

## Appendix A Table 14 (cont'd)

<b>Loma_Ridge_grassland</b>	Loma Ridge, CA	USA	California_grassland	MG-RAST	mgp1992	mgm4511062.3	5.77
<b>Mining of new genes and pathways from soil of mangrove forest</b>	Matang Mangrove Forest	Malaysia	Mangrove	MG-RAST	mgp11628	mgm4603402.3	24.38
<b>Mining of new genes and pathways from soil of mangrove forest</b>	Matang Mangrove Forest	Malaysia	Mangrove	MG-RAST	mgp11628	mgm4603270.3	24.54
<b>NEON</b>	Disney Wilderness Preserve, FL	USA	Disney_preserve	MG-RAST	mgp13948	mgm4664918.3	11.20
<b>NEON</b>	Disney Wilderness Preserve, FL	USA	Disney_preserve	MG-RAST	mgp13948	mgm4664925.3	4.14
<b>Permafrost sediments, North-East Siberia, Kolyma lowland</b>	Kolyma river lowland	Russia	Permafrost_Russia	MG-RAST	mgp7176	mgm4546813.3	19.20
<b>Ungulate Exclosure 2015</b>	Wyoming	USA	Wyoming_soil	MG-RAST	mgp15600	mgm4670120.3	6.41
<b>Surface soil microbial communities from Centralia Pennsylvania</b>	Centralia, PA	USA	Centralia_recovered	JGI	Gp0112853	Cen01	23
<b>Surface soil microbial communities from Centralia Pennsylvania</b>	Centralia, PA	USA	Centralia_recovered	JGI	Gp0112853	Cen03	26
<b>Surface soil microbial communities from Centralia Pennsylvania</b>	Centralia, PA	USA	Centralia_recovered	JGI	Gp0112853	Cen04	25
<b>Surface soil microbial communities from Centralia Pennsylvania</b>	Centralia, PA	USA	Centralia_recovered	JGI	Gp0112853	Cen05	25
<b>Surface soil microbial communities from Centralia Pennsylvania</b>	Centralia, PA	USA	Centralia_fire-affected	JGI	Gp0112853	Cen06	22
<b>Surface soil microbial communities from Centralia Pennsylvania</b>	Centralia, PA	USA	Centralia_recovered	JGI	Gp0112853	Cen07	21
<b>Surface soil microbial communities from Centralia Pennsylvania</b>	Centralia, PA	USA	Centralia_fire-affected	JGI	Gp0112853	Cen10	36
<b>Surface soil microbial communities from Centralia Pennsylvania</b>	Centralia, PA	USA	Centralia_fire-affected	JGI	Gp0112853	Cen12	24
<b>Surface soil microbial communities from Centralia Pennsylvania</b>	Centralia, PA	USA	Centralia_fire-affected	JGI	Gp0112853	Cen14	24

## Appendix A Table 14 (cont'd)

Surface soil microbial communities from Centralia Pennsylvania	Centralia, PA	USA	Centralia_fire-affected	JGI	Gp0112853	Cen15	20
Surface soil microbial communities from Centralia Pennsylvania	Centralia, PA	USA	Centralia_fire-affected	JGI	Gp0112853	Cen16	51
Surface soil microbial communities from Centralia Pennsylvania	Centralia, PA	USA	Centralia_reference	JGI	Gp0112853	Cen17	24
Surface soil microbial communities from an active vent of coal mine fire in Centralia Pennsylvania, Sep 20 '18	Centralia, PA	USA	Centralia_fire-affected	NCBI	SRR7882662	Cen13	56

**Appendix A Table 15. Phylum-level summary of arsenic related genes in RefSoil+ chromosomes and plasmids.**

Number of phylum representatives are shown in parentheses, and percentage of RefSoil+ organisms with respective arsenic related genes are shown.

PHYLUM	CHROMOSOME								PLASMID					
	<i>arsB</i>	<i>acr3</i>	<i>arsC</i> (grx)	<i>arsC</i> (trx)	<i>arsM</i>	<i>aioA</i>	<i>arxA</i>	<i>arrA</i>	<i>arsB</i>	<i>acr3</i>	<i>arsC</i> (grx)	<i>arsC</i> (trx)	<i>arsM</i>	<i>aioA</i>
ACIDOBACTERIA (7)	57.1	100	0	14.3	14.3	0	0	0	0	0	0	0	0	0
ACTINOBACTERIA (118)	44.1	61.9	33.1	5.9	1.7	0	0	0	0	4.2	0	0.8	0.8	0
BACTEROIDETES (19)	0	73.7	31.6	36.8	21.1	0	0	0	0	5.3	0	0	0	0
CHLOROFLEXI (9)	88.9	88.9	0	0	100	0	0	0	0	0	0	0	0	0
CYANOBACTERIA (26)	3.8	76.9	11.5	34.6	3.8	0	0	0	0	0	0	3.8	0	0
DEINOCOCCUS-THERMUS (6)	50	0	0	16.7	0	0	0	0	16.7	0	0	16.7	0	0
FIRMICUTES (207)	65.7	44.4	1	45.9	5.8	0	0	1.4	0	3.4	0	1.9	0	0
PROTEOBACTERIA (531)	33.1	40.7	79.5	4	1.9	2.4	0.2	0.9	2.1	4.1	5.6	0	0	0.4
SPIROCHAETES (12)	0	25	0	16.7	0	0	0	0	0	0	0	0	0	0

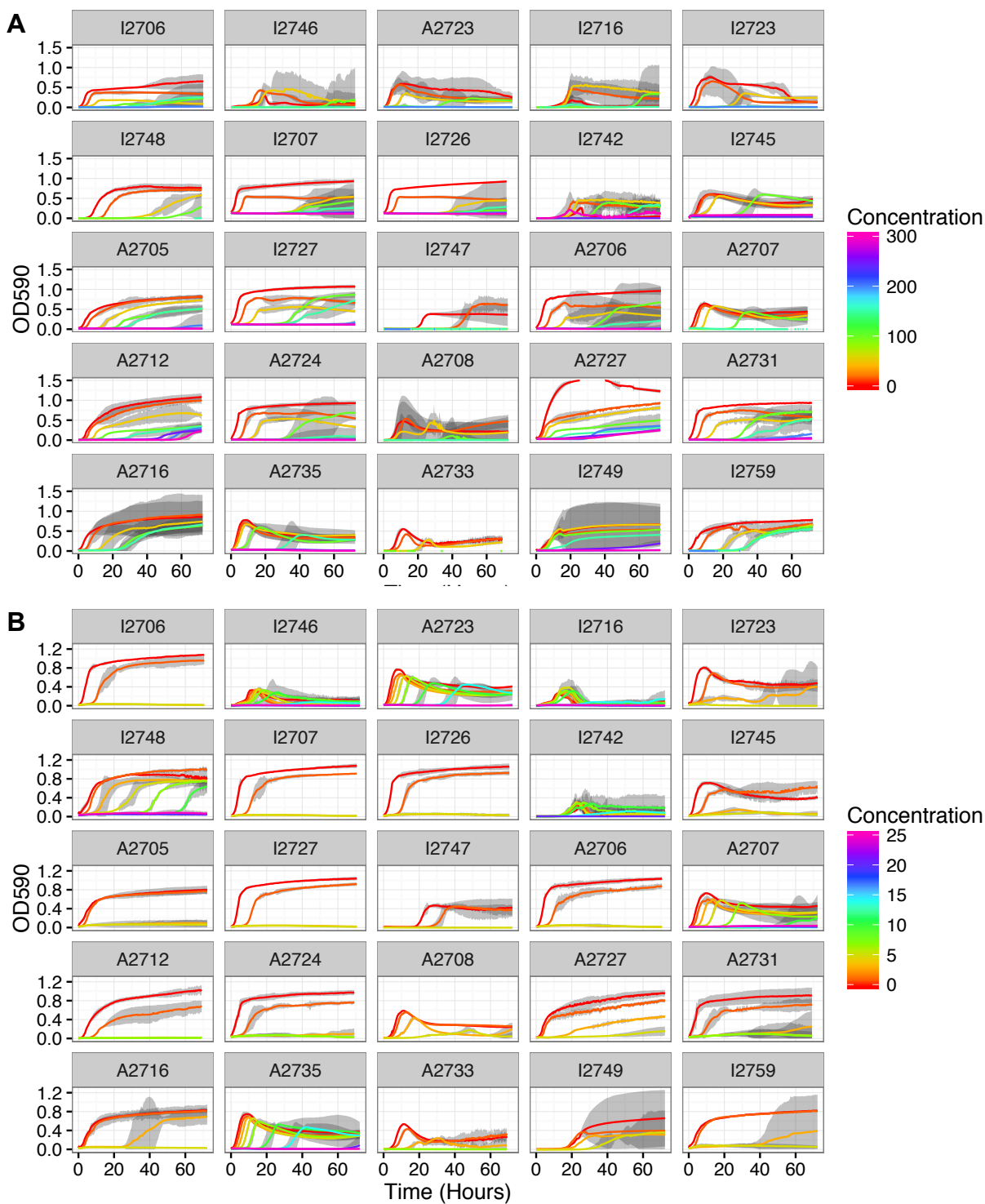
**Appendix A Table 16. Summary of endemic arsenic related gene sequences.**

A sequence was considered endemic if it was present in less than three different soil sites.

<b>Gene</b>	<b>Number of sequences</b>	<b>Number of endemic sequences</b>	<b>Percent endemic</b>
<i>acr3</i>	610	607	99.5
<i>aioA</i>	63	62	98.4
<i>arrA</i>	63	63	100
<i>arsB</i>	8	8	100
<i>arsC</i> (grx)	1316	1299	98.7
<i>arsC</i> (trx)	292	291	99.7
<i>arsD</i>	64	64	100
<i>arsM</i>	1193	1191	99.8
<i>arxA</i>	12	12	100
<b>Totals</b>	3621	3597	99.3

## **APPENDIX B: Supplementary Figures**

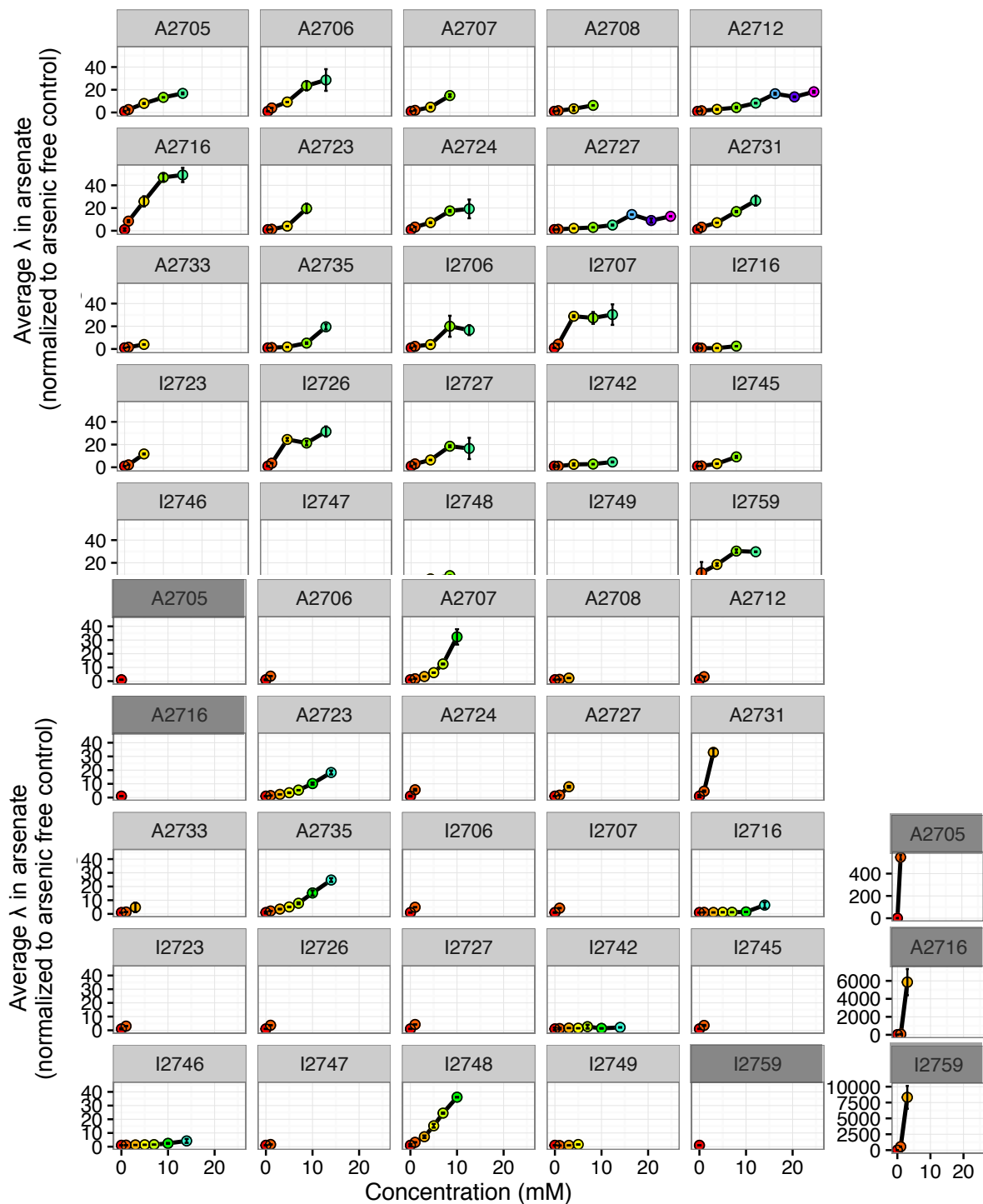




**Appendix B Figure 1. Average OD<sub>590</sub> over 72 h in TSB50 with increasing concentrations of arsenate A) or arsenite B).**

### **Appendix B Figure 1 (cont'd)**

Grey ribbon represents 95% confidence intervals from three replicates. Note the difference in color scales for **A** and **B**.

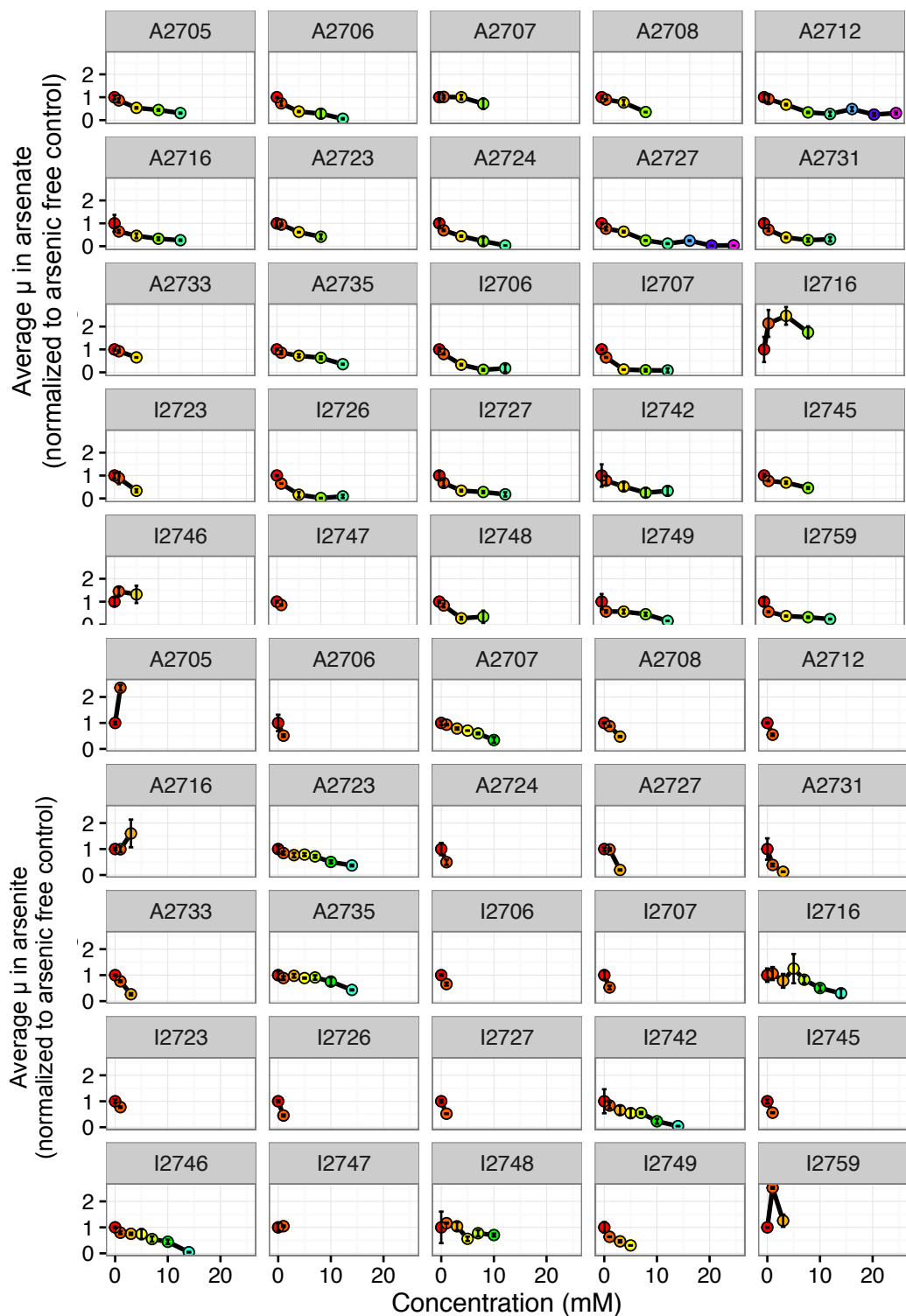


**Appendix B Figure 2. Lag time in TSB50 with increasing concentrations of arsenate and arsenite normalized to growth in TSB50 without arsenic.**

## **Appendix B Figure 2 (cont'd)**

Points are averages from three technical replicates, and error bars show standard deviation.

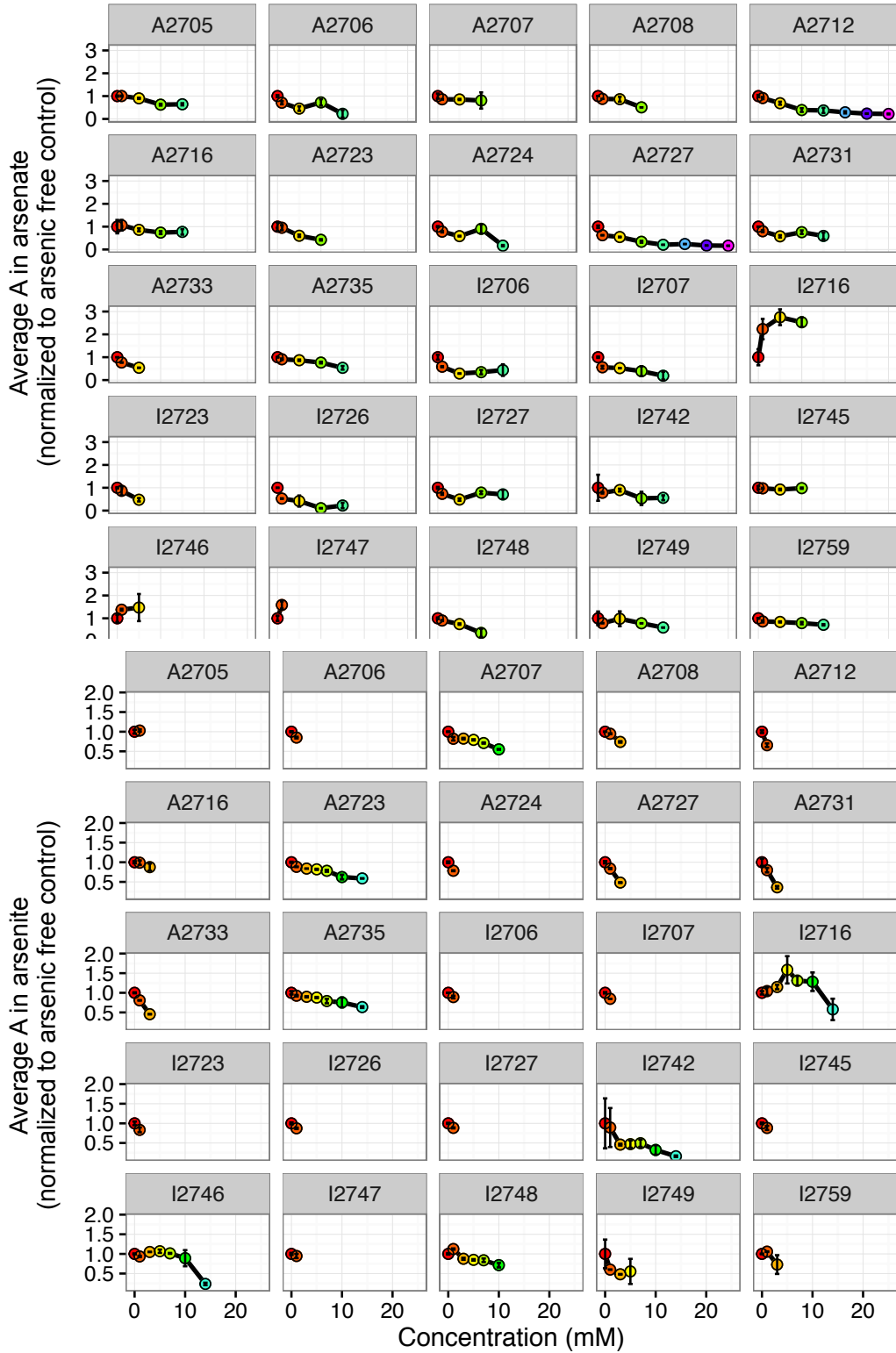
Note the different scale for  $\lambda$  in arsenite for isolates A2705, A2716, and I2759.



**Appendix B Figure 3. Growth rate in TSB50 with increasing concentrations of arsenate and arsenite normalized to growth in TSB50 without arsenic.**

### **Appendix B Figure 3 (cont'd)**

Points are averages from three technical replicates, and error bars show standard deviation.

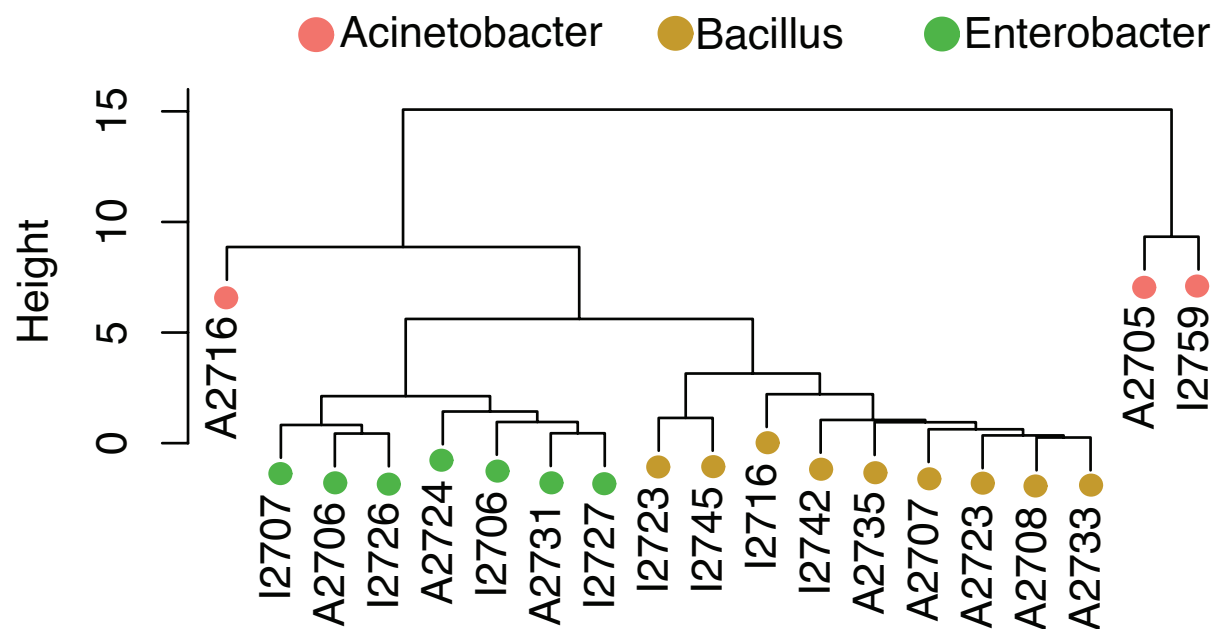


**Appendix B Figure 4. Maximum OD<sub>590</sub> in TSB50 with increasing concentrations of arsenate and arsenite normalized to growth in TSB50 without arsenic.**

#### **Appendix B Figure 4 (cont'd)**

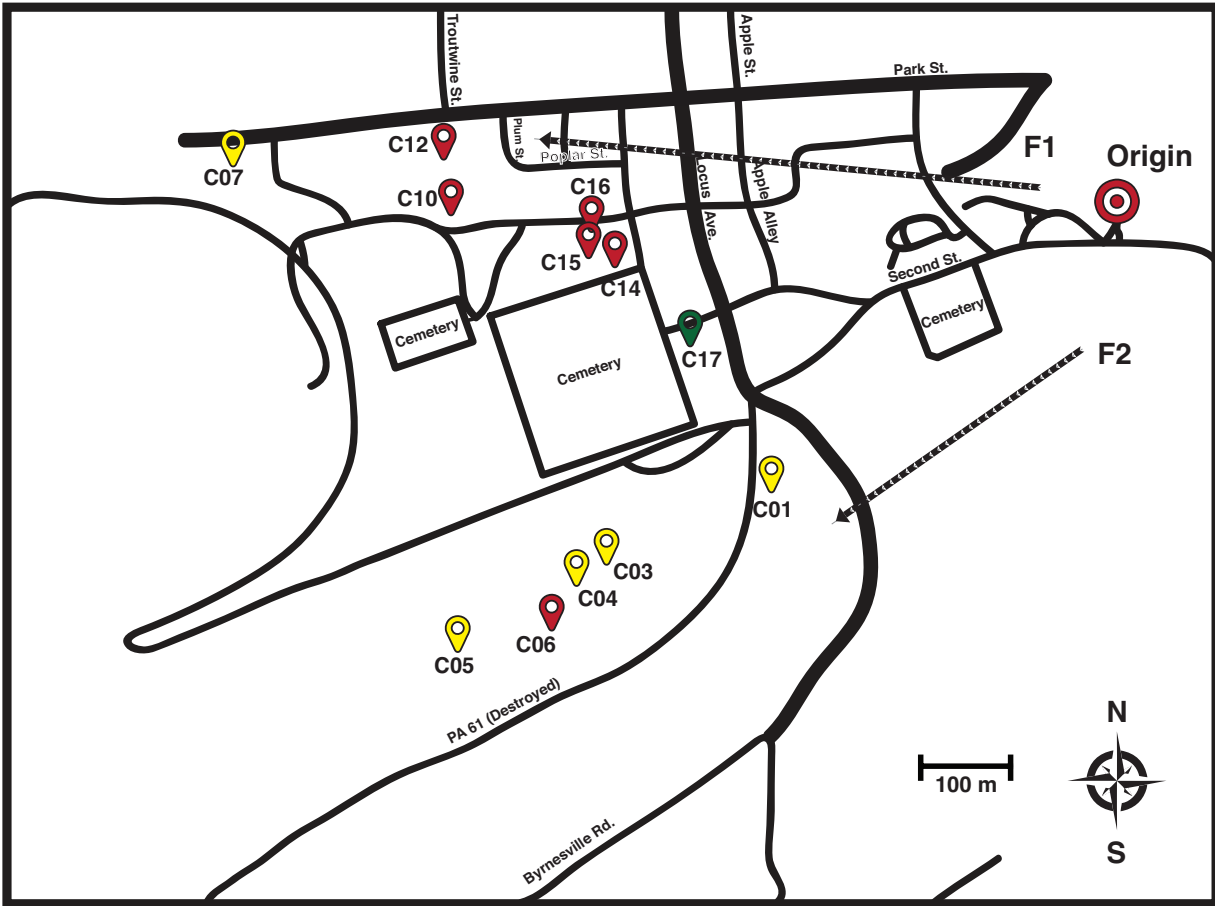
Points are averages from three technical replicates, and error bars show standard deviation.





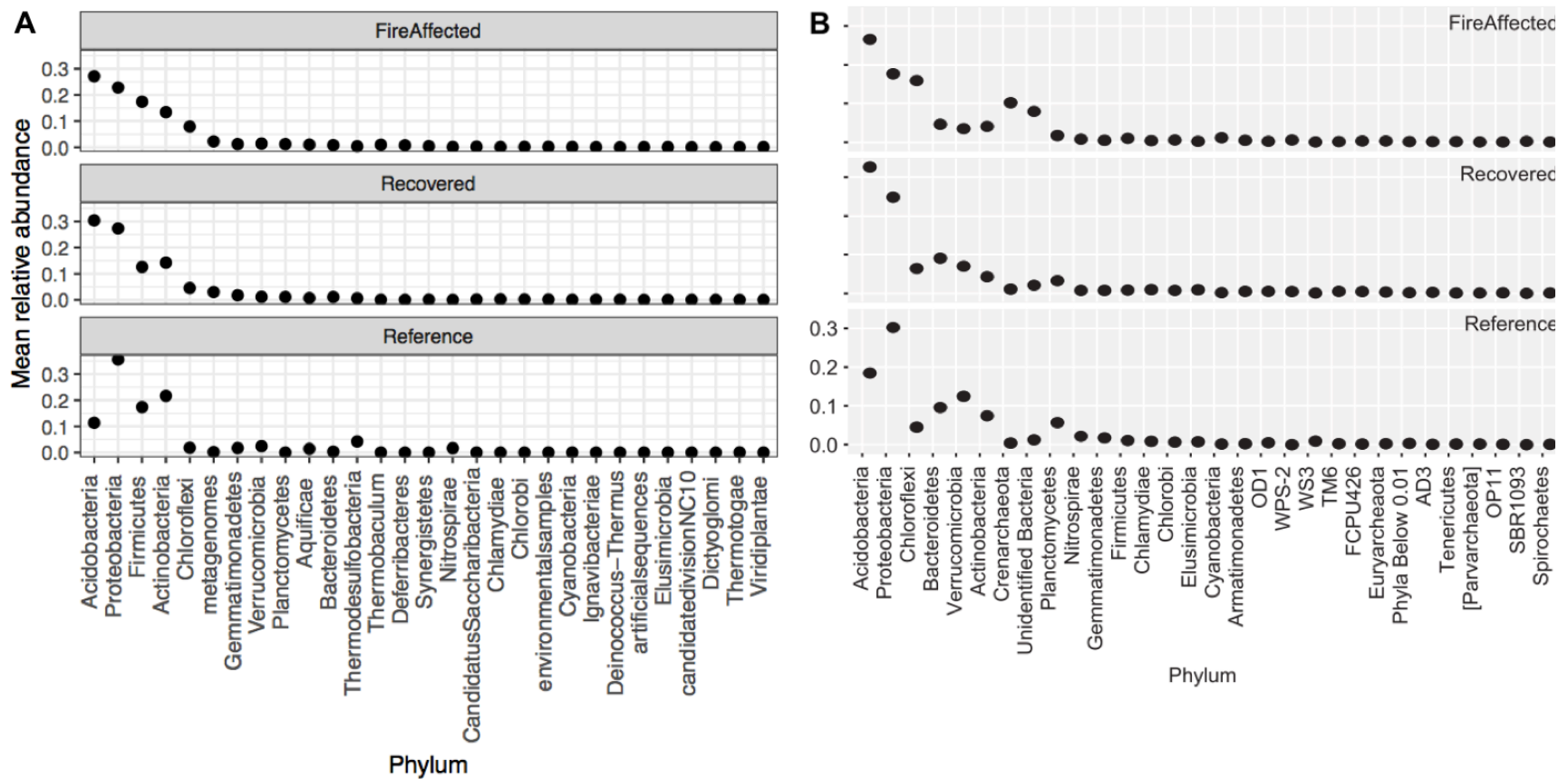
**Appendix B Figure 5. Dendrogram of isolate growth phenotypes in arsenic.**

Only isolates belonging to genera with  $n > 2$  are included. Growth parameters ( $\lambda$ ,  $\mu$ ,  $A$ ) in 1 mM sodium arsenite and 10 mM sodium arsenate were normalized to those with no arsenic controls and used for clustering. Color indicates isolate genus.



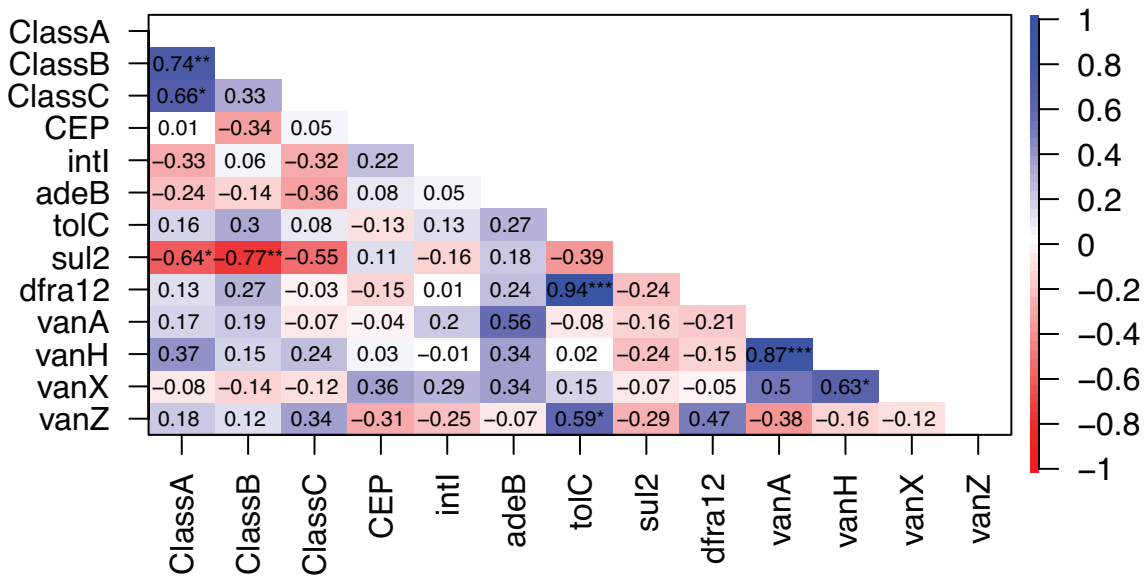
**Appendix B Figure 6. Sampling strategy along the Centralia temperature gradient.**

Twelve surface soils were collected along two fire fronts. Sampling sites are classified based on historical fire activity (Elick 2011) and observations of fire activity at the time of sampling: fire affected (red), recovered (yellow), and reference (green). Red bullseye indicates fire origin, and fire fronts one and two are indicated with arrows F1 and F2, respectively.



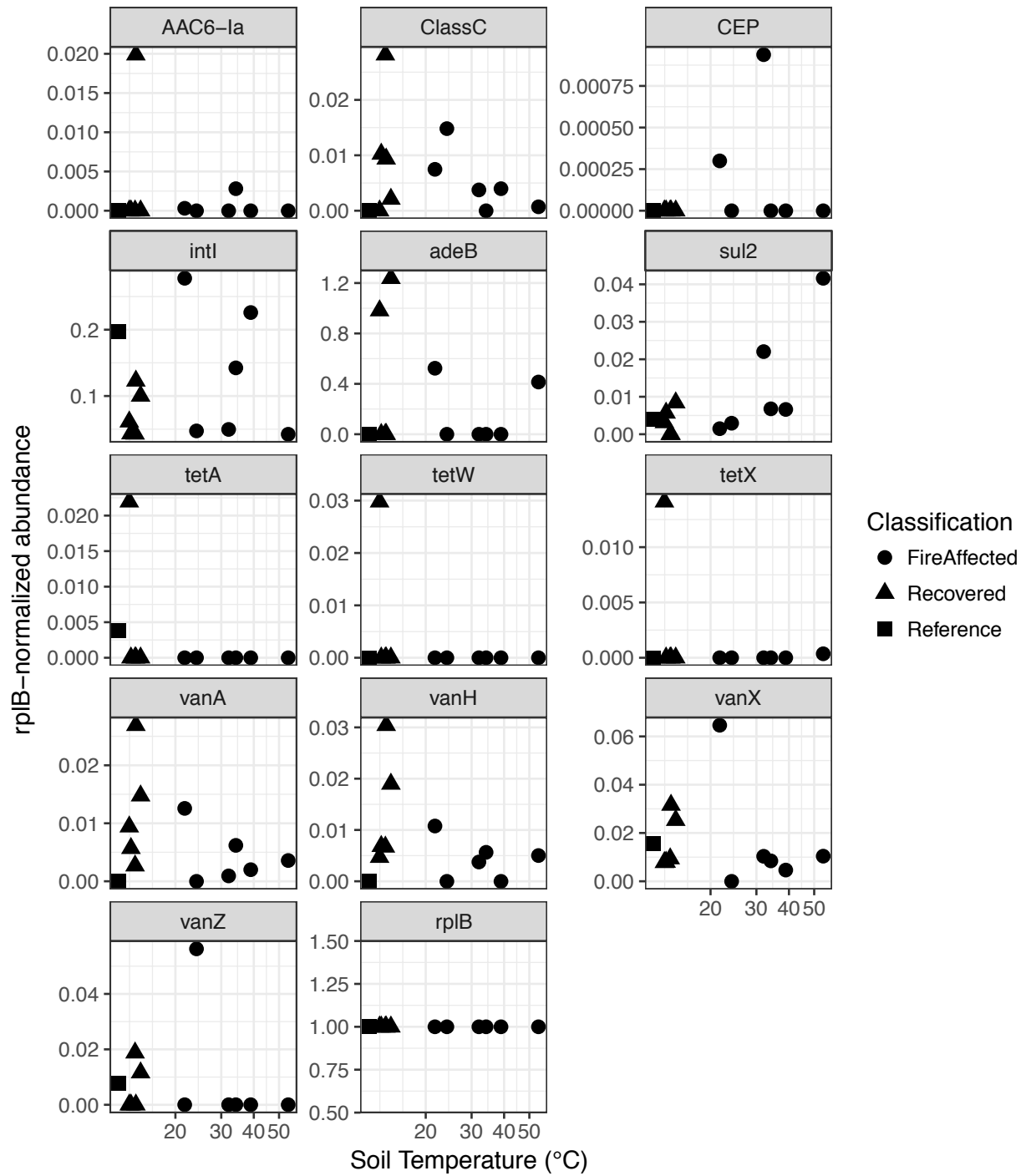
**Appendix B Figure 7. Comparison of community structure assessed using two different methods.**

Community structure determined by *rplB* (**A**) is similar to previously described community structure determined by 16S rRNA gene sequencing reported in Lee and Sorensen et al. 2017 (**B**). Samples are classified by their fire history: fire affected (n = 6), recovered (n = 5), and reference (n = 1). Note the differences in x-axes.



**Appendix B Figure 8. Pair-wise Spearman's correlations of normalized ARG abundances in Centralia.**

Spearman's rho is indicated in each cell and by color, where negative correlations are red and positive correlations are blue. False discovery rate adjusted significance is noted by asterisks.

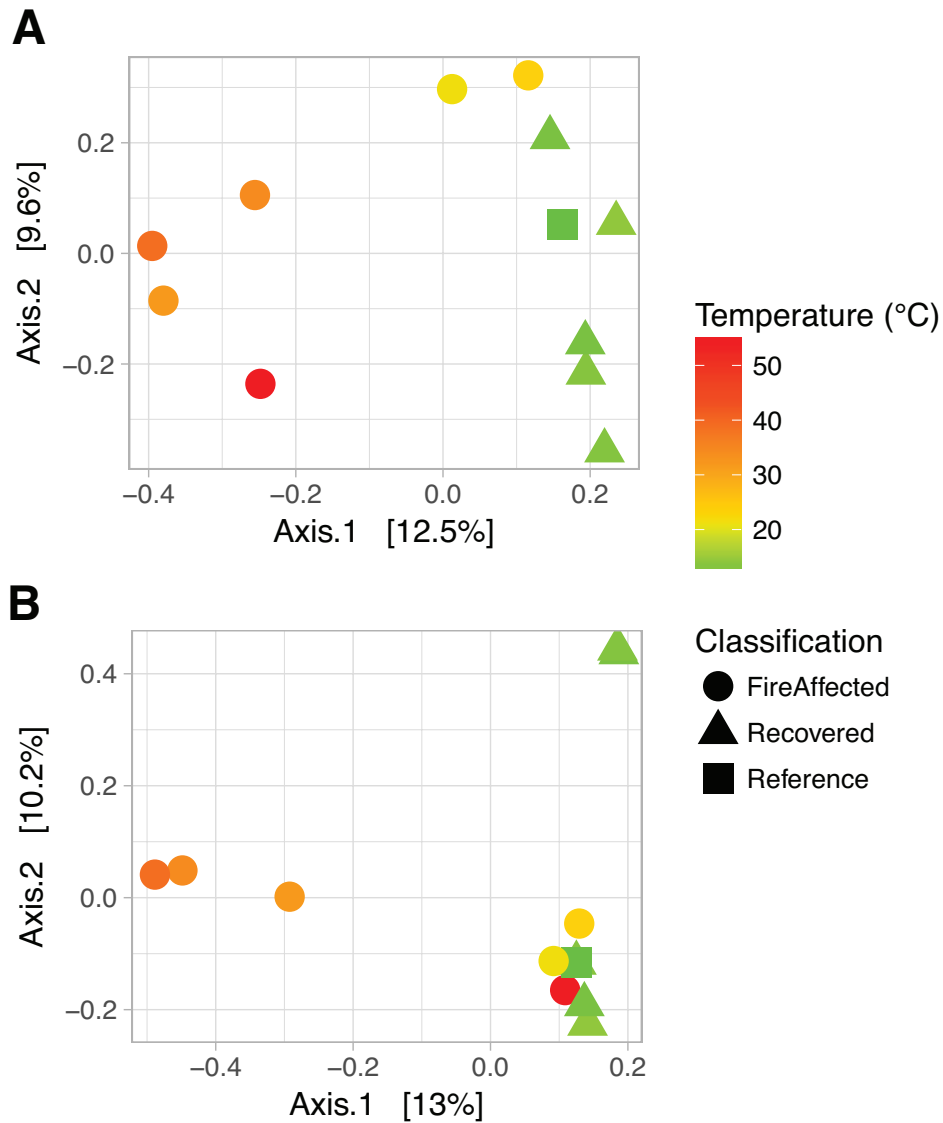


**Appendix B Figure 9. Relationship between normalized abundance of ARGs and soil temperature.**

Point shape indicates soil fire classification. Coverage-adjusted abundance for each gene was

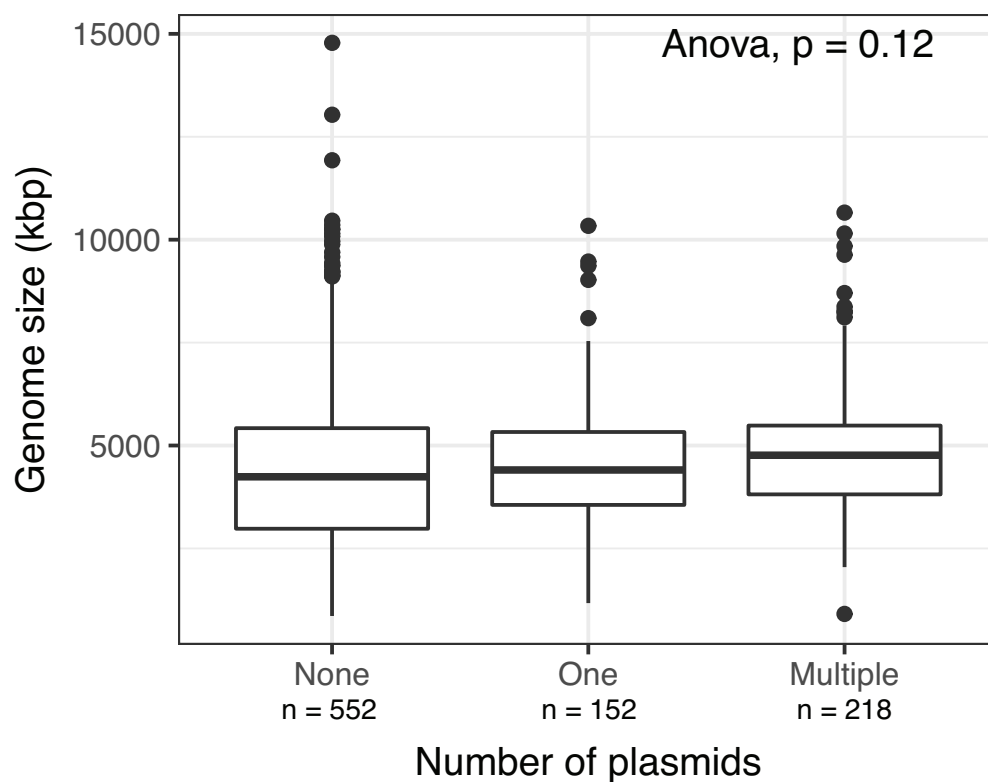
### **Appendix B Figure 9 (cont'd)**

normalized to total abundance of single copy gene *rplB*. Normalized abundance is plotted against soil temperature. Note the differences in y-axes. Shape indicates soil classification based on fire history.



**Appendix B Figure 10. Beta diversity of *Centrulia* microbial communities with *rplB* and ARGs.**

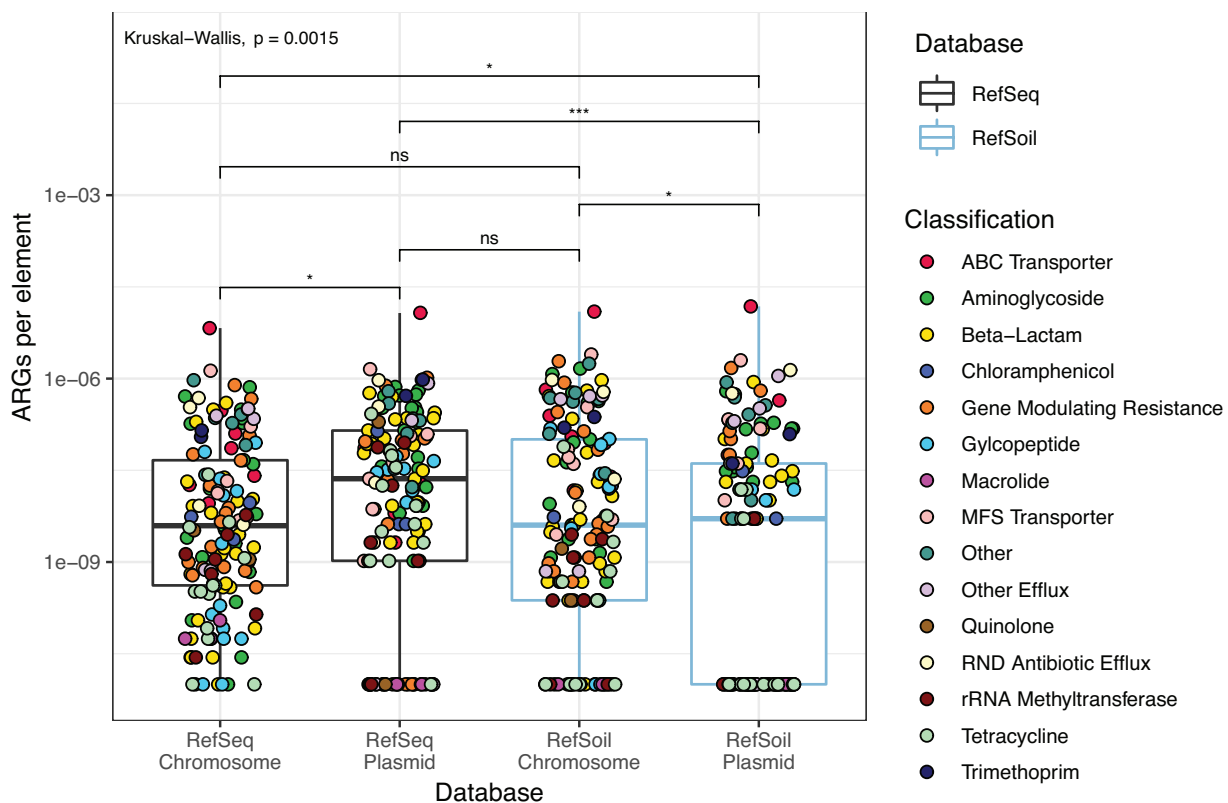
Principal coordinate analysis (PCoA) based on weighted Bray-Curtis distances of community structure (A) and ARG structure (B). Colors represent soil temperature, and shape indicates soil classification based on fire history.



**Appendix B Figure 11. Relationship between plasmid number and chromosome size.**

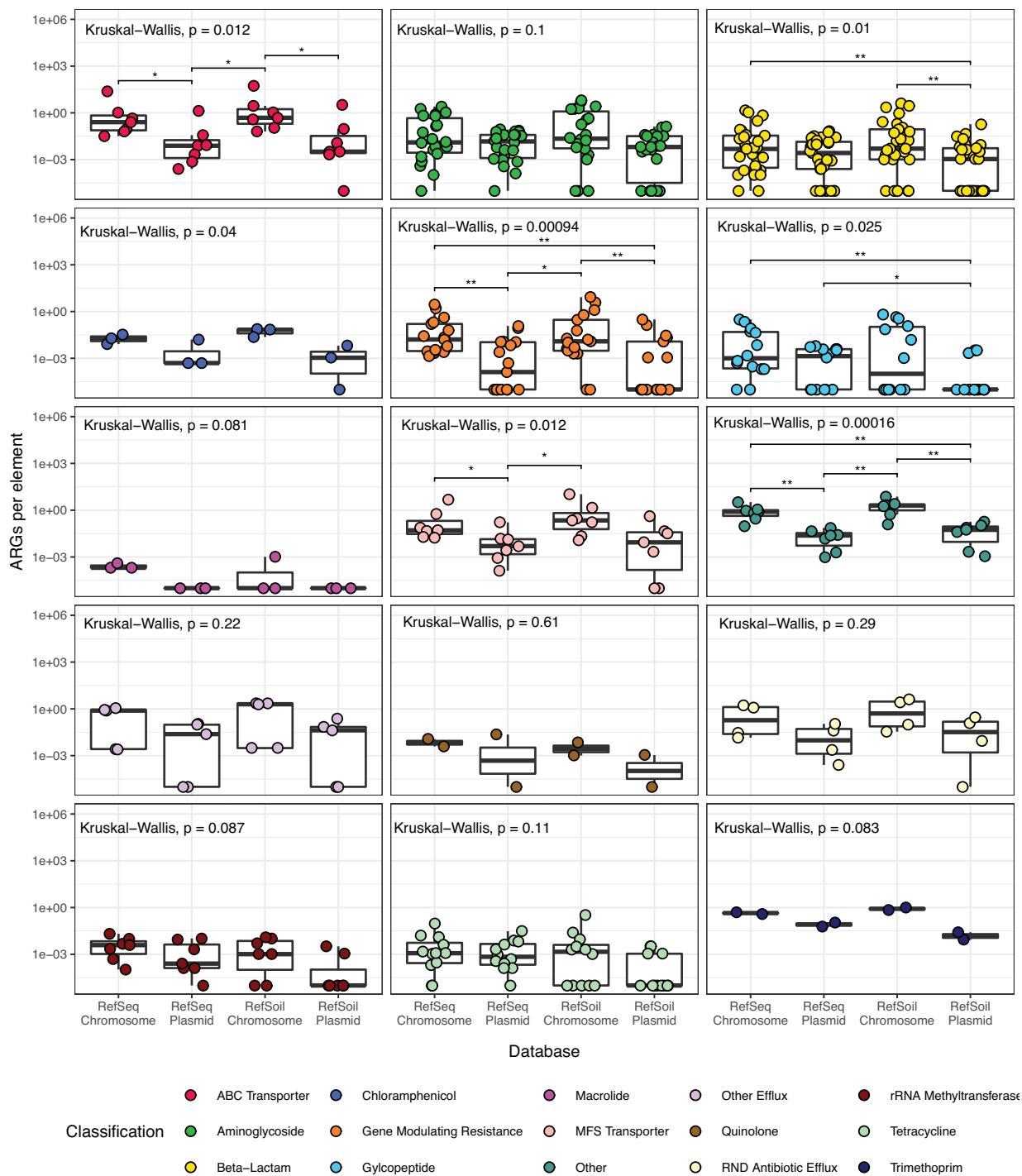
Boxplots showing the distribution of genome sizes based on the number of plasmids. Numbers of microorganisms in that category are shown below each category name. P value from ANOVA is also shown.





**Appendix B Figure 12. Proportion of ARGs on genomes and plasmids in RefSeq+ and RefSeq databases normalized to base pairs.**

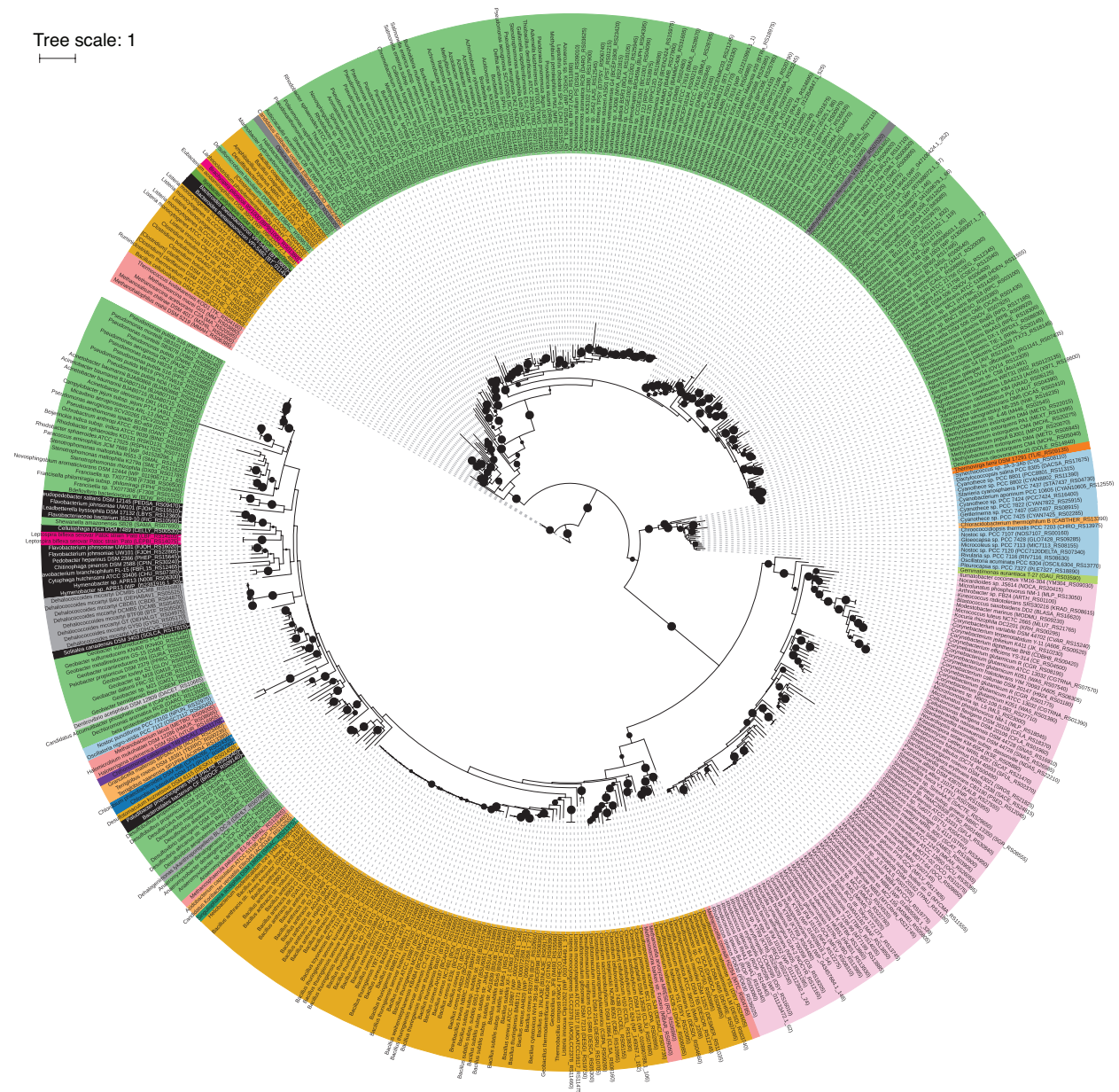
Numbers of ARGs were normalized to total numbers of base pairs. Boxplots are colored by database. Points represent individual ARGs and are colored based on classification. Kruskal-Wallis test results are shown in addition to significant results from pairwise Mann-Whitney U tests.



**Appendix B Figure 13. Proportion of ARGs by classification in RefSeq and RefSoil databases.**

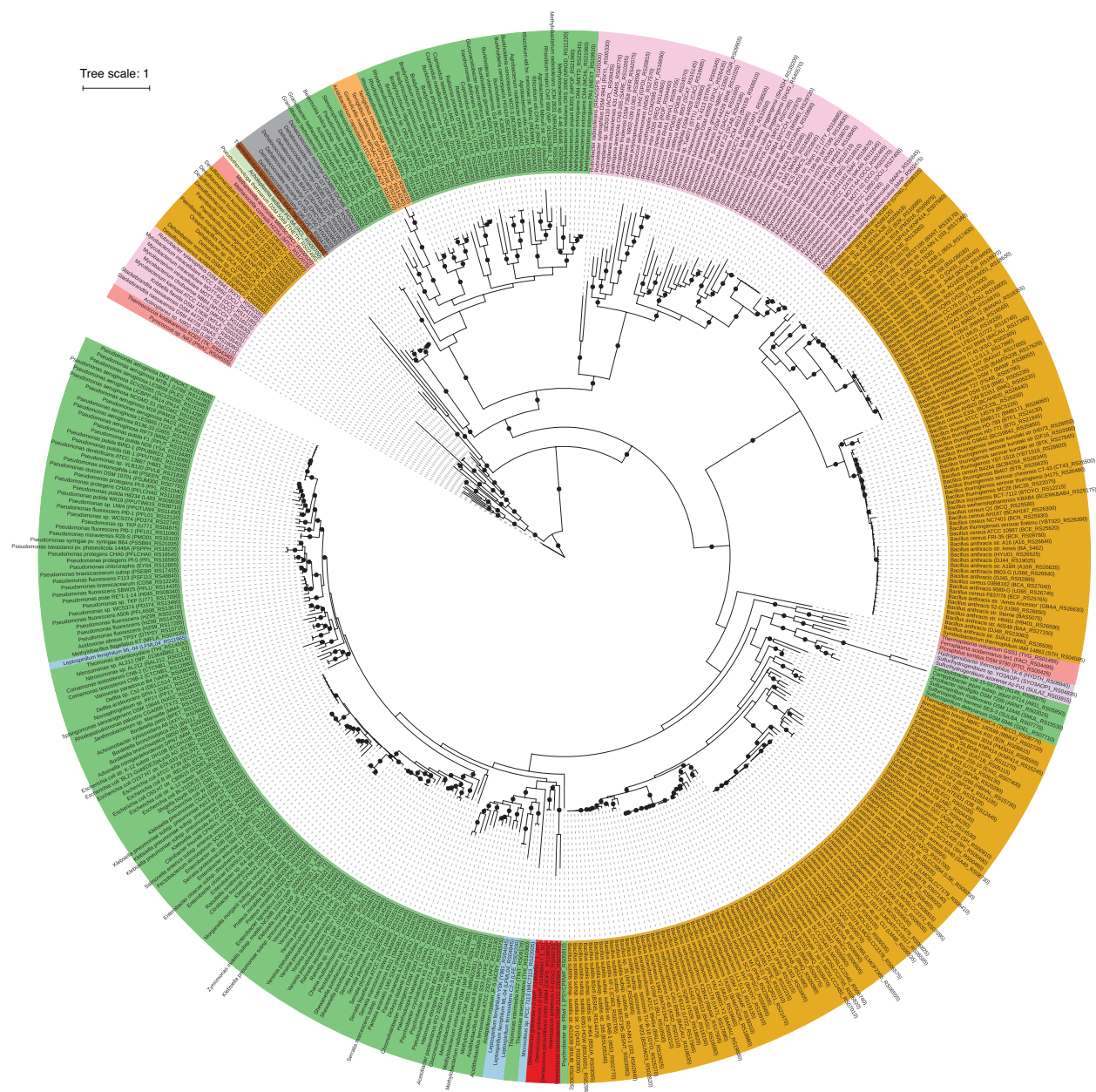
## Appendix B Figure 13 (cont'd)

Boxplots of the proportion of ARGs per genetic element. Each ARG was normalized to the number of genetic elements in the database. Points are colored by ARG category. Kruskal-Wallis P values are shown, and where applicable, significant Mann-Whitney U test results are shown (ns,  $P > 0.05$ ; \*,  $P \leq 0.05$ ; \*\*,  $P \leq 0.01$ ; \*\*\*,  $P \leq 0.001$ ; \*\*\*\*,  $P \leq 0.0001$ ).



**Appendix B Figure 14. Phylogeny of Acr3 in RefSoil+ organisms.**

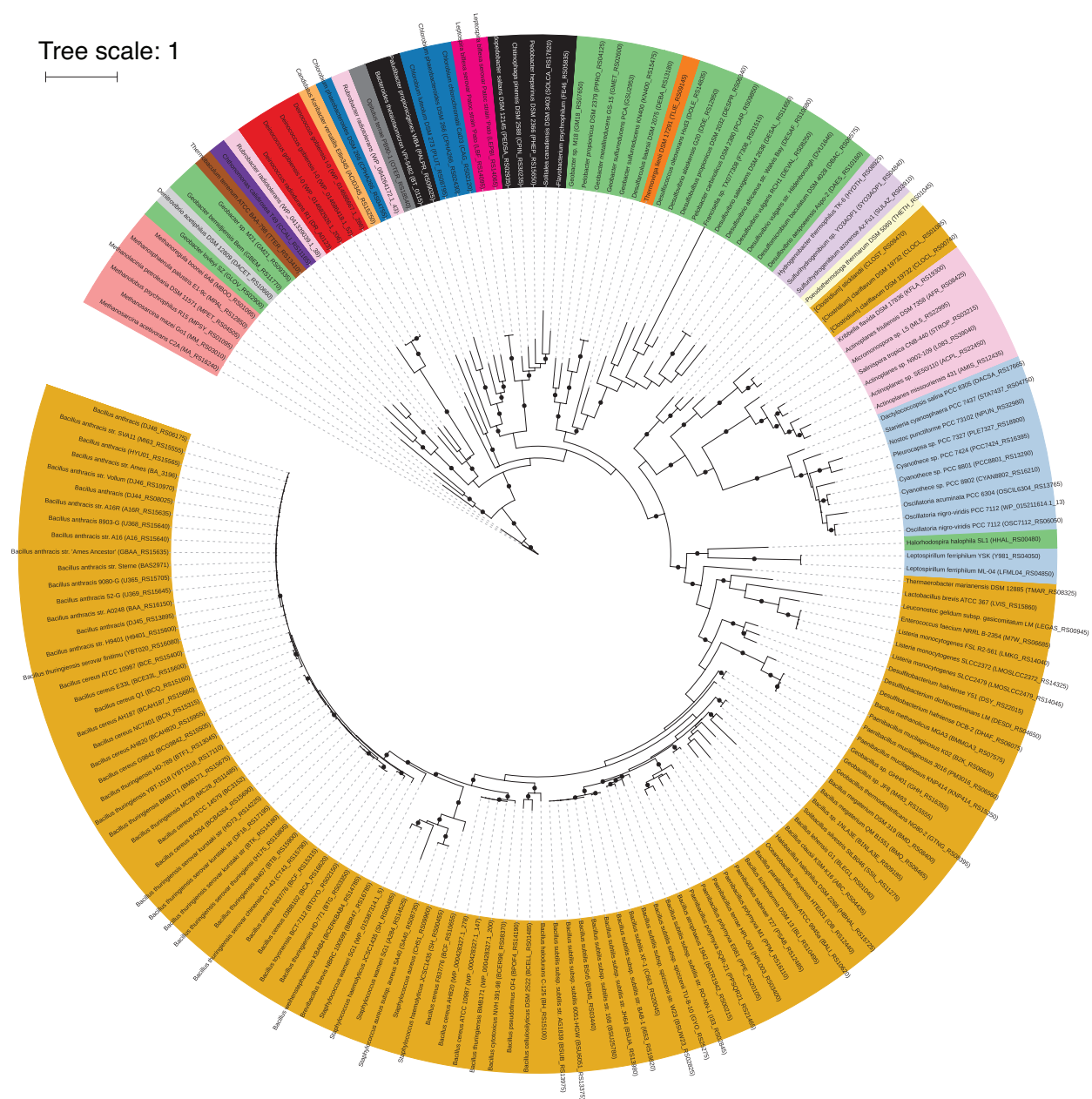
Maximum likelihood tree with 100 bootstrap replications of Acr3 sequences from RefSoil+ organisms. Leaf tips show the name of the RefSoil+ organisms and background color indicates phylum-level taxonomy. Bootstrap values  $> 50$  are represented by black circles within the tree.



**Appendix B Figure 15. Phylogeny of ArsB in RefSoil+ organisms.**


Maximum likelihood tree with 100 bootstrap replications of ArsB sequences from RefSoil+ organisms. Leaf tips show the name of the RefSoil+ organisms and background color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the tree.

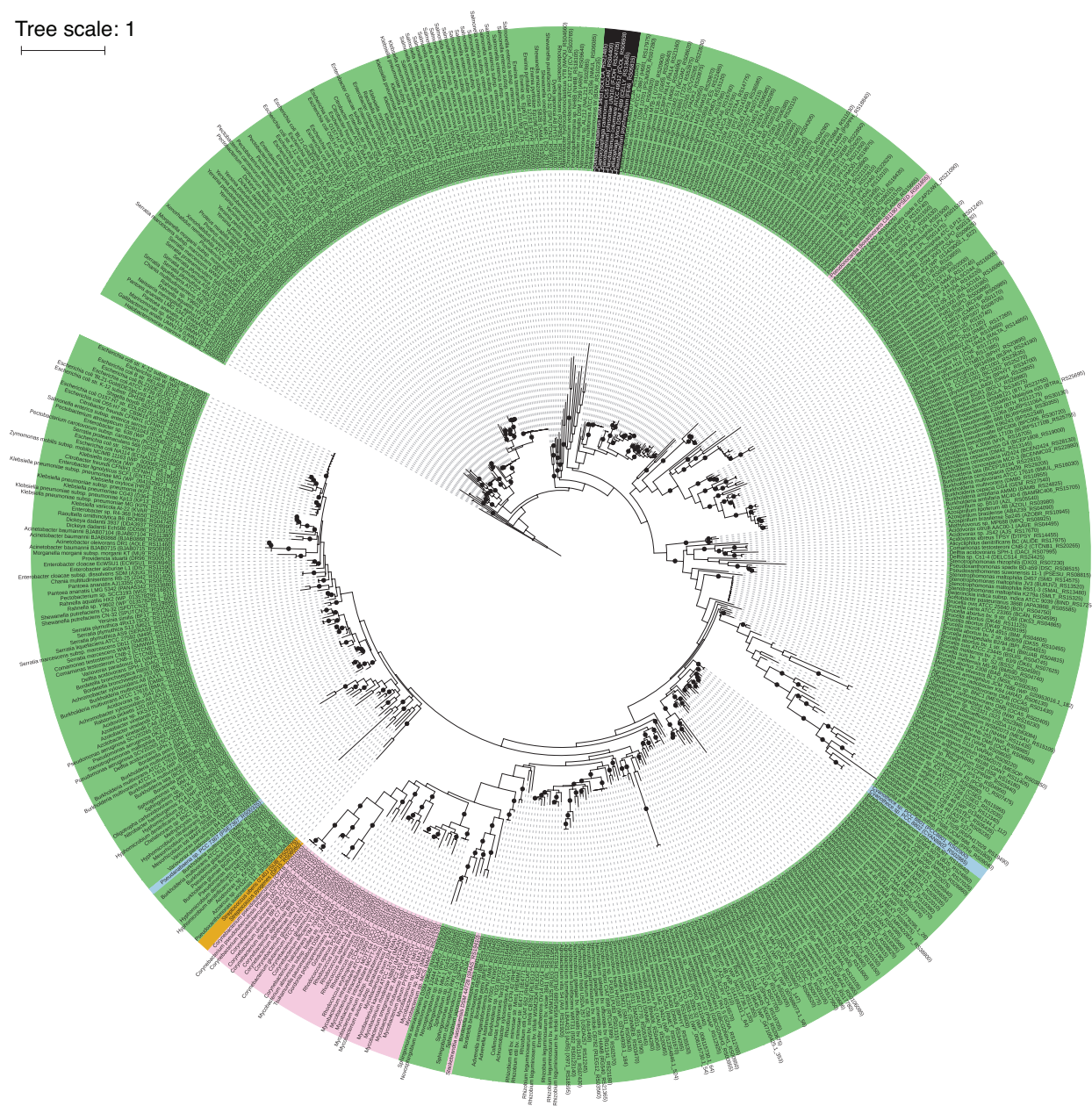




**Appendix B Figure 16. Phylogeny of *ArsC* (trx) in RefSoil+ organisms.**

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

Tree scale: 1  


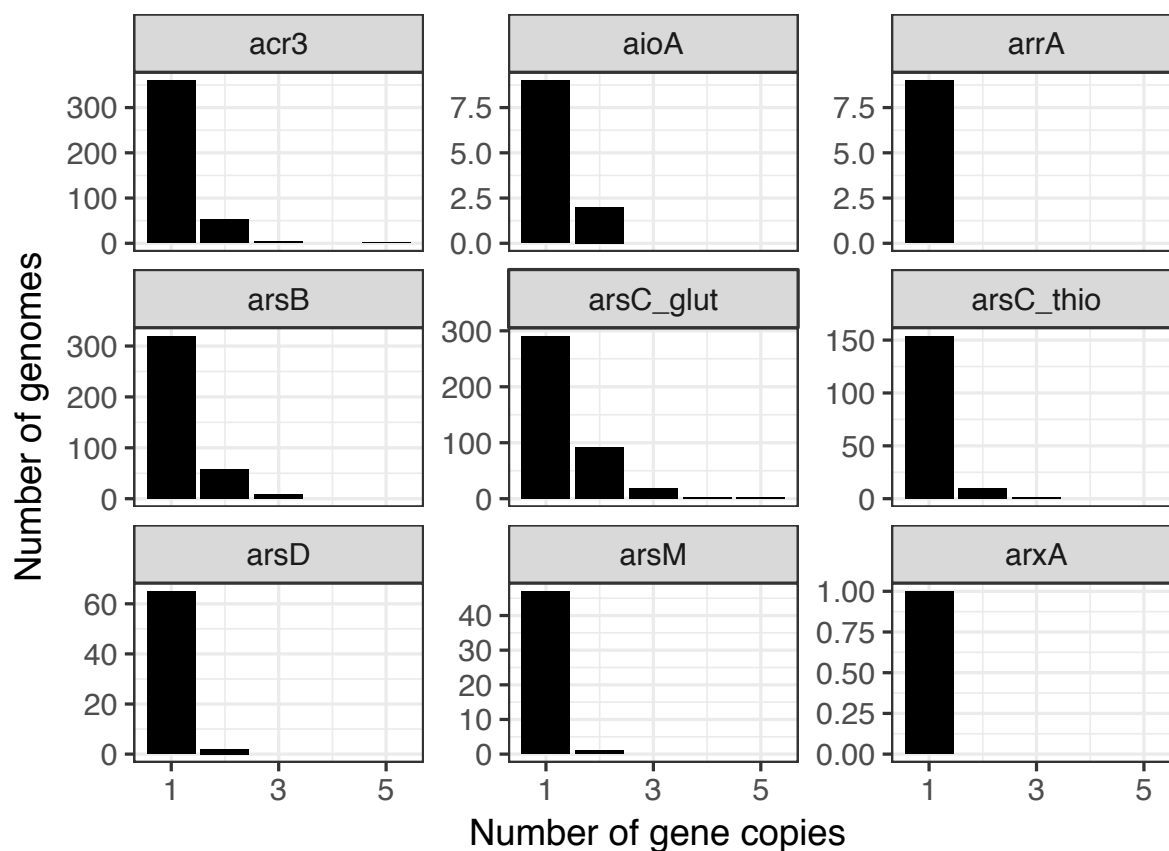


**Appendix B Figure 17. Phylogeny of ArsC (grx) in RefSoil+ organisms.**

Maximum likelihood tree with 100 bootstrap replications of ArsC (grx) sequences from RefSoil+ organisms. Leaf tips show the name of the RefSoil+ organisms and background color indicates phylum-level taxonomy. Bootstrap values > 50 are represented by black circles within the tree.

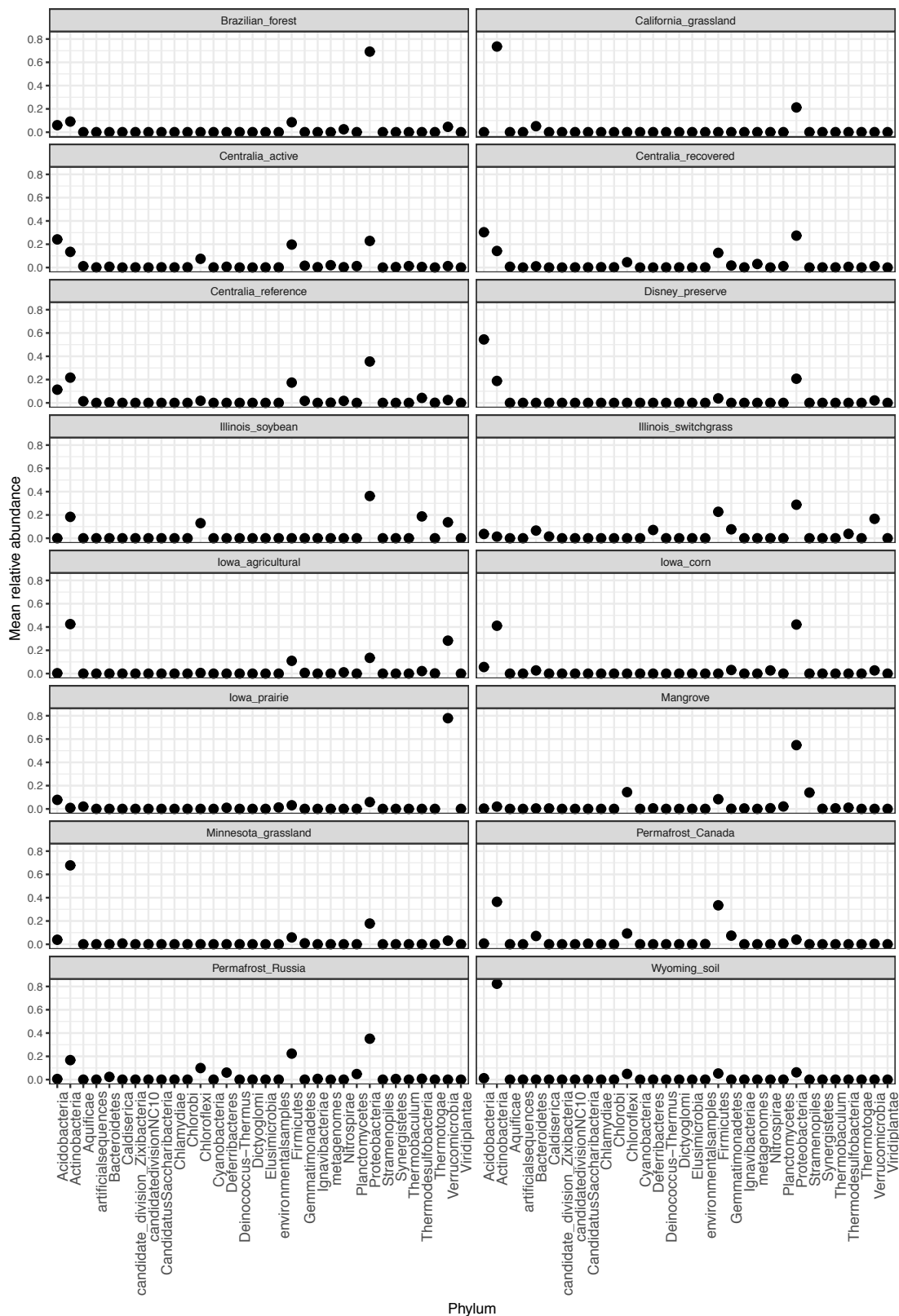






**Appendix B Figure 19. Histogram of arsenic related gene copy numbers in RefSoil+ organisms.**

Total copy number is based on hits from both chromosomes and plasmids from the same organism.



Appendix B Figure 20. Phylum-level community structure of soil metagenomes.

## REFERENCES

## REFERENCES

1. Zhu YG, Xue XM, Kappler A, Rosen BP, Meharg AA. 2017. Linking Genes to Microbial Biogeochemical Cycling: Lessons from Arsenic. *Environ Sci Technol* 51:7326–7339.
2. Summers A, Jacoby GA. 1977. Plasmid-Determined Resistance to Tellurium Compounds. *J Bacteriol* 129:276–281.
3. Ghosh A, Singh A, Ramteke PW, Singh VP. 2000. Characterization of Large Plasmids Encoding Resistance to Toxic Heavy Metals in *Salmonella abortus equi*. *Biochem Biophys Res Commun* 272:6–11.
4. Whelan KF, Sherburne RK, Taylor DE. 1997. Characterization of a region of the IncHI2 plasmid R478 which protects *Escherichia coli* from toxic effects specified by components of the tellurite, phage, and colicin resistance cluster. *J Bacteriol* 178:63–71.
5. Environmental Protection Agency. 2014. Title 40 - Protection of Environment: Appendix A to Part 423 - 126 Priority Pollutants. Code Fed Regul 29.
6. Jackson CR, Dugas SL, Harrison KG. 2005. Enumeration and characterization of arsenate-resistant bacteria in arsenic free soils. *Soil Biol Biochem* 37:2319–2322.
7. Wang P, Sun G, Jia Y, Meharg AA, Zhu Y. 2014. A review on completing arsenic biogeochemical cycle: Microbial volatilization of arsines in environment. *J Environ Sci (China)* 26:371–381.
8. Rosen BP. 2002. Biochemistry of arsenic detoxification. *FEBS Lett* 529:86–92.
9. Naujokas MF, Anderson B, Ahsan H, Vasken Aposhian H, Graziano JH, Thompson C, Suk WA. 2013. The broad scope of health effects from chronic arsenic exposure: Update on a worldwide public health problem. *Environ Health Perspect* 121:295–302.
10. WHO. 2008. Guidelines for Drinking-water Quality.
11. Yamamura S, Amachi S. 2014. Microbiology of inorganic arsenic: From metabolism to bioremediation. *J Biosci Bioeng* 118:1–9.
12. Cullen WR, Reimer KJ. 1989. Arsenic speciation in the environment. *Chem Rev* 89:713–

764.

13. Smith E, Naidu R, Alston AM, Donald LS. 1998. Arsenic in the Soil Environment: A Review. *Adv Agron* Volume 64:149–195.
14. Huang JH. 2014. Impact of microorganisms on arsenic biogeochemistry: A review. *Water Air Soil Pollut* 225:1848.
15. Zhu Y-G, Yoshinaga M, Zhao F-J, Rosen BP. 2014. Earth Abides Arsenic Biotransformations. *Annu Rev Earth Planet Sci* 42:443–467.
16. Jackson CR, Dugas SL. 2003. Phylogenetic analysis of bacterial and archaeal *arsC* gene sequences suggests an ancient, common origin for arsenate reductase. *BMC Evol Biol* 3:18.
17. Sforza 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. *Nat Geosci* 7:811–815.
18. van Lis R, Nitschke W, Duval S, Schoepp-Cothenet B. 2013. Arsenic as bioenergetic substrates. *Biochim Biophys Acta* 1827:176–188.
19. Chen S-C, Sun G-X, Rosen BP, Zhang S-Y, Deng Y, Zhu B-K, Rensing C, Zhu Y-G. 2017. Recurrent horizontal transfer of arsenite methyltransferase genes facilitated adaptation of life to arsenic. *Sci Rep* 7:7741.
20. Lebrun E, Brugna M, Baymann F, Muller D, Lièvremon D, Lett M-C, Nitschke W. 2003. Arsenite oxidase, an ancient bioenergetic enzyme. *Mol Biol Evol* 20:686–693.
21. Andres J, Bertin PN. 2016. The microbial genomics of arsenic. *FEMS Microbiol Rev* 40:299–322.
22. Van Der Merwe JA, Deane SM, Rawlings DE. 2010. The chromosomal arsenic resistance genes of *Sulfobacillus thermosulfidooxidans*. *Hydrometallurgy* 104:477–482.
23. Rosen BP. 1999. Families of arsenic transporters. *Trends Microbiol* 7:207–212.
24. Dey S, Dou D, Rosen BP. 1994. ATP-dependent arsenite transport in everted membrane vesicles of *Escherichia coli*. *J Biol Chem* 269:25442–25446.

25. Ajees AA, Yang J, Rosen BP. 2011. The ArsD as(III) metallochaperone. *BioMetals* 24:391–399.
26. Fu HL, Meng Y, Ordóñez E, Villadangos AF, Bhattacharjee H, Gil JA, Mateos LM, Rosen BP. 2009. Properties of arsenite efflux permeases (Acr3) from *Alkaliphilus metalliredigens* and *Corynebacterium glutamicum*. *J Biol Chem* 284:19887–19895.
27. Tatusov RL, Galperin MY, Natale DA, Koonin E V. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 28:33–36.
28. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin E V., Krylov DM, Mazumder R, Smirnov S, Nikolskaya AN, Rao BS, Mekhedov SL, Sverlov A V., Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2003. The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4:1–14.
29. Achour AR, Bauda P, Billard P. 2007. Diversity of arsenite transporter genes from arsenic-resistant soil bacteria. *Res Microbiol* 158:128–137.
30. Cai L, Liu G, Rensing C, Wang G. 2009. Genes involved in arsenic transformation and resistance associated with different levels of arsenic-contaminated soils. *BMC Microbiol* 9:4.
31. Kurth D, Amadio A, Ordoñez OF, Albarracín VH, Gärtner W, Farías ME. 2017. Arsenic metabolism in high altitude modern stromatolites revealed by metagenomic analysis. *Sci Rep* 7:1024.
32. Zegers I, Martins JC, Willem R, Wyns L, Messens J. 2001. Arsenate reductase from *S. aureus* plasmid pI258 is a phosphatase drafted for redox duty. *Nat Struct Biol* 8:843–847.
33. Ryan D, Collieran E. 2002. Arsenical resistance in the IncHI2 plasmids. *Plasmid* 47:234–240.
34. Villegas-Torres MF, Bedoya-Reina OC, Salazar C, Vives-Florez MJ, Dussan J. 2011. Horizontal *arsC* gene transfer among microorganisms isolated from arsenic polluted soil. *Int Biodeterior Biodegrad* 65:147–152.
35. Oden KL, Gladysheva TB, Rosen BP. 1994. Arsenate reduction mediated by the plasmid-encoded ArsC protein is coupled to glutathione. *Mol Microbiol* 12:301–306.
36. Martin P, DeMel S, Shi J, Gladysheva T, Gatti DL, Rosen BP, Edwards BFP. 2001.

Insights into the structure, solvation, and mechanism of ArsC arsenate reductase, a novel arsenic detoxification enzyme. *Structure* 9:1071–1081.

37. Escudero L V, Casamayor EO, Chong G, Pedrós-Alió C, Demergasso C. 2013. Distribution of microbial arsenic reduction, oxidation and extrusion genes along a wide range of environmental arsenic concentrations. *PLoS One* 8:e78890.
38. Marapakala K, Packianathan C, Ajees AA, Dheeman DS, Sankaran B, Kandavelu P, Rosen BP. 2015. A disulfide-bond cascade mechanism for arsenic(III) S-adenosylmethionine methyltransferase. *Acta Crystallogr Sect D Biol Crystallogr* 71:505–515.
39. Bhattacharjee H, Rosen BP. 2007. Arsenic metabolism in prokaryotic and eukaryotic microbes, p. 371–406. *In* Nies, D, Silver, S (eds.), *Molecular Microbiology of Heavy Metals*, 6th ed. Springer-Verlag, Berlin, Heidelberg.
40. Palmgren M, Engström K, Hallström BM, Wahlberg K, Søndergaard DA, Sall T, Vahter M, Broberg K. 2017. AS3MT-mediated tolerance to arsenic evolved by multiple independent horizontal gene transfers from bacteria to eukaryotes. *PLoS One* 12:e0175422.
41. Engel AS, Johnson LR, Porter ML. 2013. Arsenite oxidase gene diversity among Chloroflexi and Proteobacteria from El Tatio Geyser Field, Chile. *FEMS Microbiol Ecol* 83:745–756.
42. Macur RE, Jackson CR, Botero LM, McDermott T., Inskeep W. 2004. Bacterial populations associated with the oxidation and reduction of arsenic in an unsaturated soil RN - *Environ. Sci. Technol.*, vol. 38, pp. 104-111 38:104–111.
43. Heinrich-Salmeron A, Cordi A, Brochier-Armanet C, Halter D, Pagnout C, Abbaszadeh-Fard E, Montaut D, Seby F, Bertin PN, Bauda P, Arsene-Ploetze F. 2011. Unsuspected diversity of arsenite-oxidizing bacteria as revealed by widespread distribution of the *aoxB* Gene in prokaryotes. *Appl Environ Microbiol* 77:4685–4692.
44. Lett M-C, Muller D, Lièvreumont D, Silver S, Santini J. 2012. Unified nomenclature for genes involved in prokaryotic aerobic arsenite oxidation. *J Bacteriol* 194:207–208.
45. Lebrun E, Brugna M, Baymann F, Muller D, Lett M-C, Nitschke W. 2003. Arsenite Oxidase , an Ancient Bioenergetic Enzyme. *Mol Biol Evol* 20:686–693.
46. Zargar K, Conrad A, Bernick DL, Lowe TM, Stolc V, Hoeft S, Oremland RS, Stolz JF,

- Saltikov CW. 2012. ArxA, a new clade of arsenite oxidase within the DMSO reductase family of molybdenum oxidoreductases. *Environ Microbiol* 14:1635–1645.
47. Zargar K, Hoefft S, Oremland R, Saltikov CW. 2010. Identification of a novel arsenite oxidase gene, *arxA*, in the Haloalkaliphilic, arsenite-Oxidizing bacterium *Alkalilimnicola ehrlichii* strain MLHE-1. *J Bacteriol* 192:3755–3762.
  48. Hug K, Maher WA, Stott MB, Krikowa F, Foster S, Moreau JW. 2014. Microbial contributions to coupled arsenic and sulfur cycling in the acid-sulfide hot spring Champagne Pool, New Zealand. *Front Microbiol* 5:1–14.
  49. Oremland RS, Stolz JF. 2003. The Ecology of Arsenic. *Source Sci New Ser* 300:939–944.
  50. Mukhopadhyay R, Rosen BP, Phung LT, Silver S. 2002. Microbial arsenic: From geocycles to genes and enzymes. *FEMS Microbiol Rev* 26:311–325.
  51. Kavitha Marapakala, Jie Qin BPR. 2012. Identification of catalytic residues in the As(III) S-Adenosylmethionine Methyltransferase. *Biochemistry* 51:944–951.
  52. Masafumi Yoshinaga, Yong Cai BPR. 2011. Demethylation of methylarsonic acid by a microbial community. *Environ Microbiol* 13:1205–1215.
  53. Mandal BK, Suzuki KT. 2002. Arsenic round the world: A review. *Talanta* 58:201–235.
  54. 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.
  55. Zhang SY, Zhao FJ, Sun GX, Su JQ, Yang XR, Li H, Zhu YG. 2015. Diversity and abundance of arsenic biotransformation genes in paddy soils from southern china. *Environ Sci Technol* 49:4138–4146.
  56. Desoeuvre A, Casiot C, Héry M. 2015. Diversity and Distribution of Arsenic-Related Genes Along a Pollution Gradient in a River Affected by Acid Mine Drainage. *Microb Ecol* 71:672–685.
  57. Hu M, Sun W, Krumins V, Li F. 2019. Arsenic contamination influences microbial community structure and putative arsenic metabolism gene abundance in iron plaque on paddy rice root. *Sci Total Environ* 649:405–412.
  58. Costa PS, Reis MP, Avila MP, Leite LR, De Araujo FMG, Salim ACM, Oliveira G,



- Barbosa F, Chartone-Souza E, Nascimento AMA. 2015. Metagenome of a microbial community inhabiting a metal-rich tropical stream sediment. *PLoS One* 10:e0119465.
59. Lozupone CA, Knight R. 2007. Global patterns in bacterial diversity. *Proc Natl Acad Sci* 104:11436–11440.
  60. Fierer N, Jackson RB. 2006. The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci* 103:626–631.
  61. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciulek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauser A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Rivera JLA, Al-Moosawi L, Alverdy J, Amato KR, Andras J, Angenent LT, Antonopoulos DA, Apprill A, Armitage D, Ballantine K, Bárta J, Baum JK, Berry A, Bhatnagar A, Bhatnagar M, Biddle JF, Bittner L, Boldgiv B, Bottos E, Boyer DM, Braun J, Brazelton W, Brearley FQ, Campbell AH, Caporaso JG, Cardona C, Carroll J, Cary SC, Casper BB, Charles TC, Chu H, Claar DC, Clark RG, Clayton JB, Clemente JC, Cochran A, Coleman ML, Collins G, Colwell RR, Contreras M, Crary BB, Creer S, Cristol DA, Crump BC, Cui D, Daly SE, Davalos L, Dawson RD, Defazio J, Delsuc F, Dionisi HM, Dominguez-Bello MG, Dowell R, Dubinsky EA, Dunn PO, Ercolini D, Espinoza RE, Ezenwa V, Fenner N, Findlay HS, Fleming ID, Fogliano V, Forsman A, Freeman C, Friedman ES, Galindo G, Garcia L, Garcia-Amado MA, Garshelis D, Gasser RB, Gerdts G, Gibson MK, Gifford I, Gill RT, Giray T, Gittel A, Golyshin P, Gong D, Grossart H-P, Guyton K, Haig S-J, Hale V, Hall RS, Hallam SJ, Handley KM, Hasan NA, Haydon SR, Hickman JE, Hidalgo G, Hofmockel KS, Hooker J, Hulth S, Hultman J, Hyde E, Ibáñez-Álamo JD, Jastrow JD, Jex AR, Johnson LS, Johnston ER, Joseph S, Jurburg SD, Jurelevicius D, Karlsson A, Karlsson R, Kauppinen S, Kellogg CTE, Kennedy SJ, Kerkhof LJ, King GM, Kling GW, Koehler A V., Krezalek M, Kueneman J, Lamendella R, Landon EM, Lane-deGraaf K, LaRoche J, Larsen P, Laverock B, Lax S, Lentino M, Levin II, Liancourt P, Liang W, Linz AM, Lipson DA, Liu Y, Lladser ME, Lozada M, Spirito CM, MacCormack WP, MacRae-Crerar A, Magris M, Martín-Platero AM, Martín-Vivaldi M, Martínez LM, Martínez-Bueno M, Marzinelli EM, Mason OU, Mayer GD, McDevitt-Irwin JM, McDonald JE, McGuire KL, McMahon KD, McMinds R, Medina M, Mendelson JR, Metcalf JL, Meyer F, Michelangeli F, Miller K, Mills DA, Minich J, Mocali S, Moitinho-Silva L, Moore A, Morgan-Kiss RM, Munroe P, Myrold D, Neufeld JD, Ni Y, Nicol GW, Nielsen S, Nissimov JI, Niu K, Nolan MJ, Noyce K, O'Brien SL, Okamoto N, Orlando L, Castellano YO, Osuolale O, Oswald W, Parnell J, Peralta-Sánchez JM, Petraitis P, Pfister C, Pilon-Smits E, Piombino P, Pointing SB, Pollock FJ, Potter C, Prithiviraj B, Quince C, Rani A, Ranjan R, Rao S, Rees AP, Richardson M, Riebesell U, Robinson C, Rockne KJ, Rodriguez SM, Rohwer F, Roundstone W, Safran RJ, Sangwan N, Sanz V, Schrenk M, Schrenzel MD, Scott NM, Seger RL, Seguin-Orlando A, Seldin L, Seyler LM, Shakhsher B, Sheets GM, Shen C, Shi Y, Shin H, Shogan BD, Shutler D, Siegel J, Simmons S, Sjöling S, Smith DP, Soler JJ, Sperling M,

- Steinberg PD, Stephens B, Stevens MA, Taghavi S, Tai V, Tait K, Tan CL, Tas, N, Taylor DL, Thomas T, Timling I, Turner BL, Urich T, Ursell LK, van der Lelie D, Van Treuren W, van Zwieten L, Vargas-Robles D, Thurber RV, Vitaglione P, Walker DA, Walters WA, Wang S, Wang T, Weaver T, Webster NS, Wehrle B, Weisenhorn P, Weiss S, Werner JJ, West K, Whitehead A, Whitehead SR, Whittingham LA, Willerslev E, Williams AE, Wood SA, Woodhams DC, Yang Y, Zaneveld J, Zarraonaindia I, Zhang Q, Zhao H. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551:457–463.
62. 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. *Environ Pollut* 211:1–8.
  63. Edwardson CF, Hollibaugh JT. 2017. Metatranscriptomic analysis of prokaryotic communities active in sulfur and arsenic cycling in Mono Lake, California, USA. *ISME J* 11:2195–2208.
  64. Lauber CL, Hamady M, Knight R, Fierer N. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* 75:5111–5120.
  65. Baker-Austin C, Wright MS, Stepanauskas R, McArthur J V. 2006. Co-selection of antibiotic and metal resistance. *Trends Microbiol* 14:176–182.
  66. Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DGJ. 2015. Co-occurrence of resistance genes to antibiotics, biocides and metals reveals novel insights into their co-selection potential. *BMC Genomics* 16:964.
  67. Li L-G, Xia Y, Zhang T. 2017. Co-occurrence of antibiotic and metal resistance genes revealed in complete genome collection. *ISME J* 11:651–662.
  68. Li A-D, Li L-G, Zhang T. 2015. Exploring antibiotic resistance genes and metal resistance genes in plasmid metagenomes from wastewater treatment plants. *Front Microbiol* 6:1025.
  69. Ma Y, Metch JW, Yang Y, Pruden A, Zhang T. 2016. Shift in antibiotic resistance gene profiles associated with nanosilver during wastewater treatment. *FEMS Microbiol Ecol* 92:1–8.
  70. Andrea W, Calderón G, Mercedes M, Vargas B, Alirio W, Suárez B, Andrés L, Rodríguez Y. 2013. Horizontal transfer of heavy metal and antibiotic-resistance markers between indigenous bacteria, colonizing mercury contaminated tailing ponds in southern

- Venezuela, and human pathogens. *Rev la Soc Venez Microbiol* 33:110–115.
71. Centers for Disease Control and Prevention US. 2013. Antibiotic Resistance Threats in United States, 2013.
  72. Forsberg KJ, Reyes A, Wang B, Selleck EM, Sommer MO, Dantas G. 2012. The shared antibiotic resistome of soil bacteria and human pathogens. *Science* (80- ) 337:1107–1111.
  73. Poirel L, Liard A, Nordmann P, Mammeri H. 2005. Origin of Plasmid-Mediated Quinolone Resistance Determinant QnrA. *Antimicrob Agents Chemother* 49:3523–3525.
  74. Patel R, Piper K, Cockerill FR, Steckelberg JM, Yousten AA. 2000. The biopesticide *Paenibacillus popilliae* has a vancomycin resistance gene cluster homologous to the enterococcal VanA vancomycin resistance gene cluster. *Antimicrob Agents Chemother* 44:705–709.
  75. Nesme J, Simonet P. 2015. The soil resistome: A critical review on antibiotic resistance origins, ecology and dissemination potential in telluric bacteria. *Environ Microbiol* 17:913–930.
  76. Finley RL, Collignon P, Larsson DGJ, McEwen SA, Li X-Z, Gaze WH, Reid-Smith R, Timinouni M, Graham DW, Topp E. 2013. The Scourge of Antibiotic Resistance: The Important Role of the Environment. *Clin Infect Dis* 57:704–710.
  77. Berendonk TU, Manaia CM, Merlin C, Fatta-Kassinos D, Cytryn E, Walsh F, Bürgmann H, Sørum H, Norström M, Pons M-N, Kreuzinger N, Huovinen P, Stefani S, Schwartz T, Kisand V, Baquero F, Martinez JL. 2015. Tackling antibiotic resistance: the environmental framework. *Nat Rev Microbiol* 13:310–317.
  78. Allen HK, Donato J, Wang HH, Cloud-Hansen KA, Davies J, Handelsman J. 2010. Call of the wild: antibiotic resistance genes in natural environments. *Nat Rev Microbiol* 8:251–259.
  79. Laine A-L, Hiltunen T, Virta M. 2016. Antibiotic resistance in the wild: an eco-evolutionary perspective. *Philos Trans R Soc B Biol Sci* 372:20160039.
  80. Aminov RI, Mackie RI. 2007. Evolution and ecology of antibiotic resistance genes. *FEMS Microbiol Lett* 271:147–161.
  81. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA,

- Wall DH, Caporaso JG. 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci* 109:21390–21395.
82. Forsberg KJ, Patel S, Gibson MK, Lauber CL, Fierer N, Dantas G. 2014. Bacterial phylogeny structures soil resistomes across habitats. *Nature* 509:612–616.
  83. Tian Z, Zhang Y, Yu B, Yang M. 2016. Changes of resistome, mobilome and potential hosts of antibiotic resistance genes during the transformation of anaerobic digestion from mesophilic to thermophilic. *Water Res* 98:261–269.
  84. Qian X, Sun W, Gu J, Wang XJ, Zhang YJ, Duan ML, Li HC, Zhang RR. 2016. Reducing antibiotic resistance genes, integrons, and pathogens in dairy manure by continuous thermophilic composting. *Bioresour Technol* 220:425–432.
  85. van Elsas JD, Bailey MJ. 2002. The ecology of transfer of mobile genetic elements. *FEMS Microb Ecol* 42:187–197.
  86. Achour AR, Bauda P, Billard P. 2007. Diversity of arsenite transporter genes from arsenic-resistant soil bacteria. *Res Microbiol* 158:128–137.
  87. Jackson CR, Harrison KG, Dugas SL. 2005. Enumeration and characterization of culturable arsenate resistant bacteria in a large estuary. *Syst Appl Microbiol* 28:727–734.
  88. D’Costa VM, King CE, Kalan L, Morar M, Sung WW, Schwarz C, Froese D, Zazula G, Calmels F, Debruyne R, Golding GB, Poinar HN, Wright GD. 2011. Antibiotic resistance is ancient. *Nature* 477:457–461.
  89. Bhullar K, Waglechner N, Pawlowski A, Koteva K, Banks ED, Johnston MD, Barton HA, Wright GD. 2012. Antibiotic resistance is prevalent in an isolated cave microbiome. *PLoS One* 7:e34953.
  90. Jia Y, Huang H, Zhong M, Wang F, Zhang L, Zhu Y-G. 2013. Microbial arsenic methylation in soil and rice rhizosphere. *Environ Sci Technol* 47:3141–3148.
  91. Gómez P, Buckling A. 2013. Real-time microbial adaptive diversification in soil. *Ecol Lett* 16:650–655.
  92. Aertsen A, Michiels CW. 2005. Diversify or die: generation of diversity in response to stress. *Crit Rev Microbiol* 31:69–78.
  93. Aminov RI. 2011. Horizontal gene exchange in environmental microbiota. *Front*

Microbiol 2:1–19.

94. Williams HG, Day MJ, Fry JC, Stewart GJ. 1996. Natural transformation in river epilithion. *Appl Environ Microbiol* 62:2994–2998.
95. Elick JM. 2011. Mapping the coal fire at Centralia, Pa using thermal infrared imagery. *Int J Coal Geol* 87:197–203.
96. Melody SM, Johnston FH. 2015. Coal mine fires and human health: What do we know? *Int J Coal Geol* 152, Part:1–14.
97. Janzen C, Tobin-Janzen T. 2008. Microbial Communities in Fire-Affected Soils, p. 299–316. *In* Dion, P, Nautiyal, CS (eds.), *Microbiology of Extreme Soils*. Springer Berlin Heidelberg.
98. Pone JDN, Hein K a a, Stracher GB, Annegarn HJ, Finkleman RB, Blake DR, McCormack JK, Schroeder P. 2007. The spontaneous combustion of coal and its by-products in the Witbank and Sasolburg coalfields of South Africa. *Int J Coal Geol* 72:124–140.
99. Han FX, Su Y, Monts DL, Plodinec MJ, Banin A, Triplett GE. 2003. Assessment of global industrial-age anthropogenic arsenic contamination. *Naturwissenschaften* 90:395–401.
100. Bahar MM, Megharaj M, Naidu R. 2013. Bioremediation of arsenic-contaminated water: Recent advances and future prospects. *Water Air Soil Pollut* 224:1–20.
101. Cavalca L, Zanchi R, Corsini A, Colombo M, Romagnoli C, Canzi E, Andreoni V. 2010. Arsenic-resistant bacteria associated with roots of the wild *Cirsium arvense* (L.) plant from an arsenic polluted soil, and screening of potential plant growth-promoting characteristics. *Syst Appl Microbiol* 33:154–164.
102. Jobby R, Shah K, Shah R, Jha P, Desai N. 2016. Differential Expression of Antioxidant Enzymes under Arsenic Stress in *Enterobacter* Sp. *Environ Prog Sustain Energy* 35:1642–1645.
103. Zhang Y, Chen S, Hao X, Su JQ, Xue X, Yan Y, Zhu YG, Ye J. 2016. Transcriptomic analysis reveals adaptive responses of an enterobacteriaceae strain LSJC7 to arsenic exposure. *Front Microbiol* 7:636.

104. Parvatiyar K, Alsabbagh EM, Ochsner U a, Stegemeyer M a, Smulian AG, Hwang SH, Jackson CR, Mcdermott TR, Daniel J, Hassett DJ. 2005. Global Analysis of Cellular Factors and Responses Involved in *Pseudomonas aeruginosa* Resistance to Arsenite Global Analysis of Cellular Factors and Responses Involved in *Pseudomonas aeruginosa* Resistance to Arsenite 187:4853–4864.
105. Sacheti P, Bhonsle H, Patil R, Kulkarni MJ, Srikanth R, Gade W. 2013. Arsenomics of *Exiguobacterium* sp. PS (NCIM 5463). *RSC Adv* 3:9705.
106. Belfiore C, Ordoñez OF, Farías ME. 2013. Proteomic approach of adaptive response to arsenic stress in *Exiguobacterium* sp. S17, an extremophile strain isolated from a high-altitude Andean Lake stromatolite. *Extremophiles* 17:421–431.
107. Brauner A, Fridman O, Gefen O, Balaban NQ. 2016. Distinguishing between resistance, tolerance and presistance to antibiotic treatment. *Nat Rev Microbiol* 14:320–330.
108. Anderson CR, Cook GM. 2004. Isolation and Characterization of Arsenate-Reducing Bacteria from Arsenic-Contaminated Sites in New Zealand 48:341–347.
109. Pepi M, Volterrani M, Renzi M, Marvasi M, Gasperini S, Franchi E, Focardi SE. 2007. Arsenic-resistant bacteria isolated from contaminated sediments of the Orbetello Lagoon, Italy, and their characterization. *J Appl Microbiol* 103:2299–2308.
110. Drewniak L, Styczek A, Majder-Lopatka M, Sklodowska A. 2008. Bacteria, hypertolerant to arsenic in the rocks of an ancient gold mine, and their potential role in dissemination of arsenic pollution. *Environ Pollut* 156:1069–1074.
111. Sarkar A, Kazy SK, Sar P. 2013. Characterization of arsenic resistant bacteria from arsenic rich groundwater of West Bengal, India. *Ecotoxicology* 22:363–376.
112. Zeng XC, E G, Wang J, Wang N, Chen X, Mu Y, Li H, Yang Y, Liu Y, Wang Y. 2016. Functions and unique diversity of genes and microorganisms involved in arsenite oxidation from the tailings of a realgar mine. *Appl Environ Microbiol* 82:7019–7029.
113. Lewis K, Epstein S, Onofrio AD, Ling LL. 2010. Uncultured microorganisms as a source of secondary metabolites. *J Antibiot (Tokyo)* 63:468–476.
114. Nichols D, Cahoon N, Trakhtenberg EM, Pham L, Mehta A, Belanger A, Kanigan T, Lewis K, Epstein SS. 2010. Use of Ichip for High-Throughput In Situ Cultivation of “Uncultivable” Microbial Species. *Appl Environ Microbiol* 76:2445–2450.

115. Salcher MM, Šimek K. 2016. Isolation and cultivation of planktonic freshwater microbes is essential for a comprehensive understanding of their ecology. *Aquat Microb Ecol* 77:183–196.
116. Prakash O, Shouche Y, Jangid K, Kostka JE. 2013. Microbial cultivation and the role of microbial resource centers in the omics era. *Appl Microbiol Biotechnol* 97:51–62.
117. Madsen EL. 2005. Identifying microorganisms responsible for ecologically significant biogeochemical processes. *Nat Rev Microbiol* 3:439–446.
118. Overmann J, Abt B, Sikorski J. 2017. Present and Future of Culturing Bacteria. *Annu Rev Microbiol* 13:711–730.
119. Smalla K, Jechalke S, Top EM. 2015. Plasmid detection, characterization and ecology. *Cancer* 121:1265–1272.
120. Tamames J, Moya A. 2008. Estimating the extent of horizontal gene transfer in metagenomic sequences. *BMC Genomics* 15:136.
121. Shade A, Hogan CS, Klimowicz AK, Linske M, Mcmanus PS, Handelsman J. 2012. Culturing captures members of the soil rare biosphere. *Environ Microbiol* 14:2247–2252.
122. Edelstein AD, Tsuchida MA, Amodaj N, Pinkard H, Vale RD, Stuurman N. 2014. Advanced methods of microscope control using µManager software. *J Biol Methods* 1:10.
123. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, Tinevez J-Y, White DJ, Hartenstein V, Eliceiri K, Tomancak P, Cardona A. 2012. Fiji: an open-source platform for biological-image analysis. *Nat Meth* 9:676–682.
124. Huang X, Madan a. 1999. CAP 3: A DNA sequence assembly program. *Genome Res* 9:868–877.
125. Baker GC, Smith JJ, Cowan DA. 2003. Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods* 55:541–555.
126. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267.
127. Kim OS, Cho YJ, Lee K, Yoon SH, Kim M, Na H, Park SC, Jeon YS, Lee JH, Yi H, Won

- S, Chun J. 2012. Introducing EzTaxon-e: A prokaryotic 16s rRNA gene sequence database with phylotypes that represent uncultured species. *Int J Syst Evol Microbiol* 62:716–721.
128. Sun Y, Polishchuk EA, Radoja U, Cullen WR. 2004. Identification and quantification of *arsC* genes in environmental samples by using real-time PCR. *J Microbiol Methods* 58:335–349.
129. Song B, Chyun E, Jaffé PR, Ward BB. 2009. Molecular methods to detect and monitor dissimilatory arsenate-respiring bacteria (DARB) in sediments. *FEMS Microbiol Ecol* 68:108–17.
130. Quemeneur M, Heinrich-Salmeron A, Muller D, Lièvremon D, Jauzein M, Bertin PN, S, Jouliau C. 2008. Diversity Surveys and Evolutionary Relationships of *aoxB* Genes in Aerobic Arsenite-Oxidizing Bacteria. *Appl Environ Microbiol* 74:4567–4573.
131. Chang J-S, Kim Y-H, Kim K-W. 2008. The *ars* genotype characterization of arsenic-resistant bacteria from arsenic-contaminated gold-silver mines in the Republic of Korea. *Appl Microbiol Biotechnol* 80:155–65.
132. Chang JS, Yoon IH, Lee JH, Kim KR, An J, Kim KW. 2010. Arsenic detoxification potential of *aox* genes in arsenite-oxidizing bacteria isolated from natural and constructed wetlands in the Republic of Korea. *Environ Geochem Health* 32:95–105.
133. Suresh K, Prabakaran SR, Sengupta S, Shivaji S. 2004. *Bacillus indicus* sp. nov., an arsenic-resistant bacterium isolated from an aquifer in West Bengal, India. *Int J Syst Evol Microbiol* 54:1369–1375.
134. Huang A, Teplitski M, Rathinasabapathi B, Ma L. 2010. Characterization of arsenic-resistant bacteria from the rhizosphere of arsenic hyperaccumulator *Pteris vittata*. *Can J Microbiol* 56:236–246.
135. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res* 42:633–642.
136. Hall T. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser*.
137. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33:1870–1874.



138. R Development Core Team. 2008. R: A language and environment for statistical computing. CRAN.
139. Klaus S, Paradis E, Martins L de O, Potts A, White TW, Stachniss C, Kendall M. 2017. Phylogenetic analysis in R. 2.2.0. CRAN.
140. Lee S-H, Sorensen JW, Grady KL, Tobin TC, Shade A. 2017. Divergent extremes but convergent recovery of bacterial and archaeal soil communities to an ongoing subterranean coal mine fire. *ISME J* 11:1447–1459.
141. Simeonova DD, Lièvreumont D, Lagarde F, Muller DAE, Groudeva VI, Lett M-C. 2004. Microplate screening assay for the detection of arsenite-oxidizing and arsenate-reducing bacteria. *FEMS Microbiol Lett* 237:249–253.
142. Kahm M, Hasenbrink G, Lichtenberg-frate H, Ludwig J, Kschischo M. 2010. Grofit: Fitting biological growth curves. *J Stat Softw* 33:1–21.
143. Akaike H. 1973. Information theory and an extension of the maximum likelihood principle, p. 267–281. *In* Second International Symposium on Information Theory.
144. Wang L, Chen S, Xiao X, Huang X, You D, Zhou X, Deng Z. 2006. arsRBOCT arsenic resistance system encoded by linear plasmid pHZ227 in *Streptomyces* sp. strain FR-008. *Appl Environ Microbiol* 72:3738–3742.
145. Das S, Jean JS, Kar S, Chou ML, Chen CY. 2014. Screening of plant growth-promoting traits in arsenic-resistant bacteria isolated from agricultural soil and their potential implication for arsenic bioremediation. *J Hazard Mater* 272:112–120.
146. Ruta M, Pepi M, Gaggi C, Bernardini E, Focardi S, Magaldi E, Gasperini S, Volterrani M, Zanini A, Focardi SE. 2011. As(V)-reduction to As(III) by arsenic-resistant *Bacillus* spp. bacterial strains isolated from low-contaminated sediments of the Oliveri-Tindari Lagoon, Italy. *Chem Ecol* 27:207–219.
147. Pepi M, Protano G, Ruta M, Nicolardi V, Bernardini E, Focardi SE, Gaggi C. 2011. Arsenic-resistant *Pseudomonas* spp. and *Bacillus* sp. bacterial strains reducing As(V) to As(III), isolated from Alps soils, Italy. *Folia Microbiol (Praha)* 56:29–35.
148. Valverde A, Gonzalez-Tirante M, Medina-Sierra M, Santa-Regina I, Garcia-Sanchez A, Igual JM. 2011. Diversity and community structure of culturable arsenic-resistant bacteria across a soil arsenic gradient at an abandoned tungsten-tin mining area. *Chemosphere* 85:129–134.

149. Nelson LM, Parkinson D. 1978. Effect of freezing and thawing on survival of three bacterial isolates from an arctic soil. *Can J Microbiol* 24:1468–74.
150. Banerjee S, Datta S, Chattyopadhyay D, Sarkar P. 2011. Arsenic accumulating and transforming bacteria isolated from contaminated soil for potential use in bioremediation. *J Environ Sci Health A Tox Hazard Subst Environ Eng* 46:1736–47.
151. Pepi M, Borra M, Tamburrino S, Saggiomo M, Viola A, Biffali E, Balestra C, Sprovieri M, Casotti R. 2016. A *Bacillus* sp. isolated from sediments of the Sarno River mouth, Gulf of Naples (Italy) produces a biofilm biosorbing Pb(II). *Sci Total Environ* 562:588–595.
152. Kahn LH. 2016. One Health Initiative: Antimicrobial Resistance in the Environment.
153. Lang KS, Anderson JM, Schwarz S, Williamson L, Handelsman J, Singer RS. 2010. Novel florfenicol and chloramphenicol resistance gene discovered in alaskan soil by using functional metagenomics. *Appl Environ Microbiol* 76:5321–5326.
154. Fitzpatrick D, Walsh F. 2016. Antibiotic resistance genes across a wide variety of metagenomes. *FEMS Microbiol Ecol* 92:1–8.
155. Gibson MK, Forsberg KJ, Dantas G. 2014. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J* 9:1–10.
156. Wang FH, Qiao M, Chen Z, Su JQ, Zhu YG. 2015. Antibiotic resistance genes in manure-amended soil and vegetables at harvest. *J Hazard Mater* 299:215–221.
157. Hiltunen T, Virta M, Laine A-L. 2017. Antibiotic resistance in the wild : an eco-evolutionary perspective. *Philos Trans R Soc B* 372:20160039.
158. Rizzo L, Manaia C, Merlin C, Schwartz T, Dagot C, Ploy MC, Michael I, Fatta-Kassinos D. 2013. Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: A review. *Sci Total Environ* 447:345–360.
159. Kumar K, C. Gupta S, Chander Y, Singh AK. 2005. Antibiotic Use in Agriculture and Its Impact on the Terrestrial Environment, p. 1–54. *In* *Advances in Agronomy*.
160. Van Hoek AHAM, Mevius D, Guerra B, Mullany P, Roberts AP, Aarts HJM. 2011. Acquired antibiotic resistance genes: An overview. *Front Microbiol* 2:1–27.

161. Nunes I, Jacquiod S, Brejnrod A, Holm PE, Brandt KK, Priemé A, Sørensen SJ. 2016. Coping with copper : legacy effect of copper on potential activity of soil bacteria following a century of exposure. *FEMS Microbiol Ecol* 92:fiw175.
162. Garner E, Wallace JS, Argoty GA, Wilkinson C, Fahrenfeld N, Heath LS, Zhang L, Arabi M, Aga DS, Pruden A. 2016. Metagenomic profiling of historic Colorado Front Range flood impact on distribution of riverine antibiotic resistance genes. *Sci Rep* 6:38432.
163. Shade A, Peter H, Allison SD, Baho DL, Berga M, Burgmann H, Huber DH, Langenheder S, Lennon JT, Martiny JBH, Matulich KL, Schmidt TM, Handelsman J. 2012. Fundamentals of microbial community resistance and resilience. *Front Microbiol* 3:1–19.
164. Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM, Cole JR. 2013. FunGene: The functional gene pipeline and repository. *Front Microbiol* 4:291.
165. Wang Q, Fish JA, Gilman M, Sun Y, Brown CT, Tiedje JM, Cole JR. 2015. Xander : employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome* 3:32.
166. Rodriguez-R LM, Konstantinidis KT. 2014. Nonpareil: A redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* 30:629–635.
167. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2008. BLAST+: architecture and applications. *BMC Bioinformatics* 10.
168. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267.
169. Oksanen J, Guillaume Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymus P, Stevens MHH, Szoecs E, Wagner H. 2017. Community Ecology Package, Package ‘vegan’.
170. McMurdie PJ, Holmes S. 2013. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* 8:e61217.
171. Revelle W. 2017. Procedures for Psychological, Psychometric, and Personality Research. 1.7.8.
172. Chamberlain S, Szoecs E, Foster Z, Boettiger C, Ram K, Bartomeus I, Baumgartner J,

- O'Donnell J, Oksanen J. 2017. Taxonomic information from around the web. 0.9.0. CRAN.
173. Chen H, Boutros PC. 2011. VennDiagram: A package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* 12:35.
  174. Kolde R. 2015. Package 'pheatmap'. 1.0.8.
  175. Rubin BER, Sanders JG, Hampton-Marcell J, Owens SM, Gilbert JA, Moreau CS. 2014. DNA extraction protocols cause differences in 16S rRNA amplicon sequencing efficiency but not in community profile composition or structure. *Microbiologyopen* 3:910–921.
  176. Périchon B, Courvalin P. 2009. VanA-type vancomycin-resistant *Staphylococcus aureus*. *Antimicrob Agents Chemother* 53:4580–4587.
  177. Wannaprasat W, Padungtod P, Chuanchuen R. 2011. Class 1 integrons and virulence genes in *Salmonella enterica* isolates from pork and humans. *Int J Antimicrob Agents* 37:457–461.
  178. Johnson TA, Stedtfeld RD, Wang Q, Cole JR, Hashsham SA, Looft T, Zhu YG, Tiedje JM. 2016. Clusters of antibiotic resistance genes enriched together stay together in swine agriculture. *MBio* 7.
  179. Diehl DL, Lapara TM. 2010. Effect of temperature on the fate of genes encoding TC resistance and the integrase of class 1 integrons within anaerobic and aerobic digesters treating municipal wastewater solids.pdf. *Environ Sci Technol* 44:9128–9133.
  180. Oliveira-Pinto C, Costa PS, Reis MP, Chartone-Souza E, Nascimento AMA. 2016. Diversity of gene cassettes and the abundance of the class 1 integron-integrase gene in sediment polluted by metals. *Extremophiles* 20:283–289.
  181. Goltsman DSA, Denef VJ, Singer SW, VerBerkmoes NC, Lefsrud M, Mueller RS, Dick GJ, Sun CL, Wheeler KE, Zemla A, Baker BJ, Hauser L, Land M, Shah MB, Thelen MP, Hettich RL, Banfield JF. 2009. Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing “*Leptospirillum rubrum*” (Group II) and “*Leptospirillum ferrodiazotrophum*” (Group III) bacteria in acid mine drainage biofilms. *Appl Environ Microbiol* 75:4599–4615.
  182. Glick BR. 1995. The enhancement of plant growth by free-living bacteria. *Can J Microbiol* 41:109–117.

183. Hu J, Wei Z, Friman VP, Gu SH, Wang XF, Eisenhauer N, Yang TJ, Ma J, Shen QR, Xu YC, Jousset A. 2016. Probiotic diversity enhances rhizosphere microbiome function and plant disease suppression. *MBio* 7:1–8.
184. Falkowski PG, Fenchel T, Delong EF. 2008. The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science* (80- ) 320.
185. Smillie C, Garcillan-Barcia MP, Francia M V., Rocha EPC, de la Cruz F. 2010. Mobility of Plasmids. *Microbiol Mol Biol Rev* 74:434–452.
186. Heuer H, Smalla K. 2012. Plasmids foster diversification and adaptation of bacterial populations in soil. *FEMS Microbiol Rev* 36:1083–1104.
187. Sentchilo V, Mayer AP, Guy L, Miyazaki R, Green Tringe S, Barry K, Malfatti S, Goessmann A, Robinson-Rechavi M, van der Meer JR. 2013. Community-wide plasmid gene mobilization and selection. *ISME J* 7:1173–86.
188. Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* 3:711–721.
189. Schloss PD, Girard RA, Martin T, Edwards J, Thrash JC. 2016. Status of the archaeal and bacterial census: An update. *MBio* 7:1–10.
190. Krawczyk PS, Lipinski L, Dziembowski A. 2018. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res* 46:e35.
191. Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E, Mizrahi I, Shamir R. 2016. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* 33:475–482.
192. Burton JN, Liachko I, Dunham MJ, Shendure J. 2014. Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps. *G3Genes|Genomes|Genetics* 4:1339–1346.
193. Choi J, Yang F, Stepanauskas R, Cardenas E, Garoutte A, Williams R, Flater J, Tiedje JM, Hofmockel KS, Gelder B, Howe A. 2017. Strategies to improve reference databases for soil microbiomes. *ISME J* 11:829–834.
194. O'Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova

- O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44:D733–D745.
195. R Core Team. 2017. R: A Language and Environment for Statistical Computing. CRAN, Vienna, Austria.
  196. Hartigan JA, Hartigan PM. 1985. The dip test of unimodality. *Ann Stat* 13:70–84.
  197. Ellison AM. 1987. Effect of Seed Dimorphism on the Density-Dependent Dynamics of Experimental Populations of *Atriplex triangularis* ( *Chenopodiaceae* ). *Am J Bot* 74:1280–1288.
  198. Johnson L, Eddy S, Portugaly E. 2011. Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure. *BMC Bioinformatics* 39:431.
  199. Luo W, Xu Z, Riber L, Hansen LH, Sørensen SJ. 2016. Diverse gene functions in a soil mobilome. *Soil Biol Biochem* 101:175–183.
  200. Shintani M, Sanchez ZK, Kimbara K. 2015. Genomics of microbial plasmids: Classification and identification based on replication and transfer systems and host taxonomy. *Front Microbiol* 6:1–16.
  201. Garcillán-Barcia MP, Alvarado A, De la Cruz F. 2011. Identification of bacterial plasmids based on mobility and plasmid population biology. *FEMS Microbiol Rev* 35:936–956.
  202. Beszteri B, Temperton B, Frickenhaus S, Giovannoni SJ. 2010. Average genome size: a potential source of bias in comparative metagenomics. *ISME J* 4:1075–1077.
  203. Nayfach S, Pollard KS. 2015. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol* 16:51.
  204. Sorensen JW, Dunivin TK, Tobin TC, Shade A. 2019. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. *Nat Microbiol* 4:55–61.

205. Lee C, Kim J, Shin SG, Hwang S. 2006. Absolute and relative QPCR quantification of plasmid copy number in *Escherichia coli*. *J Biotechnol* 123:273–280.
206. Murawska E, Fiedoruk K, Bideshi DK, Swiecicka I. 2013. Complete Genome Sequence of *Bacillus thuringiensis* subsp. *thuringiensis* Strain IS5056, an Isolate Highly Toxic to *Trichoplusia ni*. *Genome Announc* 1:e00108-13-e00108-13.
207. Dunivin TK, Shade A. 2018. Community structure explains antibiotic resistance gene dynamics over a temperature gradient in soil. *FEMS Microbiol Ecol* 94.
208. Sørensen SJ, Bailey M, Hansen LH, Kroer N, Wuertz S, Sorensen SJ, Bailey M, Hansen LH, Kroer N, Wuertz S. 2005. Studying plasmid horizontal transfer in situ: a critical review. *Nat Rev Microbiol* 3:700–710.
209. Sorensen JW, Dunivin TK, Tobin TC, Shade A. 2018. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. *Nat Microbiol* 4:55–61.
210. Dunivin TK, Yeh SS, Shade A. 2018. Targeting microbial arsenic resistance genes: a new bioinformatic toolkit informs arsenic ecology and evolution in soil genomes and metagenomes. *bioRxiv*.
211. Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang XS, Davis-Richardson A, Canepa R, Triplett EW, Faith JJ, Sebra R, Schadt EE, Fang G. 2018. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol* 36:61–69.
212. Jechalke S, Broszat M, Lang F, Siebe C, Smalla K, Grohmann E. 2015. Effects of 100 years wastewater irrigation on resistance genes, class 1 integrons and IncP-1 plasmids in Mexican soil. *Front Microbiol* 6:1–10.
213. Smalla K, Haines AS, Jones K, Krögerrecklenfort E, Heuer H, Schlöter M, Thomas CM. 2006. Increased abundance of IncP-1 $\beta$  plasmids and mercury resistance genes in mercury-polluted river sediments: First discovery of IncP-1 $\beta$  plasmids with a complex mer transposon as the sole accessory element. *Appl Environ Microbiol* 72:7253–7259.
214. Riber L, Burmolle M, Alm M, Milani SM, Thomsen P, Hansen LH, Sorensen SJ. 2016. Enhanced plasmid loss in bacterial populations exposed to the antimicrobial compound irgasan delivered from interpenetrating polymer network silicone hydrogels. *Plasmid* 87–88:72–78.
215. White RA, Callister SJ, Moore RJ, Baker ES, Jansson JK. 2016. The past, present and

- future of microbiome analyses. *Nat Protoc* 11:2049–2053.
216. Stepanauskas R. 2015. Wiretapping into microbial interactions by single cell genomics. *Front Microbiol* 6:2014–2016.
  217. Nelson MB, Martiny AC, Martiny JBH. 2016. Global biogeography of microbial nitrogen-cycling traits in soil. *Proc Natl Acad Sci* 113:8033–8040.
  218. Boon E, Meehan CJ, Whidden C, Wong DHJ, Langille MGI, Beiko RG. 2014. Interactions in the microbiome: Communities of organisms and communities of genes. *FEMS Microbiol Rev* 38:90–118.
  219. Páez-Espino D, Tamames J, De Lorenzo V, Cánovas D. 2009. Microbial responses to environmental arsenic. *BioMetals* 22:117–130.
  220. Watanabe T, Hirano S. 2013. Metabolism of arsenic and its toxicological relevance. *Arch Toxicol* 87:969–979.
  221. Huang H, Jia Y, Sun G-X, Zhu Y-G. 2012. Arsenic Speciation and Volatilization from Flooded Paddy Soils Amended with Different Organic Matters. *Environ Sci Technol* 46:2163–2168.
  222. Mukai H, Ambe Y, Muku T, Takeshita K, Fukuma T. 1986. Seasonal variation of methylarsenic compounds in airborne particulate matter. *Nature* 324:239–241.
  223. Li X, Zhang L, Wang G. 2014. Genomic evidence reveals the extreme diversity and wide distribution of the arsenic-related genes in Burkholderiales. *PLoS One* 9:1–11.
  224. Crognale S, Amalfitano S, Casentini B, Fazi S, Petruccioli M, Rossetti S. 2017. Arsenic-related microorganisms in groundwater: a review on distribution, metabolic activities and potential use in arsenic removal processes. *Rev Environ Sci Biotechnol* 16:647–665.
  225. Plewniak F, Crognale S, Rossetti S, Bertin PN, Marco DE, Pelaez AI. 2018. A Genomic Outlook on Bioremediation : The Case of Arsenic Removal. *Front Microbiol* 9:820.
  226. Dunivin TK, Miller J, Shade A. 2018. Taxonomically-linked growth phenotypes during arsenic stress among arsenic resistant bacteria isolated from soils overlying the Centralia coal seam fire. *PLoS One* 13:e0191893.



227. Chi L, Bian X, Gao B, Tu P, Ru H, Lu K. 2017. The effects of an environmentally relevant level of arsenic on the gut microbiome and its functional metagenome. *Toxicol Sci* 160:1–12.
228. Rascovan N, Javier M, Martín P V, María EF. 2016. Metagenomic study of red biofilms from Diamante Lake reveals ancient arsenic bioenergetics in haloarchaea Metagenomic study of red biofilms from Diamante Lake reveals ancient arsenic bioenergetics in haloarchaea. *Int Soc Microb Ecol* 109:1–11.
229. Cai L, Yu K, Yang Y, Chen BW, Li XD, Zhang T. 2013. Metagenomic exploration reveals high levels of microbial arsenic metabolism genes in activated sludge and coastal sediments. *Appl Microbiol Biotechnol* 97:9579–9588.
230. Babilonia J, Conesa A, Casaburi G, Pereira C, Louyakis AS, Reid RP, Foster JS. 2018. Comparative metagenomics provides insight into the ecosystem functioning of the Shark Bay Stromatolites, Western Australia. *Front Microbiol* 9:1359.
231. The Uniprot Consortium T. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 45.
232. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
233. Dunivin TK, Choi J, Howe A, Shade A. 2019. RefSoil+: a Reference Database for Genes and Traits of Soil Plasmids. *mSystems* 4:e00349-18.
234. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Verezhenska O, Isbandi M, Thomas AD, Ali R, Sharma K, Kyrpides NC, Reddy TBK. 2017. Genomes OnLine Database (GOLD) v.6: Data updates and feature enhancements. *Nucleic Acids Res* 45:D446–D456.
235. Stamatakis A. 2014. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30.
236. Cho J-C, Lee D-H, Cho Y-C, Cho J-C, Kim S-J. 2006. Direct Extraction of DNA from Soil for Amplification of 16S rRNA Gene Sequences by Polymerase Chain Reaction. *J Microbiol* 34:229–235.
237. Wang Q, Fish JA, Gilman M, Sun Y, Brown CT, Tiedje JM, Cole JR. 2015. Xander: employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome* 3.

238. 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. *Proc Natl Acad Sci* 103:2075–2080.
239. Muller D, Lièvreumont D, Simeonova DD, Jean-Claude H, Lett M-C. 2003. Arsenite oxidase aox genes from a metal-resistant beta-proteobacterium. *J Bacteriol* 185:135–141.
240. Saltikov CW, Newman DK. 2003. Genetic identification of a respiratory arsenate reductase. *Proc Natl Acad Sci* 100:10983–8.
241. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10:1–9.
242. Eddy SR. 1996. Hidden Markov models. *Curr Opin Struct Biol* 6:361–365.
243. Mikael Sehlin H, Börje Lindström E. 1992. Oxidation and reduction of arsenic by *Sulfolobus acidocaldarius* strain BC. *FEMS Microbiol Lett* 93:87–92.
244. Blomberg SP, Jr TG, Ives AR. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57:717–745.
245. Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, Walters WA, Knight R, Fierer N. 2011. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biol Biochem* 43:1450–1455.
246. Inskeep WP, Macur RE, Hamamura N, Warelow TP, Ward SA, Santini JM. 2007. Detection, diversity and expression of aerobic bacterial arsenite oxidase genes. *Environ Microbiol* 9:934–943.
247. Mirza BS, Sorensen DL, Dupont RR, McLean JE. 2017. New arsenate reductase gene (arrA) PCR primers for diversity assessment and quantification in environmental samples. *Appl Environ Microbiol* 83:e02725-16.
248. Malasarn D, Saltikov CW, Campbell KM, Santini JM, Hering JG, Newman DK. 2004. arrA Is a Reliable Marker for As(V) Respiration. *Science* (80- ) 306:455.
249. Villegas-Torres MF, Bedoya-Reina OC, Salazar C, Vives-Florez MJ, Dussan J. 2011. Horizontal arsC gene transfer among microorganisms isolated from arsenic polluted soil. *Int Biodeterior Biodegrad* 65:147–152.

250. Wang L, Wang J, Jing C. 2017. Comparative genomic analysis reveals organization, function and evolution of ars genes in *Pantoea* spp. *Front Microbiol* 8.
251. Lee S, Rakic-Martinez M, Graves LM, Ward TJ, Siletzky RM, Kathariou S. 2013. Genetic determinants for cadmium and arsenic resistance among *Listeria monocytogenes* serotype 4B isolates from sporadic human listeriosis patients. *Appl Environ Microbiol* 79:2471–2476.
252. Zhao FJ, Harris E, Yan J, Ma J, Wu L, Liu W, McGrath SP, Zhou J, Zhu YG. 2013. Arsenic methylation in soils and its relationship with microbial arsM abundance and diversity, and As speciation in rice. *Environ Sci Technol* 47:7147–7154.
253. Qiao JT, Li XM, Hu M, Li FB, Young LY, Sun WM, Huang W, Cui JH. 2018. Transcriptional Activity of Arsenic-Reducing Bacteria and Genes Regulated by Lactate and Biochar during Arsenic Transformation in Flooded Paddy Soil. *Environ Sci Technol* 52:61–70.
254. Li R, Chai M, Qiu GY. 2016. Distribution, fraction, and ecological assessment of heavy metals in sediment-plant system in mangrove forest, South China Sea. *PLoS One* 11:1–15.
255. Chatterjee M, Massolo S, Sarkar SK, Bhattacharya AK, Bhattacharya BD, Satpathy KK, Saha S. 2009. An assessment of trace element contamination in intertidal sediment cores of Sunderban mangrove wetland, India for evaluating sediment quality guidelines. *Environ Monit Assess* 150:307–322.
256. Chaudhuri P, Nath B, Birch G. 2014. Accumulation of trace metals in grey mangrove *Avicennia marina* fine nutritive roots: The role of rhizosphere processes. *Mar Pollut Bull* 79:284–292.
257. Fahy A, Giloteaux L, Bertin PN, Le Paslier D, Médigue C, Weissenbach J, Duran R, Lauga B. 2015. 16S rRNA and As-Related Functional Diversity: Contrasting Fingerprints in Arsenic-Rich Sediments from an Acid Mine Drainage. *Microb Ecol* 70:154–167.
258. Newton RJ, Shade A. 2016. Lifestyles of rarity: Understanding heterotrophic strategies to inform the ecology of the microbial rare biosphere. *Aquatic Microbial Ecology*.
259. Pal C, Asiani K, Arya S, Rensing C, Stekel DJ, Larsson DGJ, Hobman JL. 2017. Metal Resistance and Its Association With Antibiotic Resistance. *Advances in Microbial Physiology*, 1st ed. Elsevier Ltd.