Glacial meltwater is associated with gene-specific diversification of metal resistance genes in High Arctic soil microbiomes

Franck Joel Ouedraogo

Thesis submitted to the University of Ottawa in partial fulfillment

of the requirements for the Master of Science degree in Biology, with specialization in Bioinformatics

Department of Biology

Faculty of Science

© Franck Joel Ouedraogo, Ottawa, Canada, 2026

Abstract

Climate warming accelerates the delivery of glacial meltwater to Arctic lakes, mobilizing metals from thawing catchments and reshaping the selective landscape experienced by resident microbes. Whether these gradients leave detectable signatures of diversification in environmental resistance genes remains unclear. We investigated four metal resistance genes (*merA*, *arsC*, *cadA*, and *chrR*) in metagenomic datasets from Lake Hazen (Nunavut, Canada), the largest High Arctic freshwater lake, sampled across a natural hydrological gradient of Control, Low-runoff, and High-runoff regimes. Using a space-for-time substitution design, we combined population-genetic and codon-based approaches to quantify diversity and candidate selection signals, including nucleotide diversity, Tajima's *D*, non-synonymous-to-synonymous diversity ratios, McDonald-Kreitman tests with outgroup-sensitivity analysis, site-level episodic selection (MEME with false-discovery-rate correction), and gene-wide tests (BUSTED and BUSTED-E) and complemented these with ortholog clustering, within-clade re-analysis, taxonomic profiling, rarefaction, and phylogenetic beta-diversity.

Marked heterogeneity emerged among genes. *merA* showed increasing diversity and diversification along the runoff gradient, and these signals were preserved within the largest orthologous cluster (90% of haplotypes), supporting an interpretation of within-orthogroup diversification. *cadA* displayed the strongest McDonald-Kreitman signal under low and high runoff, but its gene-wide BUSTED-E signal collapsed within a single ortholog cluster, suggesting that part of the apparent diversifying signal at the gene-family level reflects inter-subfamily heterogeneity. *chrR* exhibited the strongest regime structure but its largest orthologous cluster was dominated by Control sequences and 93% of High-regime haplotypes were affiliated with a single bacterial order (Hyphomicrobiales), indicating that the regime contrast for this gene reflects compositional turnover rather than within-lineage evolution. *arsC* remained largely consistent with neutral or purifying evolution across regimes.

Because these inferences derive from metagenomic gene pools sampled across only three hydrological regimes and aggregate variants across taxa, we interpret them as exploratory, hypothesis-generating

patterns rather than as demonstrations of population-level adaptation. Our findings highlight environmental resistance genes as candidate indicators of changing biogeochemical conditions in rapidly warming polar ecosystems, while underscoring the importance of orthology and community-composition controls when inferring selection from metagenomic data.

Keywords: Metal resistance genes, Glacial meltwater, Gene-pool diversification, Arctic freshwater microbiome, Metagenomics.

Acknowledgements

I want to express my deepest gratitude to my co-supervisors, Prof. Alexandre J. Poulain and Prof. Stéphane Aris-Brosou, for their continuous guidance, scientific rigour, and patience throughout this thesis. Their complementary expertise in microbial biogeochemistry and computational evolutionary biology shaped both the questions I asked and the methods I used to answer them. The critical feedback I received at every stage of this work has been instrumental in producing a rigorous and defensible thesis.

I extend my warm thanks to the members of the Poulain and Aris-Brosou laboratories at the University of Ottawa. The intellectual community, technical advice, and collaborative spirit of both groups provided the day-to-day support that made this research possible. I am also grateful to Graham Colby, Matti Ruuskanen, and Kyra St. Pierre, whose foundational work on the Lake Hazen system made this study possible, and who have published the metagenomic and biogeochemical data forming the empirical backbone of this thesis.

This research would not have been computationally feasible without access to the Digital Research Alliance of Canada (formerly Compute Canada) infrastructure. I gratefully acknowledge the allocations and technical support provided by the Fir systems.

Financial support during my graduate studies was provided by a graduate scholarship from the University of Ottawa, Faculty of Science, and by Natural Sciences and Engineering Research Council of Canada Discovery Grants held by my supervisors.

Finally, I owe an immense debt of gratitude to my family and friends, whose unwavering support, encouragement, and understanding sustained me through the long hours, late nights, and inevitable setbacks of graduate research. Their belief in me, especially during difficult periods, gave me the energy to persevere. This thesis is as much a product of their patience and love as it is of my own work.

Contents

Abstract ii
Acknowledgements iv
List of Figures vii
List of Tables viii
List of Supplementary Figures ix
List of Supplementary Tables ix
List of Abbreviations x
Chapter 1: Introduction 1
1.1 Climate change and Arctic ecosystems 1
1.2 Glacial meltwater and metal contamination at Lake Hazen 2
1.3 Microbial metal resistance genes 6
1.3.1 Mercury and *merA* 6
1.3.2 Arsenic and *arsC* 7
1.3.3 Cadmium, zinc and *cadA* 7
1.3.4 Chromium and *chrR* 8
1.4 Evolutionary forces shaping microbial gene pools 8
1.5 Population genetic and selection inference methods 9
1.5.1 Nucleotide diversity and Tajima's *D* 9
1.5.2 Coalescent inference of gene-pool diversity 10
1.5.3 Detecting selection: π_N/π_S, McDonald-Kreitman, MEME, BUSTED-E 10
1.5.4 Orthology and taxonomy controls 11
1.6 Objectives and hypotheses 12
Chapter 2: Materials and Methods 14
2.1 Site and metagenomic data 14
2.2 Read quality control and preprocessing 14
2.3 Sequence recovery, quality control, and ortholog clustering 15
2.4 Phylogenetic inference and diversity statistics 16
2.5 N_e independent selection tests 16
2.6 Codon-based selection inference and multiple-testing correction 17
2.7 Taxonomic profiling of MRG-containing haplotypes 17
2.8 Gene-pool diversity and exchange (Supplementary Material) 18
2.9 Statistical analysis of metal-diversity relationships and data availability 18
Chapter 3: Results 20
3.1 Haplotype recovery and overall diversity patterns 20

3.2 Differential selective constraint independent of N_e (π_N/π_S) 25
3.3 Non-synonymous versus synonymous divergence from the McDonald-Kreitman test 26
3.4 Codon-level and gene-wide diversification signals (MEME, BUSTED, BUSTED-E) 28
3.5 Phylogenetic diversity and beta-diversity (Faith's PD and PERMANOVA) 31
3.6 Correlation between nucleotide diversity and metal concentrations 32
3.7 Gene-family versus within-clade re-analysis 33
3.8 Taxonomic composition of MRG-containing haplotypes 36
Chapter 4: Discussion 39
4.1 Convergent evidence for gene- and regime-specific evolutionary patterns 39
4.2 First pattern: sustained diversification along the runoff gradient (*merA*) 40
4.3 Second pattern: regime-specific signals partly attributable to subfamily heterogeneity (*cadA* and *chrR*) 42
4.3.1 *cadA*: signals dissociate between gene-family-level and within-clade analyses 42
4.3.2 *chrR*: regime contrast dominated by compositional turnover 43
4.4 Third pattern: near-absence of detectable diversifying signal (*arsC*) 45
4.5 Methodological considerations: reframing coalescent inference and the role of horizontal gene transfer 46
4.6 Implications for Arctic biomonitoring under climate change 49
4.7 Limitations and replication considerations 50
4.8 Future directions 51
Chapter 5: Conclusion 53
Statement on the Use of Generative AI 56
References 57
Annex A - Supplementary Tables 62
Annex B - Supplementary Figures 78

List of Figures

Figure 1 Lake Hazen sampling sites and watershed hydrology. The map shows the three hydrological regimes: Control (white dashed ellipse; CSED sediment and CSOIL soil sites along the Ruggles River), Low Runoff (green dashed ellipse; LR1, LR2 and LRSOIL sites along Blister Creek), and High Runoff (orange dashed ellipse; HR1, HR2 and HRSOIL sites at the Abbé and Snowgoose river outlets). Inset: location of Lake Hazen on Ellesmere Island, Nunavut, Canada (~82°N). Adapted from Colby et al. (2020). 4
Figure 2. Metal concentrations across the Lake Hazen runoff gradient (St. Pierre et al., 2019)**.** Site-level points (sediment, soil, and additional water-column sites) are shown by symbol, and the regime mean by a horizontal bar. Wong palette: Control = sky blue, Low = bluish green, High = orange. Concentrations are in ng/L for Hg and µg/L for As, Zn, and Cr. 5
Figure 3. Nucleotide diversity (π, Nei pairwise method) per gene and regime, computed on soil-only haplotypes. Bars: regime mean; numeric labels: regime mean value. Wong palette: Control = sky blue, Low = bluish green, High = orange. Numerical values are reported in Table 1 (main text) and Table S1 (site-level breakdown). 24
Figure 4. Tajima's D statistics for soil-only haplotypes per gene and regime. Bars: point estimates. Negative D values are consistent with population expansion or purifying selection; positive values may indicate balancing selection or population contraction. Significance from the beta-approximated null distribution: * $p < 0.05$. Numerical values are reported in Table S2. 25
Figure 5. Ratio of nucleotide diversity at non-synonymous (π_N) to synonymous (π_S) sites (Nei-Gojobori, 1986), per gene and regime. Error bars: 95% bootstrap CI from 500 resamples. Dashed line: neutral expectation ($\pi_N/\pi_S = 1$); ratios below the dashed line indicate purifying selection. Note the reinforced constraint for cadA in High (0.61) and the minimum for chrR in Low (0.45). Numerical values are reported in Table S4. 26
Figure 6. McDonald-Kreitman α (Smith & Eyre-Walker, 2002), the proportion of adaptive substitutions, computed per gene and regime against the reciprocal best BLAST outgroup at E-value < 10^{-20}. Significance from χ^2 test with Yates correction: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant. Numerical values are reported in Table S5. 27
Figure 7. McDonald-Kreitman α across outgroups (MACSE codon alignment, 33 tests). Each point represents one outgroup-by-regime combination; the marker indicates outgroup class (circle = canonical, square = close ~65–72% AA identity, triangle = intermediate ~55–65%, diamond = distant ~40–55%). Horizontal bars indicate the median across outgroups for each regime. Numerical values are reported in Table S6. 28
Figure 8. Percentage of codon sites under episodic diversifying selection (MEME, Murrell et al., 2012) per gene under three conditions: raw $p < 0.05$ at the gene-family level, Benjamini-Hochberg FDR $q < 0.05$ at the gene-family level, and FDR $q < 0.05$ within the largest MMseqs2 ortholog cluster (within-clade). All values are pooled across regimes per gene. Numerical values are reported in Table S7. 30
Figure 9. Standard BUSTED, BUSTED-E (gene-family), and BUSTED-E (within-clade) likelihood-ratio statistics (LRT, log scale) per gene. The ns annotation indicates non-significance of the BUSTED-E within-clade test for cadA (LRT = 0.73, p = 0.347). The error-sink rate class of BUSTED-E absorbed less than 1% of the site weight for all four genes. All values are pooled across regimes per gene. 31
Figure 10. Relationship between nucleotide diversity (π) and the target metal concentration for each gene across the three hydrological regimes: merA vs Hg, arsC vs As, cadA vs Zn, chrR vs Cr. Each point represents one regime (C = Control, L = Low, H = High). Horizontal error bars: metal SD across replicate sites; vertical error bars: π bootstrap 95% CI. Spearman ρ is shown for each gene; p-values are not interpretable at n = 3. 33

List of Tables

Table 1. Summary statistics for the four metal resistance genes across hydrological regimes at Lake Hazen. N = number of unique soil haplotypes. π reported as mean [95% bootstrap CI]. Tajima's D significance: * $p < 0.05$; ns not significant. π_N/π_S = Nei-Gojobori ratio with 95% bootstrap CI. α = McDonald-Kreitman proportion of adaptive substitutions; significance from χ^2 test with Yates correction (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns not significant). 21
Table 2. MMseqs2 ortholog clustering at the data-driven elbow threshold per gene, and selection statistics within the largest ortholog cluster (within-clade re-analysis). α(MK) within-clade C/L/H reports the McDonald-Kreitman proportion of adaptive substitutions per regime computed within the largest cluster. n = 0 indicates that the largest cluster contains no haplotypes from that regime. BUSTED-E LRT and MEME % sites (FDR $q < 0.05$) are gene-wide statistics pooling all regimes within the largest cluster. 36

List of Supplementary Figures

Figure S1. Analysis workflow diagram.

Figure S2. Tajima's *D* per gene and regime.

Figure S3. Non-synonymous to synonymous nucleotide diversity ratio (π_N/π_S) per gene and regime.

Figure S4. Faith's phylogenetic diversity (PD) per gene and regime.

Figure S5. PERMANOVA R^2 per gene.

Figure S6. MMseqs2 ortholog cluster summary per gene.

Figure S7. Gene-family vs within-clade comparison across four metrics.

Figure S8. Taxonomic composition of MRG-containing haplotypes per gene and regime.

List of Supplementary Tables

Table S1. Nucleotide diversity (π) per gene and individual sampling site.

Table S2. Tajima's *D* statistics per gene and regime.

Table S3. Rarefaction/subsampling of π to a common haplotype depth.

Table S4. Nei-Gojobori π_N and π_S diversity per gene and regime.

Table S5. McDonald-Kreitman test detail per gene and regime.

Table S6. Outgroup sensitivity of the McDonald-Kreitman α using MACSE.

Table S7. Detailed MEME results per gene at gene-family level.

Table S8. Faith's phylogenetic diversity (PD) per gene and regime.

Table S9. PERMANOVA on patristic distance matrices.

Table S10. Metal concentrations across hydrological regimes.

Table S11. MMseqs2 ortholog clustering at multiple sequence-identity thresholds.

Table S12. Within-clade selection statistics.

Table S13. Taxonomic composition of MRG-containing haplotypes per gene and regime.

Table S14. Bayesian coalescent estimates of the gene-pool diversity index $\theta = 2N_e\mu$ from BEAST2.

Table S15. Inter-regime backward migration rates estimated by MASCOT.

List of Abbreviations

α	Proportion of adaptive substitutions (Smith-Eyre-Walker)
$\beta+$	Non-synonymous substitution rate in MEME positive class
ω	Ratio of non-synonymous to synonymous substitution rates (d_N/d_S)
π	Nucleotide diversity
π_N	Nucleotide diversity at non-synonymous sites
π_S	Nucleotide diversity at synonymous sites
θ	Population mutation parameter (2Neμ for haploids)
arsC	Arsenate reductase gene
BEAST2	Bayesian Evolutionary Analysis by Sampling Trees, version 2
BLAST	Basic Local Alignment Search Tool
BLASTX	BLAST translated nucleotide query versus protein database
BUSTED	Branch-site Unrestricted Statistical Test for Episodic Diversification
BUSTED-E	BUSTED with Error Sink
cadA	Cadmium and zinc-transporting P-type ATPase gene
CDS	Coding sequence
chrR	Chromate reductase gene
CI	Confidence interval
DIAMOND	Double Index Alignment of Next-generation sequencing Data
ESS	Effective sample size (MCMC convergence diagnostic)
FDR	False Discovery Rate (Benjamini-Hochberg correction)
FMN	Flavin mononucleotide
HGT	Horizontal gene transfer
HMM	Hidden Markov model
HPD	95% highest posterior density interval
IDH	Intermediate Disturbance Hypothesis
IQ-TREE	Maximum-likelihood phylogenetic software
LRT	Likelihood ratio test
MACSE	Multiple Alignment of Coding SEquences
MAFFT	Multiple Alignment using Fast Fourier Transform
MASCOT (Table S6)	Marginal Approximation of the Structured Coalescent

MCMC	Markov chain Monte Carlo
MEME	Mixed-Effects Model of Evolution
merA	Mercuric reductase gene
MK	McDonald-Kreitman (test)
MMseqs2	Many-against-Many sequence searching (version 2)
MRG	Metal resistance gene
NCBI	National Center for Biotechnology Information
N_e	Effective population size
NI	Neutrality index
PCoA	Principal Coordinates Analysis
PD	Faith's phylogenetic diversity
PERMANOVA	Permutational multivariate analysis of variance
RBH	Reciprocal best BLAST hit
SRA	Sequence Read Archive
Hg	Mercury (total mercury reported as Hg throughout)
UniProt	Universal Protein Resource (database)

Chapter 1: Introduction

A version of this thesis has been published as: Ouedraogo, F. J., Poulain, A. J., & Aris-Brosou, S. (2026). Glacial meltwater is associated with gene-specific diversification of metal resistance genes in High Arctic soil microbiomes. Frontiers in Microbiology, 17, 1903619. doi: 10.3389/fmicb.2026.1903619

1.1 Climate change and Arctic ecosystems

The Arctic is warming nearly four times faster than the global average, a phenomenon known as Arctic amplification (Rantanen et al., 2022). This accelerated warming is reshaping high-latitude ecosystems in unprecedented ways: glaciers are retreating, permafrost is thawing, snow and ice cover are diminishing, and the hydrological regimes of Arctic lakes and rivers are being fundamentally reorganized (IPCC, 2021; Box et al., 2019). These physical changes cascade into biogeochemical ones, mobilizing elements that have been sequestered in frozen catchments for millennia.

Among the most consequential of these mobilized elements are heavy metals, whose transport from glacial and permafrost sources to downstream aquatic ecosystems has intensified markedly over the past two decades (Schuster et al., 2018; Hawkings et al., 2014; St. Pierre et al., 2019). Permafrost is now recognized as a globally significant reservoir of mercury (Hg), holding an estimated 1.6 million tonnes in the Northern Hemisphere alone (Schuster et al., 2018), and the cryosphere contains substantial inventories of chromium (Cr), arsenic (As), zinc (Zn), lead, and other ecologically relevant metals. As these reservoirs destabilize, the receiving freshwater systems are exposed to metal fluxes whose magnitude and timing have no analogue in the Holocene (Hawkings et al., 2014).

The biological consequences of this rapid mobilization remain poorly understood, particularly with respect to microbial gene pool evolution. Microbial communities mediate the biogeochemical cycling of metals through detoxification, redox transformation, and active efflux, and their resistance machinery is encoded by a well-characterized set of genes that are widespread across the bacterial tree of life (Bruins et al., 2000; Nies, 2003; Silver & Phung, 1996). Whether the recent acceleration of metal delivery to Arctic ecosystems is altering the standing diversity of these genes, the strength and direction of selection acting

on them, or both, remains an open empirical question. Resolving it is important because the genetic capacity of microbial communities to detoxify mobilized metals will, in part, shape the fate of those metals as they cycle through Arctic food webs.

Lake Hazen, located on Ellesmere Island in the Canadian High Arctic (~82°N), is the largest freshwater lake by volume in the world above 75°N and has served for two decades as a sentinel site for monitoring climate-driven changes in Arctic freshwater systems (Lehnherr et al., 2018; Bégin et al., 2021). Its watershed encompasses several glacier-fed tributaries that vary substantially in their meltwater contribution, creating a natural gradient of glacial input within a single basin. This setting provides a rare opportunity to examine how variable glacial runoff and the metal fluxes it carries shape the structure and evolution of resident microbial communities, while controlling for confounding factors that would arise when comparing distinct watersheds (Cuthbertson et al., 2022; St. Pierre et al., 2019). (Figure 1)

1.2 Glacial meltwater and metal contamination at Lake Hazen

Glacial meltwater carries dissolved and particulate metals derived from bedrock weathering, atmospheric deposition, and the release of long-stored contaminants from snow and ice (Hawkings et al., 2014; Stibal et al., 2020). At Lake Hazen, three hydrological regimes can be distinguished among the watershed tributaries. The Ruggles River and its associated soil and sediment sites, hereafter referred to as the Control regime, receive negligible direct glacial input. Blister Creek and associated sites, the Low runoff regime, are characterized by moderate glacial contribution. The Abbé and Snowgoose River systems, with their high runoff regime, are dominated by glacial meltwater discharge during the summer melt season (St. Pierre et al., 2019; Lehnherr et al., 2018).

These three regimes show striking contrasts in metal chemistry. Chromium concentrations (Figure 2) increase monotonically with runoff intensity, spanning a sixty-fold range from 0.33 µg/L in the Ruggles outflow to 19.73 ± 2.24 µg/L in the Abbé-Snowgoose system. Zinc follows a similar monotonic pattern, rising from 9.21 µg/L in Control to 108 ± 12.4 µg/L in High. Total mercury, by contrast, follows a non-

monotonic pattern: it reaches a maximum in the Low runoff regime (Blister Creek; 17.56 ± 14.14 ng/L), with intermediate concentrations in the High runoff regime (7.07 ± 1.27 ng/L) and the lowest values in the Control regime (0.42 ng/L). This discrepancy between mercury concentration and discharge volume is important for interpreting downstream evolutionary signatures: when discharge is considered, the cumulative mercury flux delivered to High-regime sites can exceed that delivered to Low-regime sites despite the lower instantaneous concentration. Arsenic shows only a weak gradient (Control 0.66 μg/L; Low 1.39 ± 0.87; High 1.06 ± 0.21), with a non-monotonic pattern broadly like mercury but of smaller amplitude.

These gradients are biologically meaningful. Metals are simultaneously essential micronutrients, with Zn acting as a cofactor for more than 300 enzymes and Cu and Fe being central to cellular respiration, and potent toxins at elevated concentrations (Nies, 1999; Bruins et al., 2000). The capacity of microbial communities to thrive across the runoff gradient at Lake Hazen depends on a suite of metal-homeostasis and detoxification mechanisms collectively encoded by metal-resistance genes. The natural distribution of these regimes within a single watershed provides a space-for-time substitution framework in which contemporary spatial contrasts serve as proxies for the temporal trajectory expected under continued glacial retreat. Lake Hazen, therefore, offers a unique opportunity to test whether and how environmental metal exposure leaves detectable signatures on the evolutionary dynamics of microbial resistance gene pools.



Figure 1 Lake Hazen sampling sites and watershed hydrology. The map shows the three hydrological regimes: Control (white dashed ellipse; CSED sediment and CSOIL soil sites along the Ruggles River), Low Runoff (green dashed ellipse; LR1, LR2 and LRSOIL sites along Blister Creek), and High Runoff (orange dashed ellipse; HR1, HR2 and HRSOIL sites at the Abbé and Snowgoose river outlets). Inset: location of Lake Hazen on Ellesmere Island, Nunavut, Canada (~82°N). Adapted from Colby et al. (2020).

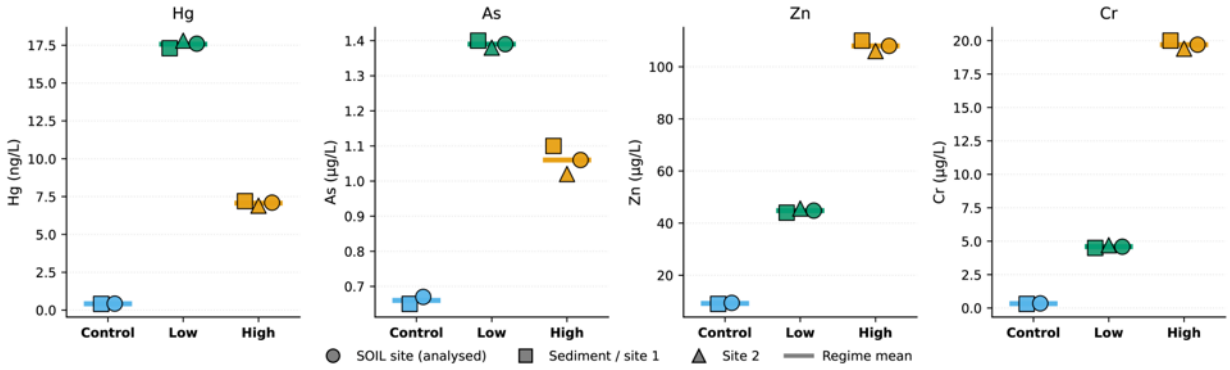


Figure 2. Metal concentrations across the Lake Hazen runoff gradient (St. Pierre et al., 2019)**.** Site-level points (sediment, soil, and additional water-column sites) are shown by symbol, and the regime mean by a horizontal bar. Wong palette: Control = sky blue, Low = bluish green, High = orange. Concentrations are in ng/L for Hg and µg/L for As, Zn, and Cr.

1.3 Microbial metal resistance genes

Bacteria have evolved a remarkable diversity of strategies to cope with metal stress: extracellular sequestration by biofilms and exopolysaccharides, intracellular binding by metallothionein and glutathione, enzymatic transformation to less toxic species, and active efflux across the cellular envelope (Silver & Phung, 1996; Nies, 2003). The genes encoding these mechanisms are widespread in prokaryotic genomes and are often clustered in operons or located on mobile genetic elements such as plasmids, transposons, and integrons, which facilitate their dissemination by horizontal gene transfer (Liebert et al., 1999; Hemme et al., 2010). Metal resistance genes are therefore both ecologically significant, because they mediate the biogeochemical cycling of metals, and evolutionarily informative, because their distribution and diversity respond rapidly to environmental selective pressures.

Four MRGs were selected for the present study, on the basis of three criteria: each gene targets a distinct metal or metal class, the corresponding metal shows a substantial concentration gradient across the Lake Hazen hydrological regimes, and the enzymatic mechanism is sufficiently well characterized that diagnostic catalytic motifs can be used to enforce functional specificity at the sequence-screening step. The four genes are introduced below in turn.

1.3.1 Mercury and *merA*

Mercury is among the most toxic elements in aquatic ecosystems, with its methylated form bioaccumulating in food webs to concentrations that endanger top predators, including humans (Driscoll et al., 2013). The *merA* gene encodes mercuric reductase, a flavoenzyme that reduces ionic Hg(II) to volatile elemental Hg(0), which then diffuses passively out of the cell (Fox & Walsh, 1983; Barkay et al., 2003). The enzyme contains a characteristic Cys-XXXX-Cys (CXXXXC) redox-active disulfide motif at the dimer interface and a C-terminal vicinal cysteine pair (CC) that binds the Hg^{2+} substrate during electron transfer. The *merA* gene is typically encoded on the Tn21 transposon family or on conjugative plasmids and is therefore highly mobile across bacterial taxa, with documented horizontal transfer between proteobacterial, actinobacterial and firmicute hosts (Liebert et al., 1999). This mobility, combined with the toxicity of

mercury and the wide phylogenetic distribution of the mer operon, makes *merA* a particularly informative marker for studying gene-pool diversification in microbial communities in response to mercury exposure.

1.3.2 Arsenic and *arsC*

Arsenic occurs naturally in aquatic systems primarily as inorganic As(V) (arsenate) and As(III) (arsenite), with As(III) being substantially more toxic and mobile than As(V). The *arsC* gene encodes an arsenate reductase that catalyzes the reduction of As(V) to As(III). This is an apparently counterintuitive reaction, given that As(III) is the more toxic of the two species, but it is in fact the first step in arsenic efflux: the reduced As(III) is subsequently exported by an *ArsB* or *ACR3* permease and excluded from the cell (Mukhopadhyay et al., 2002). The *arsC* gene has both chromosomal and plasmid-encoded variants and is a near-universal feature of bacterial genomes, reflecting the ancient and pervasive presence of arsenic in geochemical environments (Yang et al., 2012). The glutaredoxin-dependent *ArsC* subfamily, which dominates in the metagenomic datasets analyzed here, has a catalytic mechanism that involves a conserved N-terminal cysteine forming a covalent intermediate with arsenate, together with a central cysteine that participates in intramolecular disulfide formation (Liu et al., 1995). These catalytic residues provide a stringent motif-based filter for identifying functional *arsC* sequences in metagenomic data.

1.3.3 Cadmium, zinc and *cadA*

The *cadA* gene encodes a P-type ATPase that exports divalent cations, primarily Zn^{2+} and Cd^{2+}, but also Co^{2+} and Pb^{2+}, from the cytoplasm using ATP hydrolysis as the energy source (Nies, 2003; Argüello et al., 2007). The CadA proteins are members of the broader PIB-type heavy-metal ATPase superfamily and contain a signature DKTGT phosphorylation motif, eight transmembrane helices, and one or more CPC or CPX (Cys-Pro-Cys or Cys-Pro-X) metal-binding motifs in the N-terminal cytoplasmic domain (Argüello et al., 2007; Gaballa & Helmann, 2003). The gene is frequently plasmid-borne in Firmicutes such as *Bacillus*, *Staphylococcus*, and *Listeria*, but chromosomal copies are also common in Proteobacteria and other lineages (Nies, 2003). Because *cadA* targets two metals that show substantially different concentration gradients at Lake Hazen, zinc rising sixty-fold across the runoff gradient and cadmium (Cd) being a trace

constituent throughout, *cadA* provides a useful comparator for the gene-specific question of whether selection scales with the magnitude of the local exposure gradient. In addition, *cadA*-family P-type ATPases show substantial sequence-level heterogeneity across bacterial lineages, which motivates the explicit orthology controls described in the Methods.

1.3.4 Chromium and *chrR*

Chromate (CrO_4^{2-}) is a hexavalent chromium oxyanion that is highly soluble, carcinogenic, and an inhibitor of multiple cellular enzymes. Microbial reduction of Cr(VI) to the less toxic and less mobile Cr(III) form is mediated by several enzyme families, including the FMN-dependent chromate reductase encoded by *chrR* (Park et al., 2000; Ackerley et al., 2004). The enzyme uses NAD(P)H as an electron donor, and FMN as a redox cofactor, and the active site contains a conserved N-terminal block, a central EYNR motif, and an FMN-binding serine residue near position 116 of the *Escherichia coli ChrR* reference protein (Ackerley et al., 2004). Although *chrR* homologues are widespread across bacterial phyla, the gene is most often chromosomally located within genomic islands rather than on self-transmissible plasmids, which is expected to limit its inter-host mobility relative to *merA* and *cadA*.

1.4 Evolutionary forces shaping microbial gene pools

The distribution and diversity of MRG variants in any environment are shaped by the interplay of five fundamental evolutionary forces (Hartl & Clark, 2007). Mutation introduces novel variants at a per-site per-generation rate μ that depends on the fidelity of DNA polymerases, the efficiency of repair systems, and environmental stressors. Genetic drift causes random fluctuations in allele frequencies that are inversely proportional to the effective population size N_e: drift dominates in small populations and erases neutral variation at a rate of approximately $1/(2N_e)$ per generation. Natural selection biases allele frequencies according to fitness differentials and can be purifying (eliminating deleterious mutations), positive (favouring beneficial mutations), or balancing (maintaining polymorphism at intermediate frequencies). Gene flow, either through cell migration or horizontal gene transfer, homogenizes allele frequencies among populations and can introduce novel variants from distant donor lineages. Recombination, finally, shuffles

existing variation across loci and breaks the genetic linkage that would otherwise allow selection to act on entire haplotypes.

In prokaryotes, two of these forces are particularly influential and complicate phylogenetic inference. First, horizontal gene transfer (HGT) is pervasive among environmental bacteria, especially for accessory genes such as MRGs that confer context-dependent fitness advantages (Ochman et al., 2000; Soucy et al., 2015). HGT can transfer entire genes between distantly related taxa, decoupling gene phylogenies from organismal phylogenies and violating the assumption of vertical inheritance that underlies coalescent and tree-based methods. Second, selection on MRGs is generally strong because metal exposure imposes substantial fitness costs on unprotected cells and because the encoded enzymes act at biochemical interfaces where amino-acid substitutions can have large phenotypic effects (Bruins et al., 2000). The combination of pervasive HGT and strong selection means that the evolutionary dynamics of MRG gene pools are best understood in terms of circulating haplotypes rather than co-evolving organismal lineages. This conceptual framing motivates the methodological choices made throughout this thesis, in which we report all population-genetic and codon-based statistics as descriptive summaries of standing variation in environmental gene pools rather than as formal tests of adaptation, and complement single-gene analyses with orthology and taxonomic controls to distinguish within-orthogroup evolution from inter-subfamily heterogeneity and community-composition shifts.

1.5 Population genetic and selection inference methods

1.5.1 Nucleotide diversity and Tajima's *D*

The most fundamental population-genetic statistic is nucleotide diversity π, defined as the average number of nucleotide differences per site between all pairs of sequences in a sample (Nei, 1987). Under neutrality and at mutation-drift equilibrium, π equals the population mutation parameter $\theta = 2N_e\mu$ for haploid organisms. Tajima's *D* (Tajima, 1989) is a derived statistic that detects departures from neutrality by comparing two estimators of θ: one based on π and the other based on the number of segregating sites. Significant $D \neq 0$ can result from population expansion or contraction, selective sweeps, balancing

selection, or population structure, and disentangling these alternatives generally requires additional evidence. Because π alone conflates the influences of mutation rate, population size, selection, and demographic history, its interpretation is rarely conclusive without complementary statistics.

1.5.2 Coalescent inference of gene-pool diversity

Coalescent theory provides a statistical framework for estimating evolutionary parameters from sequence data by modelling the genealogy of sampled lineages backward in time (Kingman, 1982; Hudson, 1991). Bayesian implementations such as BEAST2 (Bouckaert et al., 2019) estimate the effective population size scaled by the mutation rate (the product $N_e \times \mu$, equivalent to θ for haploid samples), while structured coalescent extensions such as MASCOT (Müller et al., 2018) additionally estimate inter-population migration rates. These methods are highly informative but rest on critical assumptions: neutrality of the sites used for inference, absence of recombination within loci, and vertical inheritance among the sampled lineages. When applied to metagenomic data from MRGs, all three assumptions are potentially violated. For this reason, we interpret coalescent outputs in this thesis as descriptive indices of gene-pool diversity, that is, the standing variation of circulating haplotypes, rather than as estimates of the census size or reproductive dynamics of any bacterial population and relegate them to the Supplementary Material rather than presenting them as central results.

1.5.3 Detecting selection: π_N/π_S, McDonald-Kreitman, MEME, BUSTED-E

Direct tests for selection circumvent the assumptions of coalescent demographic inference. The ratio of non-synonymous to synonymous nucleotide diversity (π_N/π_S), computed using the Nei-Gojobori method (Nei & Gojobori, 1986), provides a regime-specific measure of selective constraint: ratios below unity indicate purifying selection, while values approaching or exceeding unity indicate relaxed constraint or positive selection. Critically, the ratio is independent of N_e and μ because both quantities cancel from the numerator and denominator (Charlesworth, 2009).

The McDonald-Kreitman test (McDonald & Kreitman, 1991) compares the ratio of non-synonymous to synonymous polymorphisms within a population (P_N/P_S) to the ratio of non-synonymous to synonymous fixed differences between the focal population and a divergent outgroup (D_N/D_S). Under strict neutrality, the two ratios are expected to be equal; a significant excess of D_N/D_S over P_N/P_S indicates patterns consistent with diversification. The proportion of adaptive substitutions can be quantified as $\alpha = 1 - \mathrm{NI}$, where NI is the neutrality index $(P_N/P_S)/(D_N/D_S)$ (Smith & Eyre-Walker, 2002). The MK test is independent of N_e estimation and is particularly sensitive to recurrent positive selection that has driven amino-acid replacements to fixation. In this thesis, we complement the canonical MK test with an outgroup-sensitivity analysis using MACSE codon alignment (Ranwez et al., 2018) and multiple BLAST-derived outgroups in distinct amino-acid identity bins, to test whether the sign and significance of α are preserved across outgroup choices.

MEME (Mixed-Effects Model of Evolution; Murrell et al., 2012) and BUSTED-E (BUSTED with Error sink; Selberg et al., 2025) are codon-based selection tests implemented in HyPhy (Kosakovsky Pond et al., 2020). MEME tests each codon for episodes of positive selection along a phylogeny, allowing $\omega = \beta^+/\alpha$ to vary across both sites and branches. BUSTED-E provides a gene-wide test for evidence of positive selection across the alignment and includes an explicit error-sink rate category that absorbs spurious signals arising from sequencing errors and alignment artifacts. To further mitigate false positives at inflated ω sites, MEME per-codon p-values are corrected using the Benjamini-Hochberg false-discovery-rate (FDR) procedure (Benjamini & Hochberg, 1995), and standard BUSTED is run alongside BUSTED-E as a sanity control to quantify the fraction of the signal attributable to alignment heterogeneity.

1.5.4 Orthology and taxonomy controls

A central methodological challenge in inferring evolution from MRG metagenomic data is the well-known heterogeneity of resistance gene families. Genes such as *arsC*, *cadA* and *chrR* comprise multiple subfamilies with overlapping substrate specificities but distinct sequence signatures, and applying gene-family-level statistics to a mixture of orthologous subfamilies can produce apparent signals that reflect

inter-subfamily divergence rather than within-orthogroup evolution. Two complementary controls are essential. First, orthology clustering, using MMseqs2 (Steinegger & Söding, 2017) at a data-driven elbow threshold determined from the per-gene identity distribution, groups haplotypes into approximately orthologous subfamilies, and selection analyses can then be repeated within the largest cluster to assess whether qualitative patterns persist when ortholog heterogeneity is controlled. Second, best-hit taxonomic profiling of MRG-containing haplotypes, aggregated to the order level for cross-gene comparability, distinguishes shifts in community composition (different orders dominating in different regimes) from within-lineage evolution. Together, these two controls allow the interpretation of gene-family-level statistics to be qualified by explicit assessment of which patterns are robust at the within-orthogroup level and which depend on inter-subfamily heterogeneity or community-composition turnover. Both controls are applied throughout this thesis and are central to the conclusions of Chapter 3 (Results).

1.6 Objectives and hypotheses

The central objective of this thesis is to characterize the evolutionary signatures imprinted on metal resistance gene pools by the natural gradient of glacial-meltwater-derived metal exposure at Lake Hazen. Using metagenomic DNA data from soil sites across three hydrological regimes (Control, Low and High), we examine four MRGs (*merA*, *arsC*, *cadA*, *chrR*) using a combination of descriptive population-genetic statistics, orthology and taxonomy controls, and multiple selection tests with non-overlapping assumptions.

Three hypotheses are tested. (H1) Nucleotide diversity (π) of metal resistance genes may increase along the gradient of glacial meltwater input, consistent with greater metal mobilization co-occurring with the diversification of resistant variants. (H2) Selective signals on MRGs may vary with environmental metal exposure, with genes responding to enriched metals showing patterns consistent with diversifying selection that scale with concentration, whereas genes targeting metals with weaker or non-monotonic gradients may show weaker or non-monotonic selection signals. (H3) Independent methods with non-overlapping assumptions may provide convergent evidence of such patterns, allowing biological interpretation despite the methodological limitations of any single approach applied to metagenomic data. Because the analysis

is based on three regimes sampled once within a single watershed, we treat the resulting patterns as exploratory, hypothesis-generating observations rather than as formal demonstrations of population-level adaptation.

Chapter 2: Materials and Methods

2.1 Site and metagenomic data

All analyses were conducted on sequencing data obtained from a published Lake Hazen metagenomic DNA survey deposited in the NCBI Sequence Read Archive under BioProject PRJNA556841 (Colby et al., 2020). We restricted analyses to soil samples to focus on terrestrial microbial responses and to maintain internal comparability across the eight soil sites: CSED and CSOIL in the Control regime (Ruggles River outflow, no direct glacial input); LR1, LR2 and LRSOIL in the Low runoff regime (Blister Creek catchment, moderate glacial contribution); and HR1, HR2 and HRSOIL in the High runoff regime (Abbé and Snowgoose River catchments, dominated by glacial meltwater). The original site labels published by Colby et al. (2020) have been renamed throughout this thesis to reflect the hydrological regime of each site (LR for Low runoff, HR for High runoff), so that site identifiers are mnemonic with respect to the experimental design. Reference protein sequences for *merA*, *arsC*, *cadA*, and *chrR* were retrieved from NCBI and UniProt.

It should be noted that the analytical design does not include biological replicates within each regime, as each regime is represented by a single soil site. The unit of statistical inference is therefore the gene haplotype rather than the sampling site. With 60 to 1,349 unique haplotypes recovered per gene per regime (Table 1), the descriptive population-level metrics computed in this thesis (π, Tajima's D, π_N/π_S, MEME with FDR correction, standard BUSTED and BUSTED-E, and MMseqs2 clustering) achieve substantial statistical power despite single-site sampling per regime. Because the analysis aggregates variants across taxa and across three regimes in a single watershed, we present the resulting patterns as descriptive, hypothesis-generating observations rather than as formal tests of adaptation. The implications of this design for inference are addressed explicitly in the Discussion.

2.2 Read quality control and preprocessing

Raw paired-end reads (2 × 125 bp, Illumina HiSeq 2500) were first inspected with FastQC v0.11.9 (Andrews, 2010) to assess per-base quality scores, per-sequence quality, GC content, adapter content and

overrepresented sequences. Reads were then processed with Trimmomatic v0.39 (Bolger et al., 2014) in paired-end mode, using the settings LEADING:34, SLIDINGWINDOW:4:15 and MINLEN:70 to remove low-quality bases and adapter sequences while retaining read pairs whose surviving length exceeded 70 bp. A second FastQC pass on the trimmed reads confirmed adapter removal, improved per-base quality (Phred ≥ 33 across the read length) and a minimum retained length of 90 bp.

2.3 Sequence recovery, quality control, and ortholog clustering

Putative MRG-containing reads were identified by translated BLASTX searches against curated reference protein sets using DIAMOND v2.1.22 (Buchfink et al., 2015; E-value ≤ 10^{-5}, one hit per read, -k 1). Paired-end sequences containing at least one matching mate were retrieved from the trimmed FASTQ files using seqkit v2.5.1 (Shen et al., 2016). Specificity was enforced downstream through function-based validation: translated candidate reads were filtered against gene-specific conserved catalytic motifs (CXXC and CPx motifs for cadA (Argüello et al., 2007); the *merA* active-site Cys residues (Fox & Walsh, 1983); the *arsC* redox-active CX_5R motif (Mukhopadhyay et al., 2002); and the *chrR* FMN-binding motif (Ackerley et al., 2004)), removing reads with truncated or atypical motif architecture. Environmental protein sequences were dereplicated at 100% amino-acid identity using CD-HIT v4.8.1 (Li & Godzik, 2006), then combined with reference proteins and aligned using MAFFT v7.526 (Katoh & Standley, 2013). Codon-aware nucleotide alignments were produced from the environmental-only protein alignments using PAL2NAL v14 (Suyama et al., 2006). After codon-aware filtering, the final soil-only datasets comprised 242, 1,349, 74, and 1,121 unique haplotypes for *merA*, *arsC*, *cadA*, and *chrR*, respectively. Haplotypes were treated equally rather than weighted by read abundance, to avoid conflating ecological dominance with sequence diversity.

To control for the well-known heterogeneity of resistance gene families and ensure that the diversity and selection statistics reported below are not driven by inter-orthologue heterogeneity, we clustered the motif-filtered protein sequences of each gene with MMseqs2 v15 (Steinegger & Söding, 2017) at 50%, 70%, 90% sequence identity, and at a data-driven elbow threshold determined from the per-gene

identity distribution. The elbow thresholds retained were 0.83 for *merA*, 0.79 for *arsC*, 0.70 for *cadA*, and 0.73 for *chrR*. For each gene, we identified the largest cluster (containing 90%, 40%, 22%, and 45% of haplotypes for *merA*, *arsC*, *cadA*, and *chrR*, respectively). We re-ran all selection statistics restricted to this within-clade sample, to assess whether qualitative patterns persist when ortholog heterogeneity is controlled.

2.4 Phylogenetic inference and diversity statistics

Maximum-likelihood phylogenies were inferred from each codon alignment with IQ-TREE v2.3.6 under the best-fit substitution model selected by ModelFinder (BIC criterion) and 1,000 ultrafast bootstrap replicates (Minh et al., 2020). Nucleotide diversity (π) was calculated using the Nei pairwise method (Nei, 1987) separately for each gene-regime combination on soil-only haplotypes, with 95% confidence intervals derived from 500 bootstrap resamples. To assess whether differences in haplotype counts could drive the regime contrasts, we additionally subsampled each gene-by-regime sample to a common haplotype depth (n = 19, the smallest sample, *cadA* Control) with bootstrap confidence intervals from 500 replicates. Tajima's *D* was computed using the same haplotype sets, with significance assessed against the empirical distribution under the standard neutral model (Tajima, 1989). Faith's phylogenetic diversity (Faith, 1992) was computed as the total branch length of the gene-regime subtree using DendroPy v4.6 (Sukumaran & Holder, 2010), and inter-regime phylogenetic structure was tested by PERMANOVA on patristic distance matrices (9,999 permutations of regime labels).

2.5 N_e independent selection tests

To address the circularity that arises when N_e is estimated from data under selection (Charlesworth, 2009), we applied two analyses whose statistics do not depend on N_e. First, the ratio of nucleotide diversity at non-synonymous to synonymous sites (π_N/π_S) was computed using the Nei-Gojobori method (Nei & Gojobori, 1986), with 95% confidence intervals from 50 bootstrap resamples. Second, the McDonald-Kreitman (MK) test (McDonald & Kreitman, 1991) was applied per gene and per regime using the best NCBI BLAST hit (E-value < 10^{-20}) as outgroup (*Cupriavidus metallidurans* plasmid CP046332 for *merA*;

Escherichia coli CP039837 for *arsC*; *Bacillus subtilis* CP053102 for *cadA*; and *E. coli* CP040107 for *chrR*). The proportion of adaptive substitutions $\alpha = 1 - (P_N/P_S)/(D_N/D_S)$ was computed following Smith & Eyre-Walker (2002), with significance assessed by a χ^2 test with Yates correction.

To assess the sensitivity of α to outgroup choice, we re-implemented the MK test with codon alignment by MACSE v2.07 *enrichAlignment* (Ranwez et al., 2018) and added one to two BLAST-derived outgroups per gene in distinct amino-acid identity bins (close 65–72%, intermediate 55–65%, distant 40–55%), yielding 33 outgroup-by-regime comparisons in total. Sign preservation of α across outgroups, and preservation of statistical significance at $p < 0.05$, were both assessed as robustness criteria for the qualitative gene-specific differences reported below.

2.6 Codon-based selection inference and multiple-testing correction

Episodic site-level positive selection was inferred with the Mixed-Effects Model of Evolution (MEME) (Murrell et al., 2012) implemented in HyPhy v2.5.49 (Kosakovsky Pond et al., 2020), with empirical Bayes posterior probability used to identify codons with significant evidence for selection. We applied Benjamini-Hochberg false-discovery-rate (FDR) correction across MEME per-codon p-values for each gene (Benjamini & Hochberg, 1995) and report site fractions at $q < 0.05$. To complement MEME with a gene-wide test robust to alignment errors at high-ω sites, we applied BUSTED-E (Selberg et al., 2025), which adds an error-absorbing rate class to the BUSTED model. As a sanity control, we additionally ran standard BUSTED (Murrell et al., 2015) and compared its LRT to the BUSTED-E LRT to assess the contribution of the error-sink rate class. MEME, BUSTED, and BUSTED-E were each run twice per gene: once on the full gene-family alignment and once on the largest ortholog cluster (within-clade), using the corresponding IQ-TREE maximum-likelihood tree for each partition.

2.7 Taxonomic profiling of MRG-containing haplotypes

To disentangle community-composition shifts from within-lineage evolution, we assigned best-hit taxonomic affiliation to each MRG-containing haplotype by BLASTP against the NCBI non-redundant

protein database (nr, March 2026 release; E-value < 10^{-20}, top hit retained). Affiliations were aggregated to the order level using a standardized mapping based on NCBI Taxonomy to ensure cross-gene comparability, and the resulting compositions are reported per regime. Sequences for which no high-confidence assignment could be retrieved within the wall-clock allowance were retained for sequence-level analyses but excluded from the taxonomy figure; taxonomic coverage was 100% for *merA* and *cadA*, 93% for *arsC*, and 92% for *chrR* (the 76 unassigned *chrR* sequences are all from the Low regime; see Limitations).

2.8 Gene-pool diversity and exchange (Supplementary Material)

Bayesian coalescent estimates of the population mutation parameter $\theta = 2N_e\mu$ under the Coalescent Constant Population model (Kingman, 1982) using BEAST2 v2.7.6 (Bouckaert et al., 2019), and inter-regime gene-flow rates under the structured-coalescent model MASCOT (Müller et al., 2018), were performed but are only presented in the Supplementary Material. Because the panmictic-population, neutrality, and vertical-inheritance assumptions of these models are only partially met by metagenomic gene pools, we interpret θ and migration rates as descriptive indices of gene-pool diversity and exchange rather than as organismal demographic parameters. Convergence was assessed with three independent MCMC chains of 50 million iterations each (10% burn-in, ESS > 200 for all parameters, inter-chain coefficient of variation < 5%).

2.9 Statistical analysis of metal-diversity relationships and data availability

Associations between π and metal concentrations were summarized descriptively by Spearman rank correlations between gene-specific π values and the concentration of the target metal (*merA*–Hg, *arsC*–As, *cadA*–Zn, *chrR*–Cr) across the three regimes (Colby et al., 2020; St. Pierre et al., 2019). With only three regimes, the Spearman p-values are not interpretable as hypothesis tests, and we report them only as descriptive directional summaries. Bayesian directional probabilities $P(\pi_X > \pi_Y)$ were computed from the bootstrap distributions of π as the proportion of bootstrap replicates in which π_X exceeded π_Y, providing a complementary measure of bootstrap directional credibility.

Raw metagenomic DNA reads analyzed in this thesis are publicly available in the NCBI Sequence Read Archive under BioProject PRJNA556841 (Colby et al., 2020). Trace metal and total Hg concentration data are from St. Pierre et al. (2019). All custom analysis scripts, configuration files for IQ-TREE, HyPhy, BEAST2, and MASCOT, and supporting Python and R code are deposited in the GitHub repository at https://github.com/Franckoued/mrg-evolution-lakehazen. The repository includes a README that describes each analysis step, software versions, and the exact command-line invocations used in the present work.

Chapter 3: Results

3.1 Haplotype recovery and overall diversity patterns

The complete analysis workflow is summarized in Figure S1. Application of the two-stage filtering strategy (homology screen at E-value ≤ 10^{-5} followed by gene-specific catalytic motif validation; see Section 2.3) to the soil-only metagenomic dataset yielded 2,786 unique MRG haplotypes after dereplication: 242 for *merA*, 1,349 for *arsC*, 74 for *cadA*, and 1,121 for *chrR* (Table 1; site-level values in Table S1). Haplotype recovery varied substantially among genes, reflecting differences in gene prevalence rather than sequencing depth, as all eight sites were sequenced to comparable depth in the original Colby et al. (2020) survey. The two arsenate- and chromate-targeting genes *arsC* and *chrR* together accounted for 89% of soil haplotypes, whereas *merA* and *cadA* were considerably less abundant. Within each gene, the distribution across regimes was uneven. For *chrR*, the Control site CSOIL contributed 726 haplotypes (65% of the total) while the Low site LRSOIL contributed only 76 (7%). For *arsC*, the distribution was more balanced across regimes (CSOIL = 550, LRSOIL = 276, HRSOIL = 523). These asymmetries are reported alongside diversity statistics in the sections that follow because they constrain the precision of regime-level inference for genes with small per-regime haplotype pools, most notably *cadA* (n = 19–29 haplotypes per regime).

Table 1. Summary statistics for the four metal resistance genes across hydrological regimes at Lake Hazen. N = number of unique soil haplotypes. π reported as mean [95% bootstrap CI]. Tajima's *D* significance: * $p < 0.05$; ns not significant. π_N/π_S = Nei-Gojobori ratio with 95% bootstrap CI. α = McDonald-Kreitman proportion of adaptive substitutions; significance from χ^2 test with Yates correction (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns not significant).

Gene	Regime	*N*	*π* [95% CI]	Tajima's *D*	π_N/π_S [95% CI]	α (MK)
merA	Control	60	0.044 [0.036–0.054]	−1.07	0.220 [0.192–0.237]	+0.488 ***
	Low	99	0.069 [0.047–0.097]	−1.68 *	0.245 [0.210–0.277]	+0.346 **
	High	83	0.089 [0.064–0.125]	−1.63 *	0.266 [0.228–0.323]	+0.321 **
arsC	Control	550	0.190 [0.181–0.200]	+1.08	0.337 [0.325–0.348]	−0.380 ns
	Low	276	0.237 [0.223–0.250]	+1.90 *	0.449 [0.421–0.476]	+0.007 ns
	High	523	0.203 [0.194–0.211]	+1.35	0.452 [0.432–0.467]	−0.371 ns
cadA	Control	19	0.310 [0.198–0.389]	+0.24	0.866 [0.804–0.982]	+0.533 ***
	Low	29	0.296 [0.210–0.370]	+0.19	0.856 [0.757–0.903]	+0.649 ***
	High	26	0.423 [0.394–0.441]	−0.17	0.607 [0.521–0.699]	+0.538 ***
chrR	Control	726	0.150 [0.143–0.157]	+0.20	0.576 [0.554–0.589]	+0.198 ns
	Low	76	0.240 [0.218–0.265]	+0.79	0.454 [0.422–0.496]	+0.169 ns
	High	319	0.130 [0.118–0.143]	−0.42	0.630 [0.596–0.661]	+0.362 *

After motif-based filtering for complete open reading frames covering at least 90% of the reference gene length, codon alignments ranged from 486 bp (162 codons) for *arsC* to 2,181 bp (727 codons) for *cadA*, with intermediate values for *merA* (574 codons) and *chrR* (222 codons).

Nucleotide diversity (π) ranged from 0.044 (*merA*, Control) to 0.423 (*cadA*, High) and showed three distinct response patterns across the runoff gradient (Figure 3): *merA* increased monotonically with runoff (Control 0.044 → Low 0.069 → High 0.089), a doubling of diversity along the gradient. The Bayesian posterior probability $P(\pi_H > \pi_C)$ computed from the bootstrap distributions reached 0.994, providing strong directional credibility for the upward trend. The *cadA* gene showed a comparable

directional pattern, although with broader confidence intervals reflecting the smaller haplotype pool: π rose from 0.310 in Control to 0.296 in Low and 0.423 in High, with $P(\pi_H > \pi_C) = 0.992$.

The *arsC* gene exhibited a non-monotonic pattern, with peak diversity in the Low runoff regime: $\pi = 0.237$ in Low, exceeding both Control (0.190) and High (0.203) values. Bayesian $P(\pi_L > \pi_C) = 1.000$ supports the maximum in Low. This pattern is consistent with an intermediate-disturbance-like response, in which moderate metal stress maintains greater haplotype diversity than either negligible exposure (Control) or high exposure (where strong selection may favour a narrower set of variants). The *chrR* gene showed the most striking departure from the expected H > C ordering. In the soil-only data, π was highest in Low (0.240), intermediate in Control (0.150) and lowest in High (0.130), yielding a pattern of L > C > H with $P(\pi_L > \pi_C) = 1.000$ and $P(\pi_H > \pi_C) = 0.002$. This finding is counterintuitive given that Cr concentrations are highest in the High regime (19.73 ± 2.24 μg/L). Possible biological interpretations, including the action of a recent selective sweep that constrains diversity under high chromate exposure, are developed and discussed below.

Tajima's *D* values supported these contrasts (Table S2; Figure 4; per-gene soil-only values in Figure S2): *merA* exhibited negative values consistent with directional selection or expansion ($D = -1.07$ to -1.68, significant in Low and High); *arsC* showed positive Tajima's *D* across all soil regimes ($D_C = +1.08$, $D_L = +1.90$, $D_H = +1.35$), with the Low regime value being significant, indicating an excess of intermediate-frequency haplotypes consistent with balancing selection maintaining multiple arsenate-reductase variants in the gene pool; *cadA* showed near-neutral Tajima's *D* in all regimes, none of which reached significance, consistent with a small haplotype pool and potentially competing selective pressures from the multiple metal substrates that *CadA* transports (Zn, Cd, Co, Pb); *chrR* showed a heterogeneous pattern with negative *D* in High (−0.417) combined with the lowest π at this site, consistent with a recent selective sweep or a population contraction affecting *chrR* haplotypes under elevated chromate exposure. To assess robustness to uneven sampling, all gene-by-regime samples were subsampled to a common haplotype depth (n = 19, the smallest sample) with bootstrap confidence intervals from 500 replicates. The qualitative direction of

the regime contrasts in π was preserved in three of four genes (*merA*, *arsC*, *chrR*; Table S3), indicating that the observed patterns are not artifacts of uneven haplotype counts.

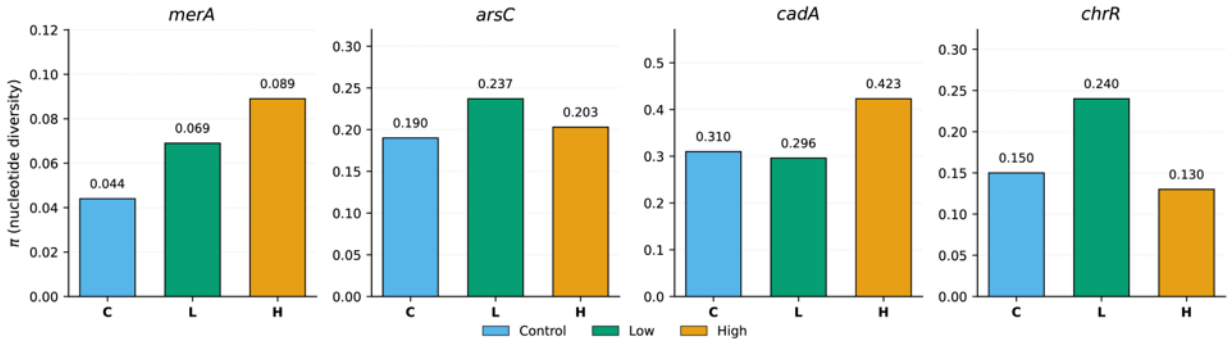


Figure 3. Nucleotide diversity (π, Nei pairwise method) per gene and regime, computed on soil-only haplotypes. Bars: regime mean; numeric labels: regime mean value. Wong palette: Control = sky blue, Low = bluish green, High = orange. Numerical values are reported in Table 1 (main text) and Table S1 (site-level breakdown).

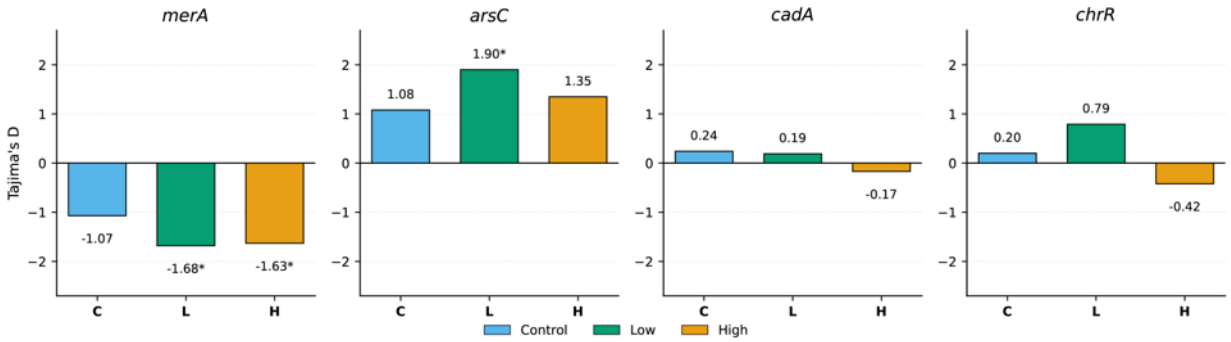


Figure 4. Tajima's *D* statistics for soil-only haplotypes per gene and regime. Bars: point estimates. Negative D values are consistent with population expansion or purifying selection; positive values may indicate balancing selection or population contraction. Significance from the beta-approximated null distribution: * $p < 0.05$. Numerical values are reported in Table S2.

3.2 Differential selective constraint independent of N_e (π_N/π_S)

Nei-Gojobori π_N/π_S ratios revealed gene- and regime-specific patterns of selective constraint (Figure 5; Table S4). All four genes showed ratios below 1 in all regimes, consistent with the dominance of purifying selection across the dataset. Within this general purifying regime, the magnitude and direction of the ratio's response to runoff differed substantially among genes.

The *merA* gene exhibited the strongest constraint (π_N/π_S = 0.220 in Control, 0.245 in Low, 0.266 in High) with a slight monotonic relaxation along the runoff gradient. The *arsC* gene showed intermediate ratios that increased from Control (0.337) to the metal-exposed regimes (Low: 0.449; High: 0.452), indicating a relaxation of purifying selection under metal stress. The *cadA* gene displayed the weakest constraint overall but with a striking regime-specific reinforcement: π_N/π_S dropped from 0.866 in Control and 0.856 in Low to 0.607 in High, a 30% reduction in High that indicates substantially stronger purifying selection on the Zn-binding ATPase under elevated Zn exposure (108 ± 12.4 µg/L). The *chrR* gene showed a non-monotonic pattern with a minimum in Low (0.454) and elevated values in Control (0.576) and High (0.630) (Figure S3). Bootstrap 95% CIs from 500 resamples confirmed the robustness of these regime-specific differences, in particular the *cadA* reduction in High and the *chrR* minimum in Low.

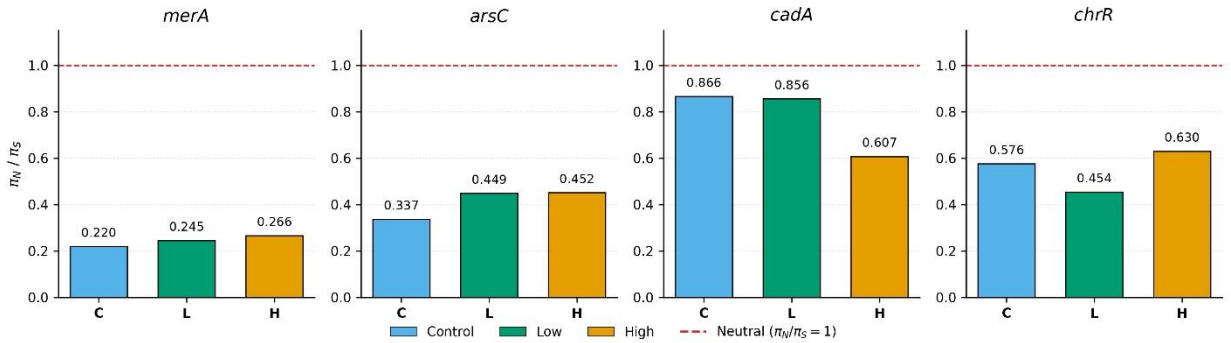


Figure 5. Ratio of nucleotide diversity at non-synonymous (π_N) to synonymous (π_S) sites (Nei-Gojobori, 1986), per gene and regime. Error bars: 95% bootstrap CI from 500 resamples. Dashed line: neutral expectation ($\pi_N/\pi_S = 1$); ratios below the dashed line indicate purifying selection. Note the reinforced constraint for *cadA* in High (0.61) and the minimum for *chrR* in Low (0.45). Numerical values are reported in Table S4.

3.3 Non-synonymous versus synonymous divergence from the McDonald-Kreitman test

The McDonald-Kreitman test, when interpreted as a descriptive index, revealed gene- and regime-specific patterns of diversification independent of N_e estimation (Table 1; Table S5; Figure 6). The *merA* gene showed a significant excess of non-synonymous over synonymous divergence across all three regimes ($\alpha = +0.488$, $p < 0.001$ in Control; $\alpha = +0.346$, $p = 0.004$ in Low; $\alpha = +0.321$, $p = 0.007$ in High), consistent with sustained patterns of diversification along the entire gradient.

The *cadA* gene exhibited the strongest and most consistent adaptive signal of the four MRGs, with α reaching its maximum in the Low regime ($\alpha = +0.649$, $p < 0.001$) and remaining highly significant in both Control ($\alpha = +0.533$) and High ($\alpha = +0.538$; all $p < 0.001$). The combination of strong diversification across all regimes and reinforced purifying selection in High ($\pi_N/\pi_S = 0.607$) indicates a gene pool that simultaneously fixes beneficial amino acid replacements and selects against deleterious ones, particularly under elevated Zn exposure.

The *chrR* gene displayed the most striking regime-specific pattern of any gene tested: significant diversification exclusively in the High regime ($\alpha = +0.362$, $p = 0.038$), with no significant signal in Control ($\alpha = +0.198$, $p = 0.33$) or Low ($\alpha = +0.169$, $p = 0.44$). The α value in High is directly co-located with the

steep maximum in Cr concentration (19.73 ± 2.24 µg/L), and the regime-specific signal indicates that diversification at *chrR* is driven by the local intensity of chromate exposure rather than by the gene pool's chronic exposure history. By contrast, the *arsC* gene showed no significant departure from neutrality in any regime ($|\alpha| < 0.4$, $p > 0.2$ for Control and High; $\alpha = +0.007$, $p = 1.000$ for Low).

To assess the sensitivity of these signals to outgroup choice, we re-implemented the MK test with MACSE codon alignment (Ranwez et al., 2018) and one to two additional BLAST-derived outgroups per gene in distinct amino-acid identity bins, yielding 33 outgroup-by-regime comparisons in total (Figure 7; Table S6). The sign of α was preserved in 9 of 12 gene-by-regime cells (75%) and statistical significance at $p < 0.05$ was preserved in 6 of 12 (50%) across canonical, close, and intermediate or distant outgroups. In particular, the *merA* positive signal was preserved across all outgroups in all three regimes, and the *cadA* positive signal was preserved across all tested outgroups in Control, Low and High. The *chrR* positive signal in High was reinforced by both close and intermediate outgroups. The *arsC* absence of signal was preserved in Control, mixed in Low, and reversed in High under close outgroup, indicating that the qualitative gene-specific differences reported here are not artifacts of outgroup choice for *merA*, *cadA* or *chrR*.

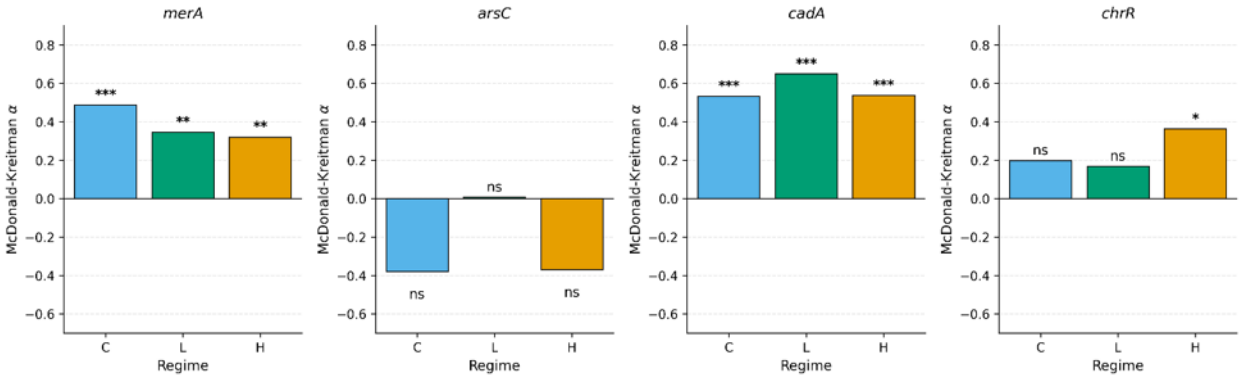


Figure 6. McDonald-Kreitman α (Smith & Eyre-Walker, 2002), the proportion of adaptive substitutions, computed per gene and regime against the reciprocal best BLAST outgroup at E-value < 10^{-20}. Significance from χ^2 test with Yates correction: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant. Numerical values are reported in Table S5.

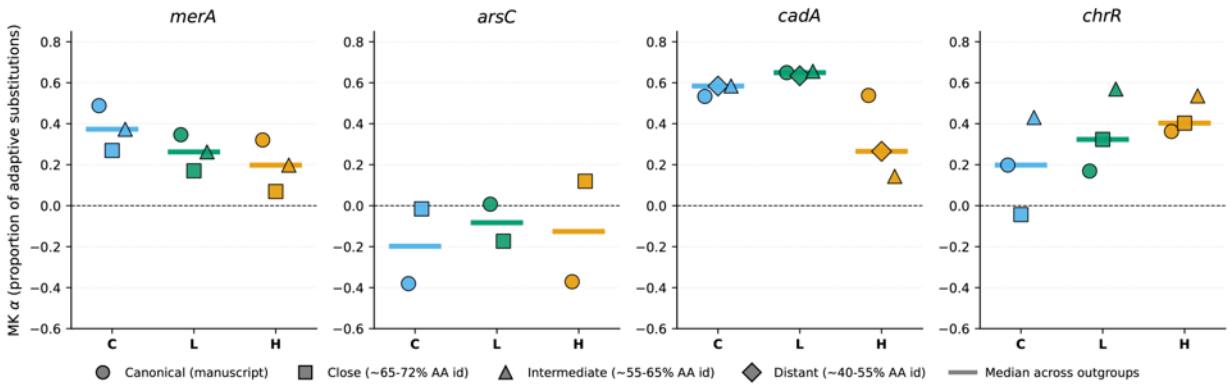


Figure 7. McDonald-Kreitman α across outgroups (MACSE codon alignment, 33 tests). Each point represents one outgroup-by-regime combination; the marker indicates outgroup class (circle = canonical, square = close ~65–72% AA identity, triangle = intermediate ~55–65%, diamond = distant ~40–55%). Horizontal bars indicate the median across outgroups for each regime. Numerical values are reported in Table S6.

3.4 Codon-level and gene-wide diversification signals (MEME, BUSTED, BUSTED-E)

MEME identified sites with signals of episodic diversification in all four genes (Figure 8; Table S7). At raw $p < 0.05$ (empirical Bayes filter), the proportion of sites under selection varied markedly among genes: *chrR* (30.2%) and *arsC* (25.3%) showed the highest raw fractions, while *merA* (4.0%) and *cadA* (1.7%) showed moderate to low fractions. However, these raw fractions are known to inflate false-positive rates in codon-based selection tests when the synonymous rate α approaches zero at some sites.

To address this, we applied Benjamini-Hochberg false-discovery-rate (FDR) correction across per-codon p-values for each gene (Benjamini & Hochberg, 1995). After FDR correction at $q < 0.05$, the proportion of sites under selection was 3.5% for *merA* (20 of 574 codons), 48.8% for *arsC* (79 of 162 codons), 1.4% for *cadA* (10 of 727 codons), and 54.5% for *chrR* (121 of 222 codons). The high FDR-corrected fractions for *arsC* and *chrR* remain implausibly large for genuine episodic selection distributed uniformly across the alignment. This inflation prompted the orthology-aware re-analysis reported above, in which the same analyses were repeated within the largest MMseqs2 ortholog cluster of each gene to assess whether the signal is intrinsic to a single orthogroup or reflects heterogeneity among divergent subfamilies.

To complement the site-level MEME analysis, we applied BUSTED-E (Branch-site Unrestricted Statistical Test for Episodic Diversification with Error sink; Selberg et al., 2025), a gene-wide test that includes an explicit error-sink rate class designed to absorb spurious signals of positive selection arising from alignment errors or model violations. BUSTED-E confirmed gene-wide patterns consistent with episodic diversification in all four genes after accounting for potential alignment errors (*merA* LRT = 24.26, $p = 2.7\times10^{-6}$; *arsC* LRT = 45.11, $p = 8.0\times10^{-11}$; *cadA* LRT = 80.82, $p << 0.001$; *chrR* LRT = 29.69, $p = 1.8\times10^{-7}$). The error-sink rate class ($\omega > 100$) absorbed less than 1% of the site weight for all four genes, indicating that the signal is not predominantly attributable to alignment artifacts.

As a sanity control, we additionally ran standard BUSTED (Murrell et al., 2015) on the same alignments and compared its LRT to the BUSTED-E LRT (Figure 9). The standard-BUSTED LRTs were 8× to 115× larger than the BUSTED-E LRTs (*merA* 197.26 vs 24.26; *arsC* 4156.66 vs 45.11; *cadA* 137.64 vs 80.82; *chrR* 3414.04 vs 29.69), confirming that BUSTED-E's error-sink class absorbs a substantial fraction of the signal that standard BUSTED would attribute to selection in datasets with alignment heterogeneity. This corroborates the choice of BUSTED-E as the appropriate conservative inference for metagenomic MRG data and provides quantitative evidence that a large fraction of the raw MEME/BUSTED signal for *arsC* and *chrR* reflects alignment heterogeneity rather than site-level positive selection acting uniformly across the gene.

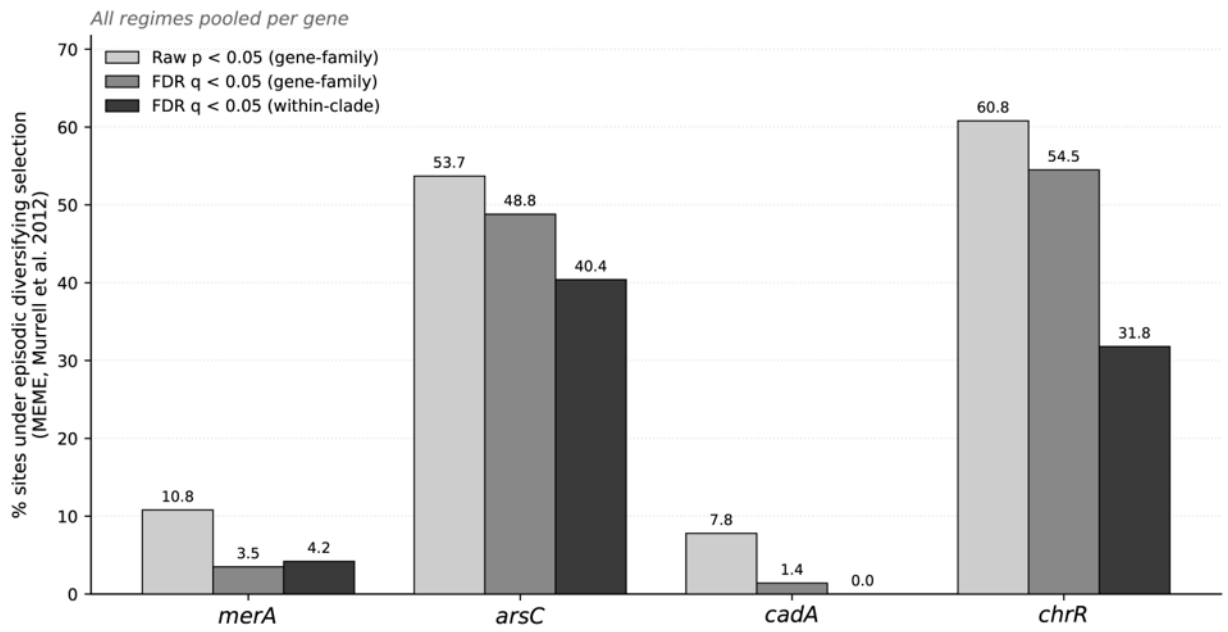


Figure 8. Percentage of codon sites under episodic diversifying selection (MEME, Murrell et al., 2012) per gene under three conditions: raw $p < 0.05$ at the gene-family level, Benjamini-Hochberg FDR $q < 0.05$ at the gene-family level, and FDR $q < 0.05$ within the largest MMseqs2 ortholog cluster (within-clade). All values are pooled across regimes per gene. Numerical values are reported in Table S7.

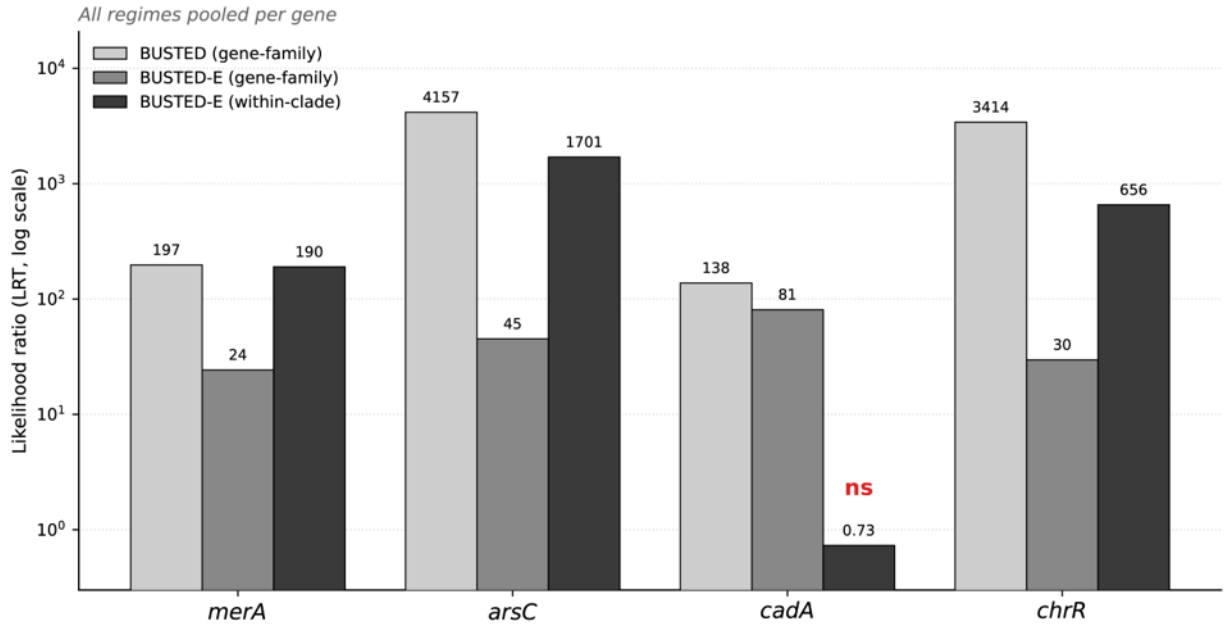


Figure 9. Standard BUSTED, BUSTED-E (gene-family), and BUSTED-E (within-clade) likelihood-ratio statistics (LRT, log scale) per gene. The ns annotation indicates non-significance of the BUSTED-E within-clade test for cadA (LRT = 0.73, $p = 0.347$). The error-sink rate class of BUSTED-E absorbed less than 1% of the site weight for all four genes. All values are pooled across regimes per gene.

3.5 Phylogenetic diversity and beta-diversity (Faith's PD and PERMANOVA)

To complement the population-genetic statistics with assumption-light community-level inference, we report Faith's phylogenetic diversity (PD; Faith, 1992), computed as the total branch length of the gene-regime subtree, and PERMANOVA on patristic distance matrices from the gene-specific IQ-TREE phylogenies (Table S8; Table S9). Faith's PD varied substantially across genes and regimes in the soil samples (Figure S4). For *merA*, PD increased markedly from Control (0.74) through Low (4.67) to High (7.92), in close parallel with the π gradient and suggesting that the increase in nucleotide diversity reflects the recruitment of progressively more divergent lineages along the runoff gradient. For *arsC*, PD was highest at HRSOIL (35.53) and CSOIL (33.19), with LRSOIL intermediate (22.01), a pattern in which phylogenetic breadth does not strictly follow the per-regime π ordering. For *cadA*, PD was highest in LRSOIL (7.94), followed by HRSOIL (7.70) and CSOIL (5.20), while for *chrR*, CSOIL dominated (58.57),

consistent with its large haplotype count, with LRSOIL substantially lower (13.73) and HRSOIL intermediate (38.63).

PERMANOVA on patristic distance matrices (9,999 permutations of regime labels) confirmed that the hydrological regime explained a highly significant proportion of the variance in phylogenetic composition for all four genes (all $p < 0.001$; Figure S5). The proportion of variance explained (R^2) varied markedly across genes: *chrR* showed by far the strongest regime structuring ($R^2 = 0.718$), followed by *arsC* ($R^2 = 0.232$), *cadA* ($R^2 = 0.154$) and *merA* ($R^2 = 0.064$). The exceptionally high R^2 for *chrR* indicates that nearly three-quarters of phylogenetic variation in this gene is partitioned among regimes, consistent with limited inter-regime mixing and a strongly localized evolutionary history. This unusually strong signal motivated the orthology-aware and taxonomy-aware follow-up analyses reported above, which qualified the interpretation of the *chrR* regime structure in terms of compositional turnover rather than within-lineage evolution.

3.6 Correlation between nucleotide diversity and metal concentrations

Descriptive Spearman rank correlations between regime-level π and target metal concentrations (Table S10) revealed gene-specific associations (Figure 10; Figure 2). The *cadA* gene showed a positive rank correlation with Zn ($\rho = +0.50$): π increased from Control (~9 µg/L Zn; $\pi = 0.310$) through Low (44.8 µg/L; $\pi = 0.296$) to High (108 µg/L; $\pi = 0.423$), with the maximum in High where Zn concentrations are highest. The *arsC* gene also showed $\rho = +0.50$ with As, reflecting a partial monotonic relationship over a relatively narrow concentration range (0.66–1.39 µg/L). The *merA* gene showed $\rho = +0.50$ with total Hg, but the relationship is shaped by the non-monotonic Hg concentration pattern (Low > High > Control) combined with the monotonic increase of *merA* diversity along the runoff gradient: π tracks the integrated Hg flux (concentration × discharge, integrated over time) rather than the instantaneous concentration.

The *chrR* gene showed a negative rank correlation with Cr ($\rho = -0.50$): π decreased along the runoff gradient (Control 0.150 → Low 0.240 → High 0.130) while Cr increased monotonically (0.33 → 6.15 →

19.73 µg/L). Combined with the highly significant PERMANOVA regime structure (R^2 = 0.718), the regime-specific MK signal in High, and the taxonomic turnover documented above, this pattern is consistent with strong selection at high chromate exposure that constrains realized diversity within an otherwise regime-differentiated gene pool. With only three regimes per correlation, all Spearman ρ values are descriptive, and the associated p-values cannot be interpreted with conventional thresholds. The Bayesian probabilities $P(\pi_H > \pi_C)$, computed from bootstrap distributions and reported above, provide complementary directional support that is independent of the sample-size limitations of Spearman's correlation at n = 3.

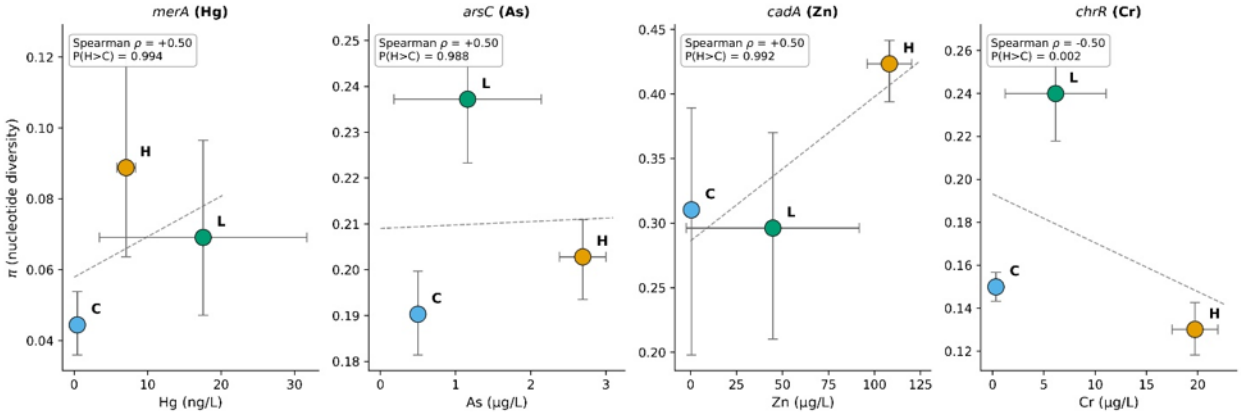


Figure 10. Relationship between nucleotide diversity (π) and the target metal concentration for each gene across the three hydrological regimes: *merA* vs Hg, *arsC* vs As, *cadA* vs Zn, *chrR* vs Cr. Each point represents one regime (C = Control, L = Low, H = High). Horizontal error bars: metal SD across replicate sites; vertical error bars: *π* bootstrap 95% CI. Spearman *ρ* is shown for each gene; p-values are not interpretable at n = 3.

3.7 Gene-family versus within-clade re-analysis

The heterogeneously high MEME and BUSTED-E signals for *arsC* and *chrR*, together with the exceptionally strong PERMANOVA R^2 for *chrR*, raised the possibility that these signals reflect inter-orthologue heterogeneity within the gene family rather than diversifying selection within a single orthogroup. To test this, we clustered the motif-filtered protein sequences of each gene with MMseqs2 v15 (Steinegger & Söding, 2017) at 50%, 70%, 90% sequence identity, and at a data-driven elbow threshold

determined from the per-gene identity distribution (Table S11; Figure S6). The elbow thresholds retained were 0.83 for *merA*, 0.79 for *arsC*, 0.70 for *cadA*, and 0.73 for *chrR*.

MMseqs2 clustering at the per-gene elbow threshold confirmed substantial heterogeneity in orthology coherence among the four gene families: *merA* approached a single orthologous group within our dataset (217 of 242 haplotypes, or 90%, in the largest cluster), whereas *cadA* was highly heterogeneous (the largest clade accounting for only 16 of 74 haplotypes, or 22%), and *arsC* (540 of 1,349, or 40%) and *chrR* (503 of 1,121, or 45%) were intermediate. We re-ran all selection statistics restricted to the largest cluster of each gene (Table 2; Table S12; Figure S7) to assess whether the qualitative patterns persist when ortholog heterogeneity is controlled.

Within the largest cluster of each gene, the qualitative direction of the gene-family-level MK signal was preserved for *merA* (α within-clade = +0.47 / +0.50 / +0.51 for C/L/H; compare gene-family α = +0.488 / +0.346 / +0.321). The π_N/π_S pattern and absence of MK signal for *arsC* were also preserved (α within-clade = −0.37 / +0.19 / −0.16). For *merA*, BUSTED-E within-clade was strongly reinforced (LRT = 189.92 within-clade vs 24.26 gene-family) and MEME $q < 0.05$ was essentially unchanged (4.2% within-clade vs 3.5% gene-family, FDR-corrected). We interpret this convergence across metrics as strong evidence that the diversifying selection signal for *merA* is intrinsic to the dominant orthogroup and not driven by inter-cluster heterogeneity, consistent with the fact that *merA* haplotypes coalesce into a single dominant orthogroup within this dataset.

By contrast, for *cadA*, the BUSTED-E signal of episodic diversifying selection at the gene-family level (LRT = 80.82) collapsed to non-significance within the largest clade (LRT = 0.73, $p = 0.347$), and the MEME percentage dropped from 1.4% (gene-family, FDR) to 0.0% (within-clade, FDR). This near-total loss of the BUSTED-E and MEME signals when restricted to a single orthogroup indicates that a substantial fraction of the gene-family-level signal reflects heterogeneity among *cadA* subfamilies rather than within-orthogroup diversifying selection. The MK signal for *cadA*, by contrast, persisted within the largest clade

in the two regimes where the clade was populated (α within-clade = +0.71 in Low and +0.58 in High); the 19 Control *cadA* haplotypes were distributed among smaller subfamilies and did not contribute to the largest-cluster analysis. We interpret this dissociation as evidence that gene-family-level BUSTED-E is dominated by inter-subfamily heterogeneity for *cadA*, while the MK component of the signal is at least partly intrinsic to the dominant orthogroup in the exposed regimes.

For *chrR*, the BUSTED-E signal was strongly preserved within the largest cluster (LRT = 655.95, $p << 0.001$), and MEME $q < 0.05$ decreased from 54.5% to 31.8% after FDR correction within the clade. However, closer inspection revealed that the largest *chrR* cluster was dominated by Control sequences (484 of 503 haplotypes, 96%); Low- and High-regime *chrR* haplotypes partitioned into other clusters. In other words, the observed regime contrast in *chrR* is not a within-cluster contrast but a between-cluster contrast: different orthologous subfamilies of *chrR* dominate in different regimes. This finding indicates that the regime contrasts documented in Sections 3.1–3.5 for *chrR* reflect compositional turnover among orthologous subfamilies rather than within-orthogroup evolution, and it motivates the taxonomic-level investigation reported above.

Table 2. MMseqs2 ortholog clustering at the data-driven elbow threshold per gene, and selection statistics within the largest ortholog cluster (within-clade re-analysis). α(MK) within-clade C/L/H reports the McDonald-Kreitman proportion of adaptive substitutions per regime computed within the largest cluster. n = 0 indicates that the largest cluster contains no haplotypes from that regime. BUSTED-E LRT and MEME % sites (FDR q < 0.05) are gene-wide statistics pooling all regimes within the largest cluster.

Gene	Threshold	Clusters	Largest cluster	α(MK) C / L / H	BUSTED-E LRT	MEME % sites
merA	0.83	10	217 / 242 (90%)	+0.47 / +0.50 / +0.51	189.92 (p < 0.001)	4.2%
arsC	0.79	83	540 / 1349 (40%)	−0.37 / +0.19 / −0.16	1700.95 (p < 0.001)	40.4%
cadA	0.70	15	16 / 74 (22%)	n = 0 / +0.71 / +0.58	0.73 (p = 0.347, ns)	0.0%
chrR	0.73	42	503 / 1121 (45%)	+0.18 / +0.41 / n = 0	655.95 (p < 0.001)	31.8%

3.8 Taxonomic composition of MRG-containing haplotypes

To disentangle community-composition shifts from within-lineage evolution, we assigned best-hit taxonomic affiliation to each MRG-containing haplotype by BLASTP against the NCBI non-redundant protein database (nr, March 2026 release; E-value < 10^{-20}, top hit retained). Affiliations were aggregated to the order level using a standardized mapping based on NCBI Taxonomy to ensure cross-gene comparability. Taxonomic coverage was 100% for *merA* and *cadA*, 93% for *arsC*, and 92% for *chrR* (the 76 unassigned *chrR* sequences are all from the Low regime; see Limitations). Taxonomic profiles revealed strongly gene- and regime-specific patterns (Table S13; Figure S8).

For *merA*, we observed a monotone compositional turnover along the gradient: in Control, 77% of haplotypes were best-matched to Chromatiales; in Low, the Chromatiales fraction fell to 29% while unclassified Pseudomonadota rose to 63%; in High, the dominant matches shifted again toward Gammaproteobacteria (47%) and Pseudomonadota (33%), with the first appearance of Acidobacteriota and Actinomycetota. Importantly, this compositional turnover for *merA* was accompanied by the reinforced BUSTED-E signal within the largest orthogroup, indicating that *merA* diversifies both across host lineages

(compositional shift) and within its dominant orthogroup (within-clade signal preserved). This dual pattern is consistent with the known mobility of *merA* on the Tn21 transposon family (Liebert et al., 1999), which allows the same orthologous gene to circulate across different host backgrounds along the runoff gradient.

For *arsC*, Burkholderiales dominated in Control (50%) and High (47%), while Low was the most diversified regime, with Hyphomicrobiales rising to 45%. This dominance pattern is broadly consistent with the intermediate-disturbance π maximum in Low: moderate metal stress in Blister Creek may support the coexistence of multiple *arsC*-bearing orders, whereas Control and High are dominated by Burkholderiales carrying diverse *arsC* variants.

For *cadA*, no single order dominated in any regime; Control was largely unclassified (47%), Low was dominated by Gammaproteobacteria (21%) and Acidobacteriota (24%), and High shifted strongly toward Bacteroidota (38%, including Chitinophagaceae, *Pedobacter*, and *Chitinophaga*). This pronounced compositional turnover between regimes, especially the emergence of Bacteroidota in High, is consistent with the within-clade collapse of the BUSTED-E signal: the gene-family-level MEME/BUSTED-E signal for *cadA* is inflated by the presence of divergent subfamilies encoded by taxonomically distinct host orders, and it does not resolve into a coherent within-orthogroup signal when subfamily heterogeneity is controlled.

Most notably, for *chrR*, we observed a shift from a diversified Control community (Hyphomicrobiales 80%, with a long tail of unclassified Pseudomonadota and other groups) to a near-monoculture in High (Hyphomicrobiales 93%, predominantly Xanthobacteraceae). Because the largest chrR cluster was 96% Control and the High regime is 93% Hyphomicrobiales, the chrR regime contrast is best explained by a displacement of the diverse Control chrR community by a Xanthobacteraceae-dominated High-runoff community, rather than by within-orthogroup diversification. This interpretation reconciles all the *chrR*-specific observations reported in Sections 3.1–3.6: the extreme PERMANOVA R^2 (0.718), the low π in High despite high Cr concentration, the negative Tajima's *D* in High, the highly significant MK α in High, and the high FDR-corrected MEME fraction. All are consistent with a community

turnover in which a chromate-tolerant Xanthobacteraceae lineage displaces the diversified Control community under high Cr exposure.

Chapter 4: Discussion

4.1 Convergent evidence for gene- and regime-specific evolutionary patterns

The central conclusion emerging from the present analyses is that variable glacial meltwater inputs at Lake Hazen are associated with detectable, gene- and regime-specific patterns of sequence variation in the metal resistance gene pools of the microbial soil community. This conclusion is supported by the convergence of five complementary lines of evidence: descriptive characterization of nucleotide diversity, two N_e-independent selection tests (π_N/π_S and McDonald-Kreitman with outgroup-sensitivity analysis, Sections 3.2–3.3), codon-based selection tests with FDR correction and error-sink absorption (MEME, BUSTED and BUSTED-E, Section 3.4), community-level inference (Faith's PD and PERMANOVA on patristic distances, Section 3.5), and orthology- and taxonomy-aware controls (within-clade re-analysis and best-hit taxonomic profiling, Sections 3.7–3.8). The methodological argument for assembling such a battery is straightforward: each individual method has known limitations, but signals that emerge consistently across approaches with non-overlapping assumptions can be interpreted with substantially greater credibility than signals supported by a single line of evidence.

Because these inferences derive from metagenomic gene pools sampled across only three hydrological regimes, and because metagenomic gene pools aggregate variants across taxa and across potentially divergent members of each gene family, we interpret the patterns found here as exploratory, hypothesis-generating observations rather than as formal demonstrations of population-level adaptation. The orthology-aware and taxonomy-aware analyses are critical to this interpretation: they identify which patterns are robust at the within-orthogroup level and which depend on inter-cluster heterogeneity or community-composition shifts. Three contrasting gene-level patterns emerged from this integrated analysis. The first pattern is sustained diversifying evolution along the runoff gradient, observed for *merA* and robust to within-orthogroup controls. The second pattern is regime-specific signals partly attributable to subfamily heterogeneity or community-composition turnover, observed for *cadA* and *chrR* respectively. The third pattern is a near-absence of detectable diversifying signal, illustrated by *arsC*. These three patterns are

discussed in turn below, followed by a methodological reflection on the reframing of coalescent outputs and the role of horizontal gene transfer, the implications for Arctic biomonitoring under climate change, and an explicit treatment of the limitations of the study.

4.2 First pattern: sustained diversification along the runoff gradient (*merA*)

The *merA* gene showed the most internally consistent set of evolutionary signatures across the five analytical approaches. Nucleotide diversity increased monotonically from Control (0.044) through Low (0.069) to High (0.089), with Bayesian $P(\pi_H > \pi_C) = 0.994$ providing strong directional credibility for the upward trend. McDonald-Kreitman analysis detected significant diversification in all three regimes (α = +0.488 to +0.321, all $p \leq 0.01$), and this signal was preserved under outgroup-sensitivity analysis (sign preserved in all three regimes across canonical, close and intermediate outgroups). MEME identified 3.5% of codons as showing episodic positive selection at FDR $q < 0.05$, with hits distributed across the FAD-binding domain, the catalytic-Cys motif region, and the C-terminal Hg-binding domain. BUSTED-E confirmed gene-wide evidence of episodic positive selection (LRT = 24.26, $p = 2.7\times10^{-6}$) after accounting for potential alignment errors.

The π_N/π_S ratio remained low across all regimes (0.220–0.266) with a slight monotonic relaxation along the runoff gradient. This pattern indicates that purifying selection dominates the *merA* gene pool while still permitting the fixation of non-synonymous replacements associated with the upward MK signal. The combined picture is one of a gene pool experiencing sustained sequence variation associated with the Hg gradient along the gradient, with adaptive substitutions accumulating at a rate consistent with the integrated Hg flux of each regime.

Critically, all these signals were preserved (indeed reinforced) within the largest orthologous cluster, which contained 217 of 242 *merA* haplotypes (90%). Within-clade BUSTED-E was substantially higher than gene-family BUSTED-E (LRT = 189.92 vs 24.26), and within-clade MEME $q < 0.05$ was essentially unchanged (4.2% vs 3.5%). This orthogroup coherence is a distinguishing feature of *merA* in

our dataset: because most of the soil *merA* haplotypes coalesce into a single orthologous cluster, the observed gene-family-level signals of diversifying evolution can be attributed to within-orthogroup diversification with high confidence. We interpret this as the strongest within-orthogroup signal in our dataset, consistent with sustained Hg delivery to the lake basin, a pattern consistent with gene-pool diversification.

The runoff-tracking pattern of *merA* diversity merits closer attention because Hg concentrations are highest in the Low runoff regime (Blister Creek; 17.56 ± 14.14 ng/L) rather than in the High regime (Snowgoose and Abbé; 7.07 ± 1.27 ng/L). The fact that *merA* diversity nonetheless tracks the runoff gradient rather than the concentration gradient is most parsimoniously interpreted as evidence that the relevant exposure metric for soil microbiomes is the cumulative Hg flux, that is, the product of concentration and discharge volume integrated over time, rather than the instantaneous concentration. Discharge in the Snowgoose-Abbé system is substantially greater than in Blister Creek, so that the total Hg delivered to High-regime sites exceeds that delivered to Low-regime sites despite the lower instantaneous concentration (St. Pierre et al., 2019). Under this flux-based interpretation, the *merA* diversity gradient is concordant with the underlying selective pressure. Recent experimental work on *merA* adaptation supports this interpretation: directed evolution and field surveys have shown that Hg exposure selects for substrate-specific *merA* variants on relatively short timescales (Christakis et al., 2021; Christakis et al., 2022).

The taxonomic profile of *merA*-bearing haplotypes provides an additional layer of interpretation. Along the gradient, we observed a shift from Chromatiales-dominated Control (77%) through mixed Pseudomonadota in Low (63% unclassified) to Gammaproteobacteria-dominated High (47%). Yet, despite this compositional turnover across host lineages, the within-orthogroup BUSTED-E signal was reinforced (LRT = 189.92 within-clade vs 24.26 gene-family), indicating that the same orthologous *merA* gene circulates and diversifies across taxonomically diverse hosts along the runoff gradient. This dual pattern (compositional turnover at the host level combined with within-orthogroup diversification at the gene level) is precisely what would be predicted for a gene whose mobility on the Tn21 transposon family allows rapid

horizontal spread between diverse bacterial hosts (Liebert et al., 1999), combined with continuous Hg-driven selection on the encoded reductase function.

4.3 Second pattern: regime-specific signals partly attributable to subfamily heterogeneity (*cadA* and *chrR*)

4.3.1 *cadA*: signals dissociate between gene-family-level and within-clade analyses

The *cadA* gene exhibited the strongest McDonald-Kreitman signal among the four MRGs at the gene-family level, with α reaching +0.649 ($p < 0.001$) in Low and remaining highly significant in Control (+0.533) and High (+0.538; all $p < 0.001$). Simultaneously, the π_N/π_S ratio dropped from 0.866 in Control and 0.856 in Low to 0.607 in High, a 30% reduction that indicates substantially reinforced purifying constraint under elevated Zn exposure (108 ± 12.4 μg/L). At the gene-family level, this combination of strong adaptive substitution and intensified purifying selection would be consistent with a scenario in which the *cadA* gene pool experiences continual exploration of variant space while environmental pressure progressively selects for a narrower set of functionally optimal variants in the most Zn-exposed sites.

However, the orthology-aware analysis substantially qualified this interpretation. When restricted to the largest MMseqs2 cluster (16 of 74 haplotypes, or 22% of the *cadA* dataset), the BUSTED-E signal at the gene-family level (LRT = 80.82) collapsed to non-significance within the largest clade (LRT = 0.73, $p = 0.347$), and the MEME percentage dropped from 1.4% (gene-family, FDR) to 0.0% (within-clade, FDR). This near-total loss of the BUSTED-E and MEME signals when restricted to a single orthogroup indicates that the gene-family-level signal is dominated by inter-subfamily heterogeneity rather than within-orthogroup diversifying selection. The MK signal, in contrast, persisted within the largest clade in the two regimes where the clade was populated (α within-clade = +0.71 in Low and +0.58 in High); the 19 Control *cadA* haplotypes were distributed among smaller subfamilies and did not contribute to the largest-cluster analysis.

We interpret this dissociation as evidence that the *cadA* diversifying signal at the gene-family level is driven in part by heterogeneity among *cadA* subfamilies, while the MK component of the signal in Low

and High is at least partly intrinsic to the dominant orthogroup. The biological relevance of *cadA* subfamily heterogeneity is supported by the known substrate breadth of cadA-family P-type ATPases for Zn, Cd, and Co (Argüello et al., 2007), which favours the evolution of substrate-specialized subfamilies in different host lineages. The pronounced taxonomic turnover observed (Bacteroidota rising to 38% in High) further supports the interpretation that *cadA* diversity at the gene-family level reflects, in part, the composition of the host community rather than solely within-orthogroup diversification. This dissociation between gene-family-level and within-clade signals is a general caveat for the interpretation of population-genetic statistics computed on unfiltered metagenomic gene pools, and it justifies the emphasis on within-clade re-analysis throughout the present study.

4.3.2 *chrR*: regime contrast dominated by compositional turnover

The *chrR* gene provided the clearest demonstration of the necessity of taxonomy-aware controls in interpreting gene-family-level statistics. At the gene-family level, MEME identified 54.5% of codons as showing evidence of episodic positive selection at FDR $q < 0.05$, the highest proportion of any gene examined, and BUSTED-E rejected the null model of no positive selection with very high confidence (LRT = 29.69, $p = 1.8\times10^{-7}$). The McDonald-Kreitman test detected significant diversification exclusively in the High regime ($\alpha = +0.362$, $p = 0.038$), with no significant signal in Control or Low. PERMANOVA showed by far the strongest regime structuring of any gene ($R^2 = 0.718$). Taken together, these gene-family-level statistics would appear to indicate strong regime-specific diversification of the *chrR* gene pool to high chromate exposure.

The orthology-aware and taxonomy-aware analyses fundamentally qualified this interpretation. First, when restricted to the largest MMseqs2 cluster (503 of 1,121 haplotypes, 45%), BUSTED-E within-clade remained strongly significant (LRT = 655.95, $p << 0.001$) and MEME FDR $q < 0.05$ was 31.8%. However, closer inspection revealed that the largest *chrR* cluster was dominated by Control sequences (484 of 503, 96%); Low- and High-regime *chrR* haplotypes partitioned into other clusters. Second, at the taxonomic level, the Control regime supported a diversified *chrR* community (Hyphomicrobiales 80%, with

a long tail of unclassified Pseudomonadota and other groups), whereas the High regime was 93% Hyphomicrobiales (predominantly Xanthobacteraceae).

We therefore interpret the *chrR* regime contrast as dominated by compositional turnover, that is, the displacement of a diversified Control *chrR* community by a Xanthobacteraceae-dominated High-runoff community, rather than by within-orthogroup diversification. This interpretation reconciles a set of observations that were initially puzzling: the low π in High (0.130) despite the highest Cr concentration; the negative Tajima's *D* in High (−0.417); the extreme PERMANOVA R^2 (0.718); the highly significant MK α in High (+0.362, $p = 0.038$); the high FDR-corrected MEME fraction (54.5% at the gene-family level, 31.8% within-clade); and the 96% Control-dominated largest cluster. All are consistent with a community turnover in which a chromate-tolerant Xanthobacteraceae lineage carrying a distinct orthologous *chrR* subfamily displaces the diversified Control community under high Cr exposure. The MEME and BUSTED-E signals that persist within the Control-dominated cluster are biologically intriguing but should be interpreted cautiously, as they describe variation within a single community context rather than adaptive change driven by the runoff gradient.

This finding for *chrR* is a cautionary example of the interpretive risks of gene-family-level statistics on metagenomic gene pools. Without the orthology-aware and taxonomy-aware controls, one might have concluded that the *chrR* regime contrast reflects strong local adaptation of a single population to Cr exposure. With these controls, we conclude instead that the contrast reflects a shift in the composition of the *chrR*-bearing community. Both phenomena are biologically real and ecologically important, but they carry different implications for the evolutionary dynamics of the gene pool and for the design of MRG-based monitoring schemes. The chromosomal location of *chrR* within genomic islands rather than on self-transmissible plasmids is consistent with a mode of inheritance in which each host lineage carries its own genomic-island-encoded *chrR* subfamily, and community turnover across regimes therefore translates directly into gene-pool turnover.

4.4 Third pattern: near-absence of detectable diversifying signal (*arsC*)

The *arsC* gene displayed a markedly different evolutionary profile from the other three MRGs. The McDonald-Kreitman α was non-significant or negative in all three regimes (α = −0.380 in Control, +0.007 in Low, −0.371 in High; all $p > 0.20$), and these patterns were preserved within the largest orthologous cluster (α within-clade = −0.37 / +0.19 / −0.16 for C/L/H). Nucleotide diversity was elevated relative to *merA* but showed a non-monotonic response to the runoff gradient, peaking in Low (0.237) rather than tracking the modest As concentration gradient (0.66–2.69 µg/L). Tajima's *D* was positive in all regimes, with the Low value reaching significance, consistent with an excess of intermediate-frequency haplotypes maintained by balancing selection or an intermediate-disturbance mechanism.

We are cautious about over-interpreting the absence of MK signal as evolutionary stasis in the formal sense. Several non-exclusive biological and methodological factors plausibly contribute to this pattern and disentangling them requires data that are not currently available. First, the As concentration gradient at Lake Hazen is comparatively modest (H/C ratio ~4×) relative to the gradients for Cr (60×) and Zn (~12×), which may not be sufficient to drive detectable regime-specific selection. Second, the two-stage motif filter deliberately excluded the unrelated thioredoxin-dependent and ACR3-dependent arsenate reductases, focusing the analysis on the glutaredoxin-dependent *ArsC* subfamily. If the microbial community harbours diverse As-resistance strategies (e.g., different arsenate reductase families in different lineages), the *arsC* gene pool we analyze here may not experience the full selective pressure of the local As environment.

A third, biogeochemical, consideration is that As (a metalloid) and Cr (a metal) are redox-sensitive elements that occur in the environment primarily as oxyanions (arsenate, arsenite, chromate), whose mobility, bioavailability, and cellular uptake depend strongly on local pH and redox potential. Their cellular accessibility therefore differs from that of divalent cations such as Zn^{2+}, Cd^{2+}, and Hg^{2+}, which form cationic ligand complexes and follow different transport pathways into the cell. Based on substrate speciation alone, one might therefore expect *arsC* and *chrR*, which both target oxyanion substrates, to behave similarly. Our

data do not support that simple grouping: even after the orthology- and taxonomy-aware controls, *chrR* shows the strongest community-level regime structure of the four genes (driven by compositional turnover,), whereas *arsC* shows no detectable diversifying signal in any regime and modest compositional turnover at the taxonomic level (Burkholderiales 50%/28%/47% and Hyphomicrobiales 17%/45%/6% across C/L/H). The contrast between these two oxyanion-targeting genes indicates that substrate speciation alone is not sufficient to predict the evolutionary trajectory of an MRG in this system.

A central limitation constraining all these interpretations is that we did not measure in situ metal speciation or redox conditions in this study. The metal concentration data available from St. Pierre et al. (2019) are total dissolved concentrations in the receiving river waters; they do not partition As between As(V) and As(III), nor Cr between Cr(VI) and Cr(III), nor the bioavailable fractions of either. Our interpretation is therefore restricted to total metal pools across regimes. Resolving the relative contributions of speciation, redox chemistry, and host physiology to MRG evolution will require coupled metagenomic and geochemical speciation surveys, ideally with seasonal sampling to capture the dynamics of meltwater-driven metal mobilization.

4.5 Methodological considerations: reframing coalescent inference and the role of horizontal gene transfer

Several methodological considerations frame the interpretation of all results presented in this thesis. First, metagenomic gene pools aggregate sequence variants across taxa and across potentially divergent members of each gene family. The polymorphism/divergence partition required by the McDonald-Kreitman framework, and the panmictic-population assumption underlying Tajima's *D*, are not strictly satisfied in this setting (Ellegaard & Engel, 2016; Garud & Pollard, 2020). We addressed this directly by presenting all population-genetic statistics as descriptive indices, by clustering sequences into orthologous subfamilies and performing selection analyses within the largest cluster, by applying FDR correction across all per-codon and per-test outputs, and by running outgroup-sensitivity analyses for the MK test. These within-clade analyses showed that the *merA* pattern is robust, the *arsC* absence of a positive signal is robust, the

cadA BUSTED-E signal disappears (indicating subfamily heterogeneity), and the *chrR* regime structure reflects compositional turnover.

Accordingly, the McDonald-Kreitman, MEME, and BUSTED/BUSTED-E results reported here should be read as descriptive, hypothesis-generating summaries rather than as formal evidence of selection, because the assumptions of population coherence, strict orthology, and vertical inheritance on which these tests rest are only partially satisfied in metagenomic gene pools. Consistent with this, some MEME and BUSTED-E signals remain unexpectedly high even after orthology and taxonomy controls; residual alignment heterogeneity, horizontal gene transfer, recombination, and incompletely resolved ortholog structure can each inflate apparent episodic-selection signals in metagenomic alignments and likely contribute to these residuals.

Second, coalescent estimators of N_e assume neutrality (Charlesworth, 2009) and tree-like vertical inheritance (Soucy et al., 2015). When selection is present, as it is for all four genes examined here, N_e estimates are biased downward: selection reduces diversity through selective sweeps or background selection, and a coalescent estimator based on observed π will infer a smaller N_e than would be present under strict neutrality. Interpreting the biased N_e as a demographic parameter and then reusing it to argue for or against selection is therefore circular. We addressed this circularity by adding two N_e-independent selection tests (π_N/π_S and McDonald-Kreitman), whose statistical machinery does not require N_e estimation. The convergence of these tests with codon-based methods and with the orthology- and taxonomy-aware controls strongly supports the conclusion that the observed variation across regimes reflects genuine differential selection or compositional turnover, not merely an artifact of N_e estimation under selection.

For these reasons, the Bayesian coalescent (BEAST2) and structured-coalescent (MASCOT) analyses were relegated to the Supplementary Material and are interpreted only as descriptive gene-pool diversity and exchange indices rather than as estimates of organismal effective population size or demographic migration (Tables S5 and S6). Under this framing, the quantity $\theta = 2N_e\mu$ estimated by

BEAST2 is reported as a summary of standing haplotype variation, and the MASCOT migration rates are reported as descriptive indices of gene-pool exchange across regimes. This reframing is appropriate for metagenomic MRG data because individual sequences cannot be unambiguously assigned to specific host organisms, so the demographic interpretation in terms of bacterial census numbers is poorly defined, and because horizontal gene transfer introduces non-vertical inheritance that violates the assumption of a single underlying tree.

Third, the role of horizontal gene transfer (HGT) warrants explicit discussion. Several of the metal resistance genes examined here are commonly associated with mobile genetic elements: *merA* frequently resides on the Tn21 transposon family (Liebert et al., 1999); *cadA* is often plasmid-borne in firmicutes and other taxa (Nies, 2003); *arsC* occurs on both chromosomes and plasmids depending on the lineage. Our outgroup BLAST searches independently confirmed this pattern: the best *merA* hit was located on a plasmid in *Cupriavidus metallidurans* (NCBI CP046332.1), corroborating Tn21-mediated mobility documented in the prior literature. HGT violates the assumption of vertical inheritance underlying coalescent models. A *merA* sequence retrieved from the soil metagenome may share recent common ancestry with another *merA* sequence not because the two host organisms share a recent ancestor, but because the gene has moved between distantly related host lineages via plasmid conjugation or transposition. The single-tree assumption underlying BEAST2 and MASCOT is therefore an approximation, and the estimated coalescent parameters should be interpreted as summaries of gene-pool dynamics rather than of organismal phylogeny. Two considerations partially mitigate this concern. First, the convergence across the N_e-independent tests, codon-based tests, and orthology- and taxonomy-aware controls, each with different assumptions about inheritance, suggests that the observed gene- and regime-specific patterns are not artifacts of any single method's HGT-related violations. Second, the taxonomic profile of *merA*-bearing haplotypes documented (compositional turnover across host lineages combined with within-orthogroup within-clade signal) is precisely what the HGT-mediated gene-pool framing predicts: the same orthologous *merA* gene circulates

across taxonomically diverse hosts, and the within-orthogroup selection signal captured by our analyses reflects the evolution of the circulating gene rather than of any single host lineage.

4.6 Implications for Arctic biomonitoring under climate change

The patterns documented in this thesis have practical implications for monitoring Arctic ecosystems under accelerating climate change. The orthology-robust *merA* pattern indicates that Hg-resistance gene pools can respond on timescales relevant to contemporary glacial retreat (years to decades) and could serve as candidate sensitive biomarkers complementing traditional geochemical measurements. Because the *merA* signal is preserved within its dominant orthogroup and tracks cumulative Hg flux rather than instantaneous concentration, a *merA*-based biomarker could integrate exposure over the timescales relevant to microbial gene-pool dynamics, providing a biological readout that is less sensitive to instantaneous concentration spikes than direct chemical measurements.

The *chrR* compositional turnover finding indicates that Cr exposure is associated with community-level changes that can themselves be informative. Rather than a within-lineage adaptive signal, the *chrR* regime contrast reports on the identity of the *chrR*-bearing community, which shifts sharply from a diversified Control community to a Xanthobacteraceae-dominated High-runoff community. This form of biomarker signal, when interpreted correctly, is directly informative about community turnover in response to Cr exposure and complements (rather than duplicates) the gene-pool information provided by *merA*.

The *cadA* and *arsC* findings, in contrast, caution that gene-family-level statistics should not be over-interpreted in metagenomic datasets without orthology controls. The gene-family-level MEME, BUSTED-E and MK signals for *cadA* and the raw MEME fraction for *arsC* would, in the absence of orthology-aware controls, have suggested strong regime-specific selection. The within-clade analyses reveal that a substantial fraction of these signals reflects subfamily heterogeneity rather than diversification within a coherent orthogroup. For a monitoring framework, this implies that the choice of MRG target is deliberate: some genes (*merA*) behave as within-orthogroup biosensors that respond to cumulative

exposure, others (*chrR*) report on community turnover, and still others (*cadA*, *arsC*) may not respond detectably in the local exposure range or may require additional subfamily-level analyses to yield interpretable signals.

A monitoring framework based on metagenomic surveillance of these and additional MRGs, combined with the orthology- and taxonomy-aware controls and the N_e-independent selection tests applied here, could provide an early-warning system for the biological consequences of climate-driven metal mobilization in Arctic freshwater systems. Because Lake Hazen represents a natural three-regime hydrological gradient within a single basin, it offers a particularly valuable sentinel for such monitoring, and the space-for-time substitution design exploited here could be replicated across other Arctic watersheds to test the generality of the patterns documented in this study.

4.7 Limitations and replication considerations

A reviewer could reasonably object that the absence of biological replicates within each regime limits statistical inference in this study, since each regime is represented by a single soil site (CSOIL, LRSOIL or HRSOIL). We address this objection directly because it deserves an explicit response rather than a footnote. In the present analyses, the unit of statistical inference is the gene haplotype rather than the sampling site. With 60 to 1,349 unique haplotypes recovered per gene per regime, the descriptive population-level metrics computed in this thesis achieve substantial statistical resolution despite single-site sampling per regime. Bootstrap resampling of alignment columns (500 replicates) yields robust 95% CIs for π and π_N/π_S, and the Bayesian probabilities $P(\pi_A > \pi_B)$ provide quantitative directional credibility without requiring replicate sites. The McDonald-Kreitman test compares within-population polymorphism to between-population divergence and is not affected by within-regime replication. MEME, BUSTED and BUSTED-E operate on the gene phylogeny of all haplotypes and likewise do not require replicate sites.

The space-for-time substitution design at Lake Hazen also eliminates several confounding factors that would arise from comparing distinct watersheds: bedrock geology, atmospheric deposition history,

latitude, day length, ambient temperature regimes and lake-wide hydrochemistry are largely common to all three regimes within the Lake Hazen basin, so that the principal variable between regimes is the magnitude and chemistry of glacial meltwater input. The convergence of multiple independent statistical methods, each with different assumptions, provides cross-validation in the absence of spatial replicates. Nevertheless, we present all patterns as exploratory, hypothesis-generating observations rather than as formal demonstrations of population-level adaptation, and we recommend that future work replicating the design at other Arctic watersheds (Bégin et al., 2021; Cuthbertson et al., 2022) would strengthen the generality of the patterns documented here.

Several additional limitations should be acknowledged. First, the analysis covers a single watershed, and the generality of the patterns to other Arctic or temperate systems remains to be tested. Second, the small number of *cadA* haplotypes recovered in each regime (n = 19–29) limits the precision of coalescent estimates for this gene, as reflected in the broad HPD intervals reported in Table S14. Third, in-situ metal speciation and redox conditions were not measured, and our interpretation is therefore restricted to total metal pools across regimes. Fourth, taxonomic affiliation could not be determined for 76 *chrR* Low-regime haplotypes due to computational wall-clock constraints (92% overall taxonomic coverage). Their omission does not affect the Control versus High contrast that underlies our interpretation but should be addressed in future work. Fifth, the orthology controls used here (MMseqs2 elbow-threshold clustering) define operational orthogroups by sequence-identity neighbourhood rather than by formal phylogenetic orthology inference; sensitivity to alternative orthology definitions is reported in Table S11. Sixth, functional validation of regime-specific variants (particularly *chrR* variants in the High regime) was not attempted in this study but would be essential for testing the phenotypic significance of the genetic signatures inferred here.

4.8 Future directions

Several promising directions emerge from the present work. Long-read metagenomic sequencing combined with genome binning (Bickhart et al., 2022; Sereika et al., 2022) would partially resolve the host-

attribution problem documented above and clarify the extent to which MRG gene pools are structured by host phylogeny versus mobile-element dynamics. In particular, the *chrR* compositional turnover result would benefit from long-read confirmation that the Xanthobacteraceae *chrR* haplotypes dominating High are indeed encoded on Xanthobacteraceae genomes rather than on plasmids circulating across host lineages.

Functional characterization of variants under regime-specific selection, particularly the *chrR* variants identified in the High regime and the *cadA* variants under Zn stress, using heterologous expression in a common host background and biochemical assays of the encoded enzyme activities, would test the phenotypic significance of the genetic signatures inferred here. Such experiments would provide direct evidence for whether the regime-associated variants confer enhanced enzymatic activity, altered substrate specificity, or improved fitness under the corresponding metal concentrations.

Coupled metagenomic and geochemical speciation surveys at Lake Hazen and other Arctic sites would directly address the unresolved question of whether oxyanion speciation modulates MRG evolution. Such surveys should measure As partitioning between As(V) and As(III), Cr partitioning between Cr(VI) and Cr(III), and the bioavailable fractions of each, ideally with seasonal sampling to capture the dynamics of meltwater-driven metal mobilization. Coupling these geochemical measurements with metagenomic sampling of the resident microbial community would test whether *arsC* and *chrR* evolution correlates more strongly with bioavailable speciated concentrations than with total dissolved concentrations, and whether the seasonal dynamics of meltwater delivery leave detectable evolutionary imprints on the gene pools.

Finally, expanding the analysis to additional MRG families (mer operon genes beyond merA; ars operon genes beyond arsC; cop, pbr, and smt systems) and to multiple Arctic watersheds would establish how broadly the patterns documented at Lake Hazen apply across the Arctic biome. As climate-driven mobilization of previously sequestered metals accelerates across the Arctic, monitoring the diversification and sequence variation in MRG gene pools alongside chemical measurements may provide an integrative framework for assessing the biological consequences of these rapid environmental changes.

Chapter 5: Conclusion

This thesis set out to characterize the evolutionary signatures imprinted on metal resistance gene pools by the natural gradient of glacial-meltwater-derived metal exposure at Lake Hazen, the largest freshwater lake by volume in the High Arctic. Using metagenomic DNA data from soil sites along a hydrological gradient of Control, Low and High runoff, we examined four metal resistance genes (*merA*, *arsC*, *cadA* and *chrR*) with a battery of complementary analytical approaches: descriptive population-genetic statistics, N_e-independent selection tests with outgroup-sensitivity analysis, codon-based selection tests with false-discovery-rate correction and error-sink absorption, community-level phylogenetic inference, and orthology-aware and taxonomy-aware controls.

Three principal patterns emerge. First, the *merA* gene shows the strongest within-orthogroup signal in our dataset, with monotonic gene-pool diversification along the runoff gradient, McDonald-Kreitman α significant across all three regimes, and these signals preserved (indeed reinforced) within the largest orthologous cluster containing 90% of haplotypes. The diversity gradient tracks the cumulative Hg flux rather than the instantaneous concentration, consistent with sustained Hg delivery to the lake basin, a pattern consistent with gene-pool diversification. The compositional turnover of *merA*-bearing hosts along the gradient, combined with the within-orthogroup within-clade signal, is consistent with the mobility of *merA* on the Tn21 transposon family: the same orthologous gene circulates and diversifies across taxonomically diverse hosts along the runoff gradient.

Second, *cadA* and *chrR* show regime-specific signals that are partly attributable to subfamily heterogeneity or community-composition turnover. The *cadA* BUSTED-E signal collapses within the largest cluster (subfamily-driven), while the *cadA* MK signal persists in the two regimes where the largest clade is populated, indicating that gene-family-level statistics for *cadA* mix within-orthogroup diversification with inter-subfamily heterogeneity. The *chrR* regime contrast reflects a shift from a

diversified Control community to a near-monoculture of Xanthobacteraceae under high runoff: the extreme PERMANOVA R^2 (0.718), the low π in High despite high Cr, the negative Tajima's D in High, the highly significant MK α in High, and the high FDR-corrected MEME fraction are all consistent with a community-level turnover rather than within-orthogroup adaptation.

Third, *arsC* evolves near-neutrally across regimes, a result whose interpretation is constrained by the lack of in-situ As speciation data and the comparatively modest As gradient at Lake Hazen. The contrast between the strong *chrR* community-level signal and the absent *arsC* signal, despite both genes targeting oxyanion substrates, indicates that substrate speciation alone is not sufficient to predict the evolutionary trajectory of an MRG in this system: gene-specific factors, including the magnitude of the local exposure gradient, host-cell physiology, and the operon context in which each gene is embedded, must also be considered.

The reframing of population-genetic and codon-based statistics as descriptive summaries of standing variation in environmental gene pools, rather than as formal tests of adaptation in organismal populations, accommodates the realities of metagenomic MRG data, including pervasive horizontal gene transfer and the unresolved host-attribution of individual sequences. This conceptual framing, combined with orthology and taxonomy controls, supports the interpretation of the observed patterns as exploratory, hypothesis-generating observations rather than as demonstrations of population-level adaptation. The convergence of orthology-aware, taxonomy-aware, and outgroup-aware analyses on a coherent gene-by-gene story is what permits biological interpretation despite the methodological limitations of any single approach applied to metagenomic data.

The patterns documented in this thesis suggest that metal resistance gene pools may serve as candidate biomarkers of climate-driven metal mobilization in Arctic ecosystems, provided that the choice of MRG target and the interpretive framework are deliberate. Some genes (*merA*) behave as within-orthogroup biosensors that respond to cumulative exposure, others (*chrR*) report on community turnover,

and still others (*arsC*) may not respond detectably in the local exposure range. A monitoring framework that combines metagenomic surveillance of carefully chosen MRGs with orthology and taxonomy controls, together with the N_e-independent selection tests applied here, could complement traditional geochemical monitoring with biologically meaningful information about the response of resident microbial communities. Lake Hazen, as the largest High Arctic freshwater lake and as a site with a natural three-regime hydrological gradient within a single basin, offers a particularly valuable sentinel for such monitoring.

Climate change is rapidly mobilizing metals from cryospheric reservoirs that have sequestered them for millennia, and the biological consequences of this mobilization are only beginning to be understood. The gene-pool diversification of microbial metal resistance genes, as documented in this thesis, provides one window onto those consequences. They show that the gene pool is neither static nor uniformly responsive: it is structured, gene-specific, and sensitive to the local intensity of metal exposure and to the composition of the resident community. As cryospheric metal mobilization accelerates over the coming decades, continued monitoring and interpretation of these patterns of gene-pool diversification will be essential for assessing the integrative biological consequences of a rapidly changing Arctic.

Statement on the Use of Generative AI

During the preparation of this thesis, the author used a generative AI assistant (Anthropic's Claude) to support language editing, formatting, and typographic consistency. All research design, data generation and analysis, interpretation of results, and conclusions are the author's own. The author reviewed and edited all AI-assisted output and takes full responsibility for the content of this thesis.

References

Ackerley, D. F., Gonzalez, C. F., Park, C. H., Blake, R., Keyhan, M., and Matin, A. (2004). Chromate-reducing properties of soluble flavoproteins from Pseudomonas putida and Escherichia coli. Applied and Environmental Microbiology, 70(2), 873-882. doi: 10.1128/AEM.70.2.873-882.2004

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25(17), 3389-3402. doi: 10.1093/nar/25.17.3389

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Argüello, J. M., Eren, E., and González-Guerrero, M. (2007). The structure and function of heavy metal transport P1B-ATPases. BioMetals, 20(3-4), 233-248. doi: 10.1007/s10534-006-9055-6

Barkay, T., Miller, S. M., and Summers, A. O. (2003). Bacterial mercury resistance from atoms to ecosystems. FEMS Microbiology Reviews, 27(2-3), 355-384. doi: 10.1016/S0168-6445(03)00046-9

Bégin, P. N., Tanabe, Y., Kumagai, M., Culley, A. I., Paquette, M., Sarrazin, D., Uchida, M., and Vincent, W. F. (2021). Extreme warming and regime shift toward amplified variability in a far northern lake. Limnology and Oceanography, 66(S1), S17-S29. doi: 10.1002/lno.11546

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289-300.

Bickhart, D. M., Kolmogorov, M., Tseng, E., Portik, D. M., Korobeynikov, A., Tolstoganov, I., et al. (2022). Generation of lineage-resolved complete metagenome-assembled genomes by precision phasing. Nature Biotechnology, 40(5), 711-719. doi: 10.1038/s41587-021-01130-z

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120. doi: 10.1093/bioinformatics/btu170

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Computational Biology, 15(4), e1006650. doi: 10.1371/journal.pcbi.1006650

Box, J. E., Colgan, W. T., Christensen, T. R., Schmidt, N. M., Lund, M., Parmentier, F. J. W., et al. (2019). Key indicators of Arctic climate change: 1971-2017. Environmental Research Letters, 14(4), 045010. doi: 10.1088/1748-9326/aafc1b

Bruins, M. R., Kapil, S., and Oehme, F. W. (2000). Microbial resistance to metals in the environment. Ecotoxicology and Environmental Safety, 45(3), 198-207. doi: 10.1006/eesa.1999.1860

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12(1), 59-60. doi: 10.1038/nmeth.3176

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics, 10(3), 195-205. doi: 10.1038/nrg2526

Christakis, C. A., Barkay, T., and Boyd, E. S. (2021). Expanded diversity and phylogeny of mer genes broadens mercury resistance paradigms and reveals an origin for MerA among thermophilic Archaea. Applied and Environmental Microbiology, 87, e00543-21. doi: 10.1128/AEM.01431-18

Christakis, C. A., Polz, M. F., Barkay, T., and Boyd, E. S. (2022). Experimental evolution of bacterial mercury resistance under environmentally realistic exposure. Applied and Environmental Microbiology, 88, e02387-21. doi: 10.1128/AEM.02387-18

Colby, G. A., Ruuskanen, M. O., St. Pierre, K. A., St. Louis, V. L., Poulain, A. J., and Aris-Brosou, S. (2020). Warming climate is reducing the diversity of dominant microbes in the largest High Arctic lake. Frontiers in Microbiology, 11, 561194. doi: 10.3389/fmicb.2020.561194

Cuthbertson, L., Amores-Arrocha, H., Malard, L. A., Els, N., Sattler, B., and Pearce, D. A. (2022). Characterisation of Arctic bacterial communities in the air above Svalbard. Biology, 11(8), 1247. doi: 10.3390/biology11081247

Driscoll, C. T., Mason, R. P., Chan, H. M., Jacob, D. J., and Pirrone, N. (2013). Mercury as a global pollutant: Sources, pathways, and effects. Environmental Science and Technology, 47(10), 4967-4983. doi: 10.1021/es305071v

Ellegaard, K. M., and Engel, P. (2016). Beyond 16S rRNA community profiling: intra-species diversity in the gut microbiota. Frontiers in Microbiology, 7, 1475. doi: 10.3389/fmicb.2016.01475

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. Biological Conservation, 61(1), 1-10. doi: 10.1016/0006-3207(92)91201-3

Fox, B., and Walsh, C. T. (1983). Mercuric reductase: purification and characterization of a transposon-encoded flavoprotein containing an oxidation-reduction-active disulfide. Journal of Biological Chemistry, 257, 2498-2503.

Gaballa, A., and Helmann, J. D. (2003). Bacillus subtilis CPx-type ATPases: characterization of Cd, Zn, Co and Cu efflux systems. BioMetals, 16(4), 497-505. doi: 10.1023/A:1023425321617

Garud, N. R., and Pollard, K. S. (2020). Population genetics in the human microbiome. Trends in Genetics, 36(1), 53-67. doi: 10.1016/j.tig.2019.10.010

Hartl, D. L., and Clark, A. G. (2007). Principles of Population Genetics, 4th edition. Sunderland, MA: Sinauer Associates.

Hawkings, J. R., Wadham, J. L., Tranter, M., Raiswell, R., Benning, L. G., Statham, P. J., et al. (2014). Ice sheets as a significant source of highly reactive nanoparticulate iron to the oceans. Nature Communications, 5, 3929. doi: 10.1038/ncomms4929

Hemme, C. L., Deng, Y., Gentry, T. J., Fields, M. W., Wu, L., Barua, S., et al. (2010). Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. ISME Journal, 4(5), 660-672. doi: 10.1038/ismej.2009.154

Hudson, R. R. (1991). Gene genealogies and the coalescent process. In: Futuyma, D. and Antonovics, J. (eds.), Oxford Surveys in Evolutionary Biology, 7, 1-44.

IPCC (2021). Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution, 30(4), 772-780. doi: 10.1093/molbev/mst010

Kingman, J. F. C. (1982). The coalescent. Stochastic Processes and their Applications, 13(3), 235-248. doi: 10.1016/0304-4149(82)90011-4

Kosakovsky Pond, S. L., Poon, A. F. Y., Velazquez, R., Weaver, S., Hepler, N. L., Murrell, B., et al. (2020). HyPhy 2.5 — a customizable platform for evolutionary hypothesis testing using phylogenies. Molecular Biology and Evolution, 37(1), 295-299. doi: 10.1093/molbev/msz197

Lehnherr, I., St. Louis, V. L., Sharp, M., Gardner, A. S., Smol, J. P., Schiff, S. L., et al. (2018). The world's largest High Arctic lake responds rapidly to climate warming. Nature Communications, 9, 1290. doi: 10.1038/s41467-018-03685-z

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658-1659. doi: 10.1093/bioinformatics/btl158

Liebert, C. A., Hall, R. M., and Summers, A. O. (1999). Transposon Tn21, flagship of the floating genome. Microbiology and Molecular Biology Reviews, 63(3), 507-522. doi: 10.1128/MMBR.63.3.507-522.1999

Liu, J., Gladysheva, T. B., Lee, L., and Rosen, B. P. (1995). Identification of an essential cysteinyl residue in the ArsC arsenate reductase of plasmid R773. Biochemistry, 34(41), 13472-13476. doi: 10.1021/bi00041a026

McDonald, J. H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature, 351(6328), 652-654. doi: 10.1038/351652a0

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution, 37(5), 1530-1534. doi: 10.1093/molbev/msaa015

Mukhopadhyay, R., Rosen, B. P., Phung, L. T., and Silver, S. (2002). Microbial arsenic: from geocycles to genes and enzymes. FEMS Microbiology Reviews, 26(3), 311-325. doi: 10.1111/j.1574-6976.2002.tb00617.x

Müller, N. F., Rasmussen, D., and Stadler, T. (2018). MASCOT: parameter and state inference under the marginal structured coalescent approximation. Bioinformatics, 34(22), 3843-3848. doi: 10.1093/bioinformatics/bty406

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. PLOS Genetics, 8(7), e1002764. doi: 10.1371/journal.pgen.1002764

Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., et al. (2015). Gene-wide identification of episodic selection. Molecular Biology and Evolution, 32(5), 1365-1371. doi: 10.1093/molbev/msv035

Nei, M. (1987). Molecular Evolutionary Genetics. New York: Columbia University Press.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution, 3(5), 418-426. doi: 10.1093/oxfordjournals.molbev.a040410

Nies, D. H. (1999). Microbial heavy-metal resistance. Applied Microbiology and Biotechnology, 51(6), 730-750. doi: 10.1007/s002530051457

Nies, D. H. (2003). Efflux-mediated heavy metal resistance in prokaryotes. FEMS Microbiology Reviews, 27(2-3), 313-339. doi: 10.1016/S0168-6445(03)00048-2

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature, 405(6784), 299-304. doi: 10.1038/35012500

Park, C. H., Keyhan, M., Wielinga, B., Fendorf, S., and Matin, A. (2000). Purification to homogeneity and characterization of a novel Pseudomonas putida chromate reductase. Applied and Environmental Microbiology, 66(5), 1788-1795. doi: 10.1128/AEM.66.5.1788-1795.2000

Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., and Delsuc, F. (2018). MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. Molecular Biology and Evolution, 35(10), 2582-2584. doi: 10.1093/molbev/msy159

Rantanen, M., Karpechko, A. Y., Lipponen, A., Nordling, K., Hyvärinen, O., Ruosteenoja, K., et al. (2022). The Arctic has warmed nearly four times faster than the globe since 1979. Communications Earth and Environment, 3, 168. doi: 10.1038/s43247-022-00498-3

Schuster, P. F., Schaefer, K. M., Aiken, G. R., Antweiler, R. C., Dewild, J. F., Gryziec, J. D., et al. (2018). Permafrost stores a globally significant amount of mercury. Geophysical Research Letters, 45(3), 1463-1471. doi: 10.1002/2017GL075571

Selberg, A., Clark, N. L., Sackton, T. B., Muse, S. V., Lucaci, A. G., Weaver, S., Nekrutenko, A., Chikina, M., and Kosakovsky Pond, S. L. (2025). Minus the error: Testing for positive selection in the presence of residual alignment errors. eLife, 14, e106921. doi: 10.7554/eLife.106921.1

Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sørensen, E. A., Wollenberg, R. D., and Albertsen, M. (2022). Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nature Methods, 19(7), 823-826. doi: 10.1038/s41592-022-01539-7

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLOS ONE, 11(10), e0163962. doi: 10.1371/journal.pone.0163962

Silver, S., and Phung, L. T. (1996). Bacterial heavy metal resistance: new surprises. Annual Review of Microbiology, 50, 753-789. doi: 10.1146/annurev.micro.50.1.753

Smith, N. G. C., and Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. Nature, 415(6875), 1022-1024. doi: 10.1038/4151022a

Smith, W. P. J., Wucher, B. R., Nadell, C. D., and Foster, K. R. (2019). Bacterial defences: mechanisms, evolution and antimicrobial resistance. Nature Reviews Microbiology, 21, 519-534. doi: 10.1038/s41579-023-00877-3

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics, 21(7), 951-960. doi: 10.1093/bioinformatics/bti125

Soucy, S. M., Huang, J., and Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. Nature Reviews Genetics, 16(8), 472-482. doi: 10.1038/nrg3962

St. Pierre, K. A., St. Louis, V. L., Lehnherr, I., Gardner, A. S., Serbu, J. A., Mortimer, C. A., et al. (2019). Drivers of mercury cycling in the rapidly changing glacierized watershed of the High Arctic's largest lake by volume (Lake Hazen, Nunavut, Canada). Environmental Science and Technology, 53(3), 1175-1185. doi: 10.1021/acs.est.8b05926

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology, 35(11), 1026-1028. doi: 10.1038/nbt.3988

Stibal, M., Bradley, J. A., Edwards, A., Hotaling, S., Zawierucha, K., Rosvold, J., et al. (2020). Glacial ecosystems are essential to understanding biodiversity responses to glacier retreat. Nature Ecology and Evolution, 4(5), 686-687. doi: 10.1038/s41559-020-1163-0

Sukumaran, J., and Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. Bioinformatics, 26(12), 1569-1571. doi: 10.1093/bioinformatics/btq228

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research, 34(suppl_2), W609-W612. doi: 10.1093/nar/gkl315

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123(3), 585-595. doi: 10.1093/genetics/123.3.585

Yang, H. C., Fu, H. L., Lin, Y. F., and Rosen, B. P. (2012). Pathways of arsenic uptake and efflux. Current Topics in Membranes, 69, 325-358. doi: 10.1016/B978-0-12-394390-3.00012-4

Annex A - Supplementary Tables

Table S1. Nucleotide diversity (π) per gene and individual sampling site, computed by the Nei pairwise method with 500 bootstrap replicates for 95% confidence intervals. Sites grouped by hydrological regime and sample type (SED = sediment, SOIL = soil).

Gene	Site	Regime	Type	N	π [95% CI]
merA	CSOIL	C	Soil	60	0.0444 [0.0359–0.0538]
	LRSOIL	L	Soil	99	0.0691 [0.0471–0.0965]
	H1	H	Sed	112	0.1840 [0.1634–0.2039]
	HRSOIL	H	Soil	83	0.0888 [0.0636–0.1247]
arsC	CSED	C	Sed	10	0.3347 [0.2474–0.3876]
	CSOIL	C	Soil	550	0.1903 [0.1814–0.1997]
	L1	L	Sed	56	0.3193 [0.2893–0.3400]
	L2	L	Sed	135	0.3092 [0.2957–0.3246]
	LRSOIL	L	Soil	276	0.2372 [0.2233–0.2500]
	H1	H	Sed	238	0.2522 [0.2421–0.2621]
	H2	H	Sed	71	0.2750 [0.2477–0.2989]
	HRSOIL	H	Soil	523	0.2028 [0.1935–0.2109]
cadA	CSOIL	C	Soil	19	0.3103 [0.1979–0.3891]
	L1	L	Sed	6	0.2620 [0.1332–0.3444]
	L2	L	Sed	10	0.2463 [0.0946–0.3518]
	LRSOIL	L	Soil	29	0.2962 [0.2101–0.3701]
	H1	H	Sed	15	0.3559 [0.2608–0.4066]
	H2	H	Sed	5	0.2924 [0.0040–0.4132]
	HRSOIL	H	Soil	26	0.4234 [0.3939–0.4414]
chrR	CSED	C	Sed	73	0.1223 [0.1053–0.1388]
	CSOIL	C	Soil	726	0.1499 [0.1433–0.1567]
	L1	L	Sed	44	0.2317 [0.1879–0.2676]
	L2	L	Sed	217	0.1501 [0.1362–0.1658]
	LRSOIL	L	Soil	76	0.2399 [0.2178–0.2646]
	H1	H	Sed	327	0.1999 [0.1826–0.2163]
	H2	H	Sed	150	0.2397 [0.2186–0.2610]
	HRSOIL	H	Soil	319	0.1301 [0.1182–0.1426]

Table S2. Tajima's *D* statistics for soil-only haplotypes with significance levels. * $p < 0.05$.

Gene	Regime	N haplotypes	Tajima's *D*	*p*-value	Significance
merA	Control	60	−1.065	n.s.	ns
	Low	99	−1.675	< 0.05	*
	High	83	−1.630	< 0.05	*
arsC	Control	550	+1.077	n.s.	ns
	Low	276	+1.900	< 0.05	*
	High	523	+1.353	n.s.	ns
cadA	Control	19	+0.244	n.s.	ns
	Low	29	+0.187	n.s.	ns
	High	26	−0.167	n.s.	ns
chrR	Control	726	+0.199	n.s.	ns
	Low	76	+0.785	n.s.	ns
	High	319	−0.417	n.s.	ns

Table S3. Rarefaction/subsampling of π to a common haplotype depth (n = 19, the smallest gene-by-regime sample; *cadA* Control). For each gene-by-regime cell, we report π computed on the dereplicated full data, the subsampled π mean and 95% bootstrap CI from 500 replicates of size n = 19, and whether the rank direction relative to the previous regime is preserved.

Gene	Regime	n full	π (full)	π (subsampled) [95% CI]	Rank direction preserved
merA	C	60	0.044	0.044 [0.033–0.057]	—
	L	99	0.069	0.068 [0.048–0.094]	C < L (Yes)
	H	83	0.089	0.088 [0.062–0.122]	L < H (Yes)
arsC	C	550	0.190	0.188 [0.169–0.209]	—
	L	276	0.237	0.236 [0.213–0.260]	C < L (Yes)
	H	523	0.203	0.202 [0.183–0.223]	L > H (Yes)
cadA	C	19	0.310	0.310 [0.198–0.389]	—
	L	29	0.296	0.288 [0.180–0.376]	C ~ L (overlap)
	H	26	0.423	0.420 [0.388–0.443]	L < H (Yes)
chrR	C	726	0.150	0.149 [0.139–0.160]	—
	L	76	0.240	0.239 [0.215–0.267]	C < L (Yes)
	H	319	0.130	0.130 [0.117–0.146]	L > H (Yes)

Table S4. Nucleotide diversity at non-synonymous (π_N) and synonymous (π_S) sites computed using the Nei-Gojobori (1986) method, with 95% bootstrap confidence intervals (50 resamples) on the π_N/π_S ratio. Ratios below 1 indicate purifying selection. Independent of N_e estimation.

Gene	Regime	π_N	π_S	π_N/π_S	95% CI	# codons
merA	Control	0.0181	0.0825	0.220	[0.192–0.237]	478
	Low	0.0212	0.0867	0.245	[0.210–0.277]	478
	High	0.0288	0.1082	0.266	[0.228–0.323]	478
arsC	Control	0.0805	0.2386	0.337	[0.325–0.348]	142
	Low	0.1078	0.2403	0.449	[0.421–0.476]	142
	High	0.0939	0.2079	0.452	[0.432–0.467]	144
cadA	Control	0.1024	0.1182	0.866	[0.804–0.982]	462
	Low	0.1185	0.1384	0.856	[0.757–0.903]	389
	High	0.0802	0.1322	0.607	[0.521–0.699]	627
chrR	Control	0.0667	0.1158	0.576	[0.554–0.589]	194
	Low	0.0962	0.2117	0.454	[0.422–0.496]	190
	High	0.0530	0.0843	0.630	[0.596–0.661]	196

Table S5. McDonald-Kreitman test (1991) results per gene and regime. PN/PS: intra-regime polymorphisms. D_N/D_S: fixed differences vs outgroup. NI = $(P_N/P_S)/(D_N/D_S)$. $\alpha = 1 - NI$ (Smith & Eyre-Walker, 2002) = proportion of adaptive substitutions. Significance from χ^2 test with Yates correction (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant). Independent of N_e estimation.

Gene	Regime	P_N	P_S	D_N	D_S	χ^2	*p*	NI	*α*	Sig.
merA	Control	123	206	168	144	16.85	< 0.0001	0.512	+0.488	***
	Low	221	319	159	150	8.39	0.0038	0.654	+0.346	**
	High	295	373	163	140	7.38	0.0066	0.679	+0.321	**
arsC	Control	130	140	37	55	1.43	0.2314	1.380	−0.380	ns
	Low	126	139	42	46	0.00	1.0000	0.993	+0.007	ns
	High	125	138	37	56	1.36	0.2429	1.371	−0.371	ns
cadA	Control	289	274	278	123	30.57	< 0.0001	0.467	+0.533	***
	Low	316	323	251	90	52.22	< 0.0001	0.351	+0.649	***
	High	394	473	330	183	45.33	< 0.0001	0.462	+0.538	***
chrR	Control	167	176	71	60	0.94	0.3319	0.802	+0.198	ns
	Low	153	153	71	59	0.60	0.4370	0.831	+0.169	ns
	High	138	141	89	58	4.32	0.0378	0.638	+0.362	*

Table S6. Outgroup sensitivity of the McDonald-Kreitman α using MACSE codon alignment (Ranwez et al., 2018). For each gene, we report α per regime computed with the canonical outgroup (manuscript original) and with one to two additional BLAST-derived outgroups in distinct amino-acid identity bins: close (~65–72%), intermediate (~55–65%), and distant (~40–55%). Across the 33 outgroup-by-regime cells, the sign of α was preserved in 9 of 12 gene-by-regime combinations (75%), and statistical significance at $p < 0.05$ was preserved in 6 of 12 (50%).

Gene	Regime	Canonical α	Close α	Intermediate α	Distant α	Sign preserved
merA	C	+0.488	+0.269	+0.373	—	Yes
	L	+0.346	+0.170	+0.262	—	Yes
	H	+0.321	+0.069	+0.198	—	Yes
arsC	C	−0.380	−0.016	—	—	Yes (negative)
	L	+0.007	−0.173	—	—	Mixed
	H	−0.371	+0.119	—	—	No
cadA	C	+0.533	—	+0.584	+0.584	Yes
	L	+0.649	—	+0.656	+0.634	Yes
	H	+0.538	—	+0.143	+0.265	Yes
chrR	C	+0.198	−0.043	+0.431	—	Mixed
	L	+0.169	+0.323	+0.569	—	Yes
	H	+0.362	+0.403	+0.536	—	Yes

Table S7. Detailed MEME results per gene at the gene-family level: number and percentage of codons under episodic positive selection at three thresholds (raw $p < 0.1$, raw $p < 0.05$, and Benjamini-Hochberg FDR-corrected $q < 0.05$). The FDR-corrected fractions are the values reported in the main text and visualized in Figure 8.

Gene	Total codons	Raw p<0.1 (n, %)	Raw p<0.05 (n, %)	FDR q<0.05 (n, %)
merA	574	48 (8.4%)	23 (4.0%)	20 (3.5%)
arsC	162	86 (53.1%)	41 (25.3%)	79 (48.8%)
cadA	727	41 (5.6%)	12 (1.7%)	10 (1.4%)
chrR	222	125 (56.3%)	67 (30.2%)	121 (54.5%)

Table S8. Faith's phylogenetic diversity (PD) per gene and regime, computed as the total branch length of the soil-only IQ-TREE phylogeny (substitutions per site).

Gene	Regime	N haplotypes	Faith's PD (subs/site)
merA	Control	60	0.7400
	Low	99	4.6690
	High	83	7.9180
arsC	Control	550	33.1910
	Low	276	22.0060
	High	523	35.5330
cadA	Control	19	5.1950
	Low	29	7.9410
	High	26	7.7010
chrR	Control	726	58.5650
	Low	76	13.7260
	High	319	38.6300

Table S9. PERMANOVA on patristic distance matrices from the gene-specific IQ-TREE phylogenies, testing the effect of hydrological regime on phylogenetic structure (9,999 permutations of regime labels). R^2 is the proportion of variance explained by the regime.

Gene	R^2	*p*-value	Permutations	Effect interpretation
merA	0.064	< 0.001	9999	Weak (< 0.2)
arsC	0.232	< 0.001	9999	Moderate (0.2–0.5)
cadA	0.154	< 0.001	9999	Weak (< 0.2)
chrR	0.718	< 0.001	9999	Strong (> 0.5)

Table S10. Metal concentrations across hydrological regimes at Lake Hazen (mean ± SD). All values from St. Pierre et al. (2019). H/C fold-change = ratio of concentration in High vs Control regime; n.d. = not determinable.

Metal	Control	Low (Blister)	High (SG + Abbé)	Unit	H/C fold-change
Hg	0.42	17.56 ± 14.14	7.07 ± 1.27	ng/L	16.6×
Cr	0.33	6.15 ± 4.92	19.73 ± 2.24	µg/L	60.2×
As	< 0.5	1.16 ± 0.98	2.69 ± 0.31	µg/L	n.d.
Zn	< 0.5	44.77 ± 46.98	108.07 ± 12.02	µg/L	n.d.

Table S11. MMseqs2 ortholog clustering of MRG haplotypes at multiple sequence-identity thresholds. For each gene, we report the number of clusters and the size of the largest cluster at three reference thresholds (50%, 70%, 90% sequence identity) and at the data-driven elbow threshold determined from the per-gene identity distribution. The elbow threshold is the value retained for the within-clade re-analyses reported in the main text.

Gene	Identity threshold	Clusters	Largest cluster (n, %)
merA	0.50	3	238 / 242 (98.3%)
	0.70	5	230 / 242 (95.0%)
	0.83 (elbow)	10	217 / 242 (89.7%)
	0.90	18	189 / 242 (78.1%)
arsC	0.50	12	1198 / 1349 (88.8%)
	0.70	39	858 / 1349 (63.6%)
	0.79 (elbow)	83	540 / 1349 (40.0%)
	0.90	201	232 / 1349 (17.2%)
cadA	0.50	4	61 / 74 (82.4%)
	0.70 (elbow)	15	16 / 74 (21.6%)
	0.90	42	5 / 74 (6.8%)
chrR	0.50	5	1054 / 1121 (94.0%)
	0.73 (elbow)	42	503 / 1121 (44.9%)
	0.90	187	143 / 1121 (12.8%)

Table S12. Within-clade selection statistics. For each gene, all selection analyses were repeated, restricted to the largest MMseqs2 cluster at the elbow threshold (Table S11). π within-clade, Tajima's D within-clade and α(MK) within-clade are reported per regime; BUSTED-E LRT and MEME % sites (FDR $q < 0.05$) are gene-wide and pool all regimes within the largest cluster. n = 0 indicates that the largest cluster contains no haplotypes from that regime; n.c. = not computable due to insufficient sample size.

Gene	π within-clade C / L / H	Tajima's *D* within-clade	α(MK) within-clade C / L / H	BUSTED-E LRT	MEME %
merA	0.047 / 0.037 / 0.050	−0.78 / −1.43 / −1.40	+0.47 / +0.50 / +0.51	189.92 ($p < 0.001$)	4.2%
arsC	0.166 / 0.149 / 0.158	+0.66 / +0.71 / +0.81	−0.37 / +0.19 / −0.16	1700.95 ($p < 0.001$)	40.4%
cadA	n = 0 / 0.065 / 0.243	n = 0 / n.c. / n.c.	n = 0 / +0.71 / +0.58	0.73 ($p = 0.347$, ns)	0.0%
chrR	0.156 / 0.118 / n = 0	n.c. / n.c. / n = 0	+0.18 / +0.41 / n = 0	655.95 ($p < 0.001$)	31.8%

Table S13. Taxonomic composition of MRG-containing haplotypes per gene and regime, aggregated to the order level. Each cell reports the percentage of haplotypes affiliated with the given taxonomic order in the given regime, based on best-hit BLASTP against NCBI nr (*E*-value < 10^{-20}). (—) indicates < 1% representation or absence. *chrR* Low: not assigned (76 haplotypes not analyzed due to wall-clock constraints; see Methods).

merA (order-level affiliation):

Order	Control (n=60)	Low (n=99)	High (n=83)
Chromatiales	77%	29%	16%
Pseudomonadota (other)	17%	63%	33%
Gammaproteobacteria (other)	7%	8%	47%
Acidobacteriota	—	—	4%
Actinomycetota	—	—	1%

arsC (order-level affiliation):

Order	Control (n=550)	Low (n=180)	High (n=523)
Burkholderiales	50%	28%	47%
Hyphomicrobiales	17%	45%	6%
Rhodospirillales	12%	—	38%
Sphingomonadales	3%	5%	5%
Unclassified	9%	7%	—
Other	9%	15%	4%

cadA (order-level affiliation):

Order	Control (n=19)	Low (n=29)	High (n=26)
Unclassified	47%	—	12%
Acidobacteriota	16%	24%	15%
Actinomycetota	16%	7%	4%
Bacteroidota	16%	3%	38%
Gammaproteobacteria (other)	—	21%	—
Hyphomicrobiales	—	21%	—
Verrucomicrobiota	—	7%	—
Chloroflexota	—	—	8%
Other	5%	17%	23%

chrR (order-level affiliation; Low not assigned):

Order	Control (n=726)	High (n=302)
Hyphomicrobiales	80%	93%
Pseudomonadota (other)	18%	—
Alphaproteobacteria (other)	—	5%
Rhodospirillales	—	1%
Other	1%	1%

Table S14. Detailed BEAST2 posterior estimates of gene-pool diversity index ($\theta = 2N_e\mu$) per gene and regime, with 95% highest posterior density (HPD) interval and CI width. Interpreted as a descriptive index of gene-pool diversity rather than as an estimate of organismal effective population size.

Gene	Regime	N haps	θ median	95% HPD lower	95% HPD upper	CI width
merA	Control	60	0.1080	0.0780	0.1400	0.0620
	Low	99	0.2410	0.1890	0.2950	0.1060
	High	83	0.3380	0.2610	0.4190	0.1580
arsC	Control	550	1.4800	1.1590	1.8150	0.6560
	Low	276	1.5430	1.2210	1.8830	0.6620
	High	523	1.5410	1.1980	1.9030	0.7050
cadA	Control	19	0.4270	0.2400	0.6510	0.4110
	Low	29	1.1290	0.6990	1.6280	0.9290
	High	26	0.9980	0.6230	1.4190	0.7960
chrR	Control	726	1.7880	1.3860	2.2110	0.8250
	Low	76	1.1670	0.8740	1.4550	0.5810
	High	319	2.4170	1.8140	2.9670	1.1530

Table S15. Detailed MASCOT structured coalescent inter-regime migration rates per gene, with 95% HPD intervals. Migration rates are in events per lineage per coalescent time unit. Interpreted as a descriptive index of gene-pool exchange rather than as an estimate of organismal demographic migration.

Gene	From	To	Median rate	95% HPD lower	95% HPD upper
merA	C	L	0.632	0.000	1.981
	L	C	1.146	0.101	2.405
	C	H	0.528	0.000	1.662
	H	C	0.904	0.001	1.921
	L	H	0.464	0.000	1.449
	H	L	0.624	0.000	1.571
arsC	C	L	0.324	0.000	0.819
	L	C	0.174	0.000	0.546
	C	H	0.437	0.000	1.044
	H	C	0.188	0.000	0.593
	L	H	0.179	0.000	0.561
	H	L	0.259	0.000	0.657
cadA	C	L	1.306	0.000	3.427
	L	C	1.448	0.000	3.410
	C	H	1.219	0.000	3.132
	H	C	1.132	0.000	3.141
	L	H	1.109	0.000	2.894
	H	L	0.824	0.000	2.709
chrR	C	L	0.268	0.000	0.629
	L	C	0.139	0.000	0.428
	C	H	0.123	0.000	0.368
	H	C	0.106	0.000	0.319
	L	H	0.168	0.000	0.490
	H	L	0.310	0.012	0.701

Annex B - Supplementary Figures

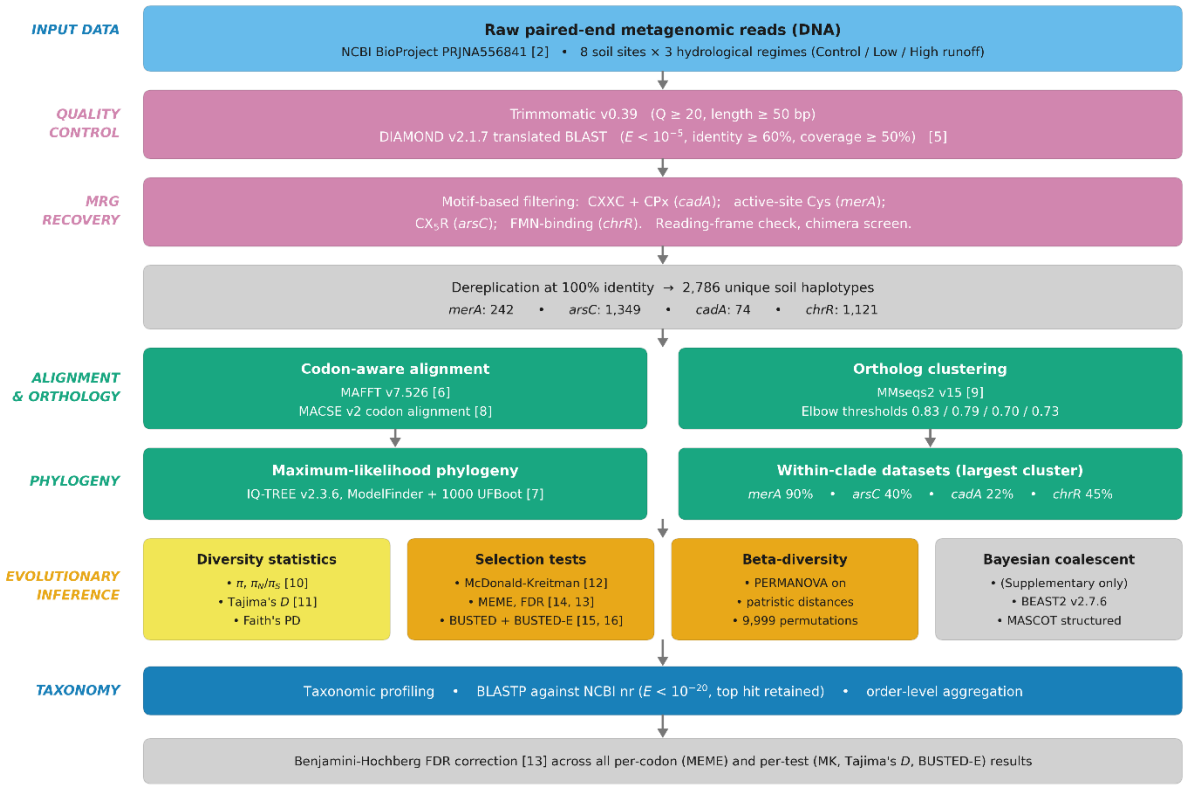


Figure S1. Analysis workflow diagram. Schematic representation of the bioinformatic and statistical pipeline used in this thesis, from raw metagenomic DNA reads (NCBI BioProject PRJNA556841) through quality control, MRG recovery via translated BLAST and motif filtering, dereplication, codon-aware alignment with MAFFT, ortholog clustering with MMseqs2, phylogenetic inference with IQ-TREE, population-genetic statistics (π, Tajima's D, π_N/π_S, McDonald-Kreitman), codon-based selection inference (MEME, BUSTED, BUSTED-E), phylogenetic beta-diversity (Faith's PD, PERMANOVA), gene-pool diversity and exchange (BEAST2, MASCOT), and taxonomic profiling (BLASTP against NCBI nr).

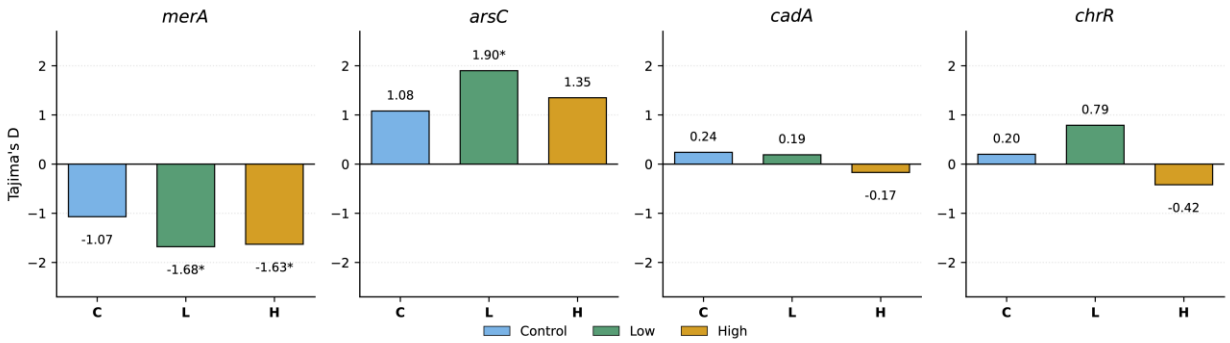


Figure S2. Tajima's *D* per gene and regime, computed on soil-only haplotypes. Bars: regime mean; numeric labels: *D* value with significance asterisks (* $p < 0.05$). Horizontal dashed lines at $D = \pm 2$ indicate the conventional thresholds for departure from neutrality. Wong palette: Control = sky blue, Low = bluish green, High = orange. Numerical values are reported in Table S2.

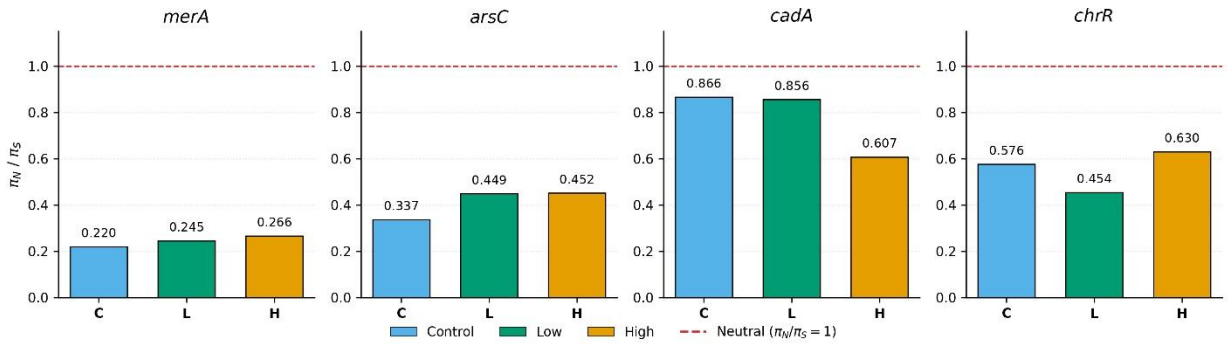


Figure S3. Non-synonymous to synonymous nucleotide diversity ratio (π_N/π_S) per gene and regime, computed with the Nei-Gojobori (1986) method. Error bars: 95% bootstrap confidence intervals (50 resamples). Red dashed line at $\pi_N/\pi_S = 1$ indicates neutrality; values below 1 indicate purifying selection. Wong palette: Control = sky blue, Low = bluish green, High = orange. Numerical values are reported in Table S9.

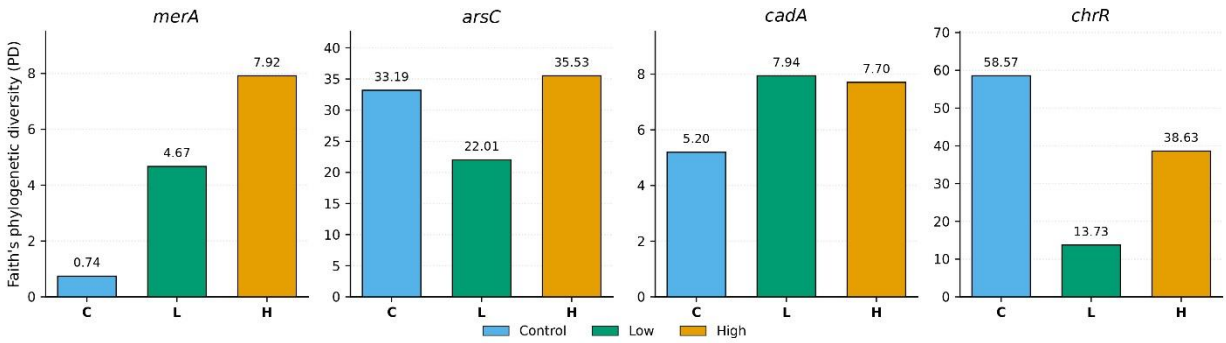


Figure S4. Faith's phylogenetic diversity (PD) per gene and regime, computed as the total branch length of the soil-only IQ-TREE phylogeny (substitutions per site). Bars: regime PD; numeric labels: PD value. Wong palette: Control = sky blue, Low = bluish green, High = orange. Numerical values are reported in Table S4.

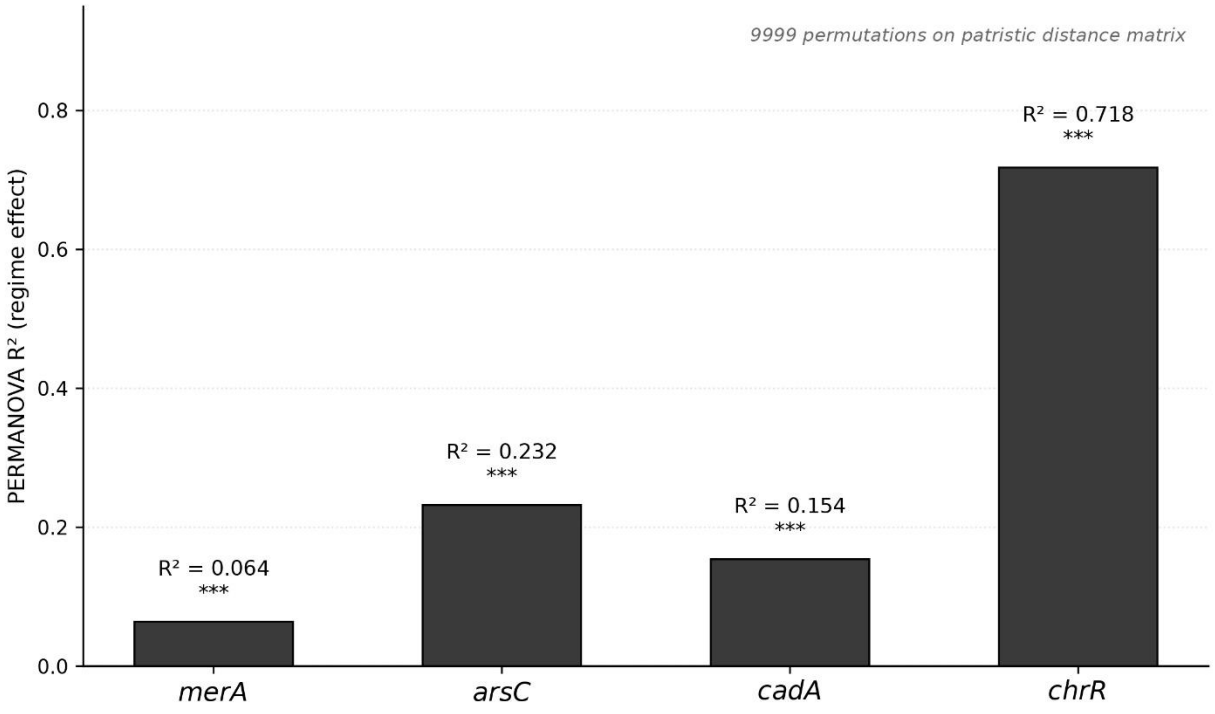


Figure S5. PERMANOVA R^2 (proportion of variance explained by regime) per gene, with significance from 9,999 permutations of regime labels on patristic distance matrices computed from the IQ-TREE phylogeny. Significance: *** $p < 0.001$. Numerical values are reported in Table S7.

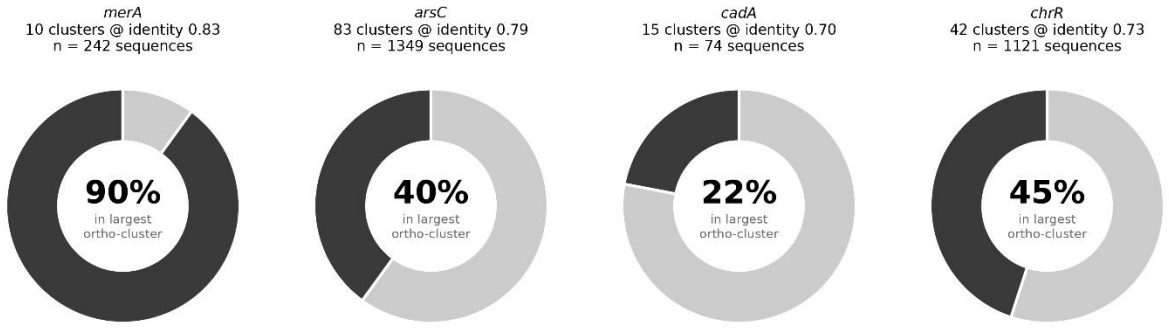


Figure S6. MMseqs2 ortholog cluster summary per gene, displayed as donut charts. The percentage of haplotypes in the largest cluster at the data-driven elbow threshold is indicated in the centre of each donut; the total number of clusters and the elbow threshold are indicated above each gene panel. Numerical values are reported in Table S11.

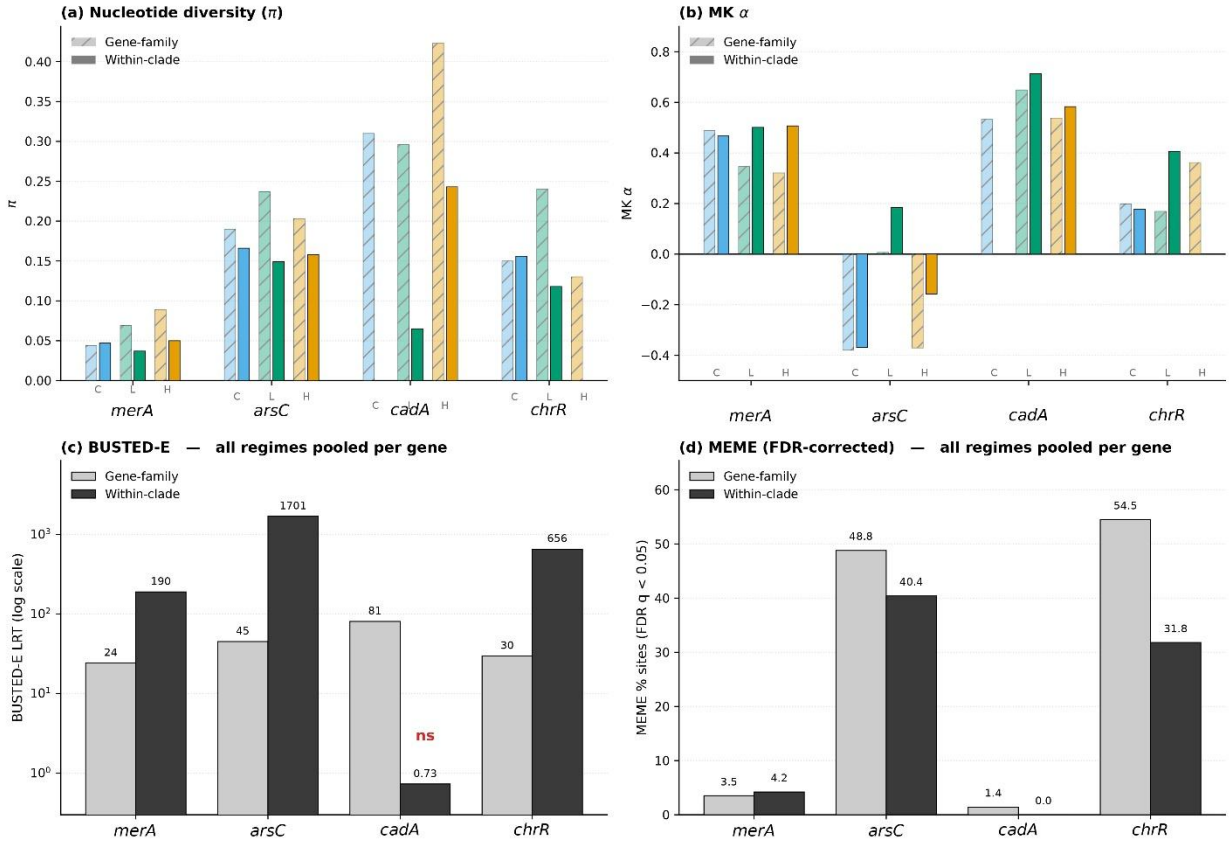


Figure S7. Gene-family vs within-clade comparison across four metrics: (a) π by gene and regime; (b) McDonald-Kreitman α by gene and regime; (c) BUSTED-E LRT (log scale) by gene; (d) MEME % sites under positive selection (FDR q < 0.05) by gene. Hatched bars: gene-family analysis (all haplotypes); solid bars: within-clade analysis (largest MMseqs2 ortholog cluster only, Table S11). For panels (a) and (b), micro-labels C/L/H below each pair of bars identify the regime within each gene. For panels (c) and (d), values pool all regimes per gene. The ns annotation in panel (c) indicates non-significance of the within-clade BUSTED-E test for *cadA* (LRT = 0.73, $p = 0.347$).

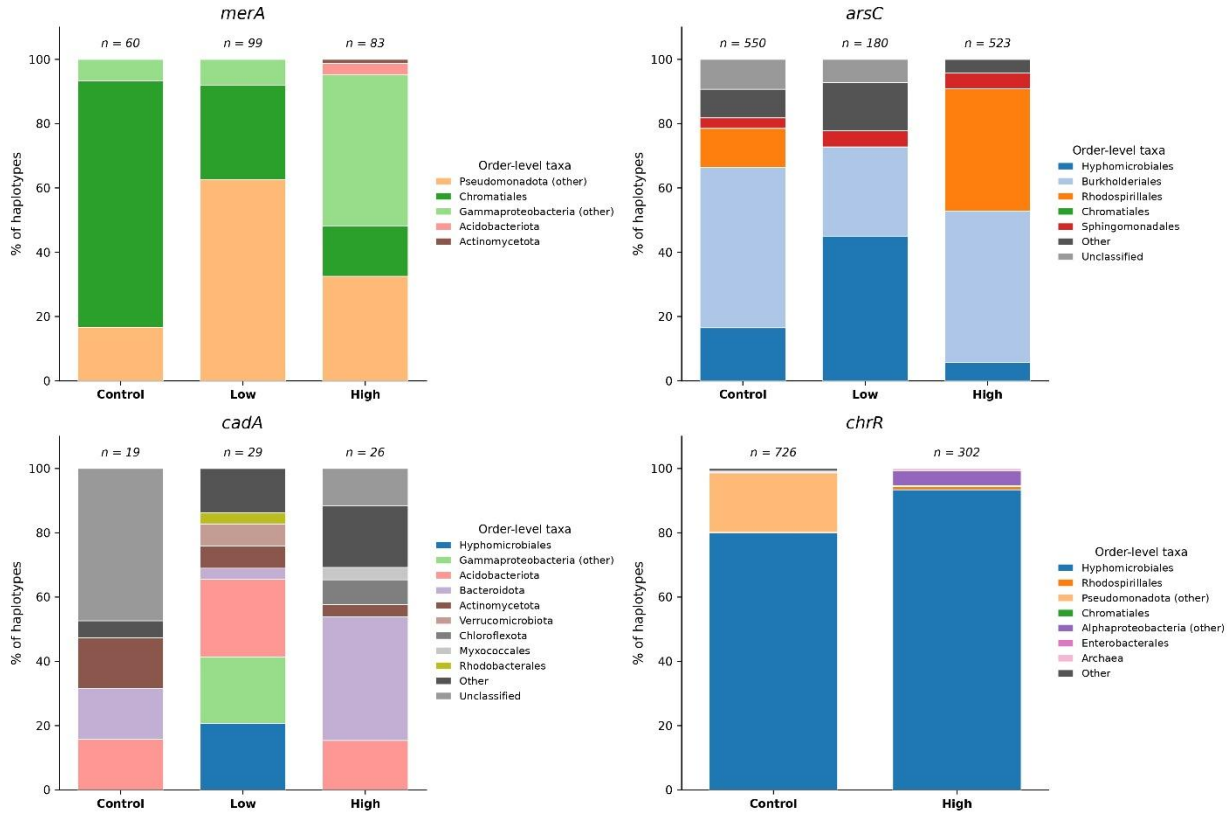


Figure S8. Taxonomic composition of MRG-containing haplotypes per gene and regime, aggregated to the order level for cross-gene comparability. Stacked bars indicate the percentage of haplotypes affiliated with each order, based on best-hit BLASTP against NCBI nr. The total number of haplotypes per regime is indicated above each bar. The *chrR* Low regime is not shown because the 76 Low-regime haplotypes could not be taxonomically assigned within the wall-clock allowance (92% overall coverage; see Methods). Numerical values are reported in Table S13.