Identifying genomic loci under selection in a widespread *Bradyrhizobium* across Australian ecosystems using landscape genomics

Jose Antonio Blanco Alcantara[1] and Anna K. Simonsen[1,2]

[1]Department of Biological Sciences, Florida International University, Miami, Florida, USA

[2]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

Corresponding author: Anna K. Simonsen (simonsen@ualberta.ca)

ABSTRACT

Climate change is reshaping soil environments, intensifying selective pressures on microbial communities that drive essential ecosystem processes. Understanding how nitrogen-fixing rhizobia adapt to environmental variation is critical for predicting ecosystem responses to global change. Here, we used redundancy analysis (RDA) to identify genomic loci associated with environmental gradients across 375 *Bradyrhizobium diazoefficiens* isolates from southwestern Australia, analyzing variants mapped to four independent reference genomes.

Climate variability emerged as the dominant driver of genomic adaptation, with annual rainfall and temperature ranking first and second among twelve environmental variables tested. These climate effects substantially exceeded those of soil chemistry factors, with rainfall explaining nearly twice the variance of the strongest soil variable (carbon content). Notably, annual temperature showed its strongest association with RDA axis 3 rather than early axes, demonstrating that limiting analysis to the first two RDA dimensions can cause researchers to miss important environmental effects.

Functional enrichment analysis revealed that signal transduction mechanisms were significantly over-represented among outlier loci, while core symbiosis genes showed consistent depletion, providing suggestive evidence for purifying selection maintaining nitrogen-fixing capacity. Within symbiosis genes, regulatory components (exoR, regB) showed 100% outlier rates while structural machinery (T4SS, nol genes) showed complete conservation, supporting a "regulatory evolution" model where adaptation occurs through expression control rather than structural changes.

These findings demonstrate that *Bradyrhizobium* populations adapt to climate heterogeneity primarily through signal transduction and regulatory networks, while core metabolic and symbiotic functions remain under strong functional constraint.

Introduction

Soil microbial communities drive essential ecosystem processes including nutrient cycling, organic matter decomposition, and plant-microbe interactions that enable plants to respond to environmental variation (Fierer & Jackson 2006; Griffiths & Philippot 2013). Understanding how these microbial communities adapt to evolving environmental pressures – including shifts in

community structure and species turnover – is essential for predicting ecosystem responses to climate change (Allison & Martiny 2008). Among soil microbes, rhizobia occupy a unique ecological position as nitrogen-fixing symbionts of legumes, playing a critical role in global nutrient cycling by transferring atmospheric nitrogen into ecosystems and fueling the biogeochemical nitrogen cycle (Galloway *et al.* 2008). However, climate change is reshaping the environmental conditions these communities experience, with shifts in temperature, precipitation, pH, and nutrient availability creating unprecedented challenges for microbial populations (Rousk *et al.* 2010).

A fundamental question in evolutionary microbiology is how bacterial populations adapt to environmental variation at the genomic level. Local adaptation occurs when natural selection favors different alleles in different environments, resulting in genotype-environment associations detectable through landscape genomics approaches (Savolainen *et al.* 2013). In soil bacteria, environmental factors impose selective pressures that can drive adaptation through several mechanisms: mutations in regulatory genes that alter gene expression thresholds, structural variations in stress-response pathways, and modifications to resource acquisition systems (Remigi *et al.* 2016). The relative importance of different environmental factors as drivers of bacterial adaptation remains poorly understood, particularly the balance between large-scale climate variability and localized soil chemistry.

Climate variables such as temperature and precipitation operate at regional to continental scales, potentially imposing consistent selective pressures across broad geographic areas. Temperature directly affects microbial metabolic rates, membrane fluidity, and protein stability, while precipitation influences soil moisture regimes and nutrient availability (Bardgett *et al.* 2008). Conversely, soil chemistry factors including pH, nutrient concentrations, and metal availability vary at much finer spatial scales, creating highly heterogeneous selective landscapes even within localized areas (Fierer & Jackson 2006). Soil pH, for example, influences nutrient solubility, metal toxicity, and cellular homeostasis, imposing strong selection on microbial physiology (Rousk *et al.* 2010). Soil nitrogen availability (ammonia and nitrate) may be particularly important for nitrogen-fixing bacteria, as high environmental nitrogen reduces the selective advantage of costly nitrogen fixation processes (Denison & Kiers 2011).

Rhizobia such as *Bradyrhizobium* provide critical ecosystem services through nitrogen fixation and plant productivity support (Galloway *et al.* 2008). These bacteria possess specialized functional genes enabling nitrogen fixation and symbiosis with legumes, such as *nif* (nitrogen fixation), *fix* (oxygen response), and *nod* (nodulation) gene families. Nitrogen fixation requires substantial energy investment, and maintaining symbiotic function may constrain overall rhizobial fitness in the free-living state (Poole *et al.* 2018). Conversely, adaptations increasing resistance to environmental stresses could alter symbiotic efficiency regarding nodulation competitiveness and nutrient uptake (Masson-Boivin *et al.* 2009). Understanding how environmental selection acts on genes involved in the free living stage and symbiosis stage – particularly whether symbiotic genes experience environmental selection or remain buffered within the plant-controlled nodule environment – is critical for predicting rhizobial responses to climate change.

The central goal of this study is to determine to what extent climate variability and soil chemistry impose selective pressures on rhizobial populations, with climate factors potentially driving broad-scale adaptation patterns while soil chemistry shapes localized responses. We hypothesize

that local adaptation to environmental stressors in rhizobia will involve changes in genomic regions related to regulatory networks and environmental sensing systems (Remigi *et al.* 2016). However, these genomic regions are also involved in the regulation of symbiosis and nitrogen fixation (Zhang *et al.* 2024), creating a genetic linkage between evolutionary responses to climate and soil chemistry variability and symbiosis function. Conversely, if core symbiosis genes themselves are under environmental selection, adaptation may be constrained by functional trade-offs.

Here, we test for the relative contributions of climate and soil chemistry in shaping adaptive genetic variation in the *Bradyrhizobium diazoefficiens* species complex, using a landscape genomics approach, where we sampled rhizobia strains spanning a large climate and soil chemistry gradient within a major climatic region (Dinnage *et al.* 2018; Simonsen *et al.* 2019), Southwest Australia. Landscape genomics provides a framework to discover adaptive genetic diversity without needing to hypothesize candidate loci *a priori* (Rellstab *et al.* 2015). This approach specifically aims to identify alleles showing increased frequency along environmental gradients, indicating putative local adaptation. Among landscape genomics methods, multivariate redundancy analysis (RDA) has emerged as an effective approach for detecting multilocus markers of local adaptation while controlling false positive rates (Capblancq & Forester 2021; Forester *et al.* 2018). RDA explains linear relationships between environmental predictors and allele frequencies, isolating genotype-environment associations within environmentally variable landscapes.

RDA offers several advantages over univariate genotype-environment association methods: (1) it accounts for correlations among environmental variables through multivariate ordination, reducing false positives from collinear predictors; (2) it detects loci responding to combinations of environmental factors rather than single variables; (3) it provides a statistically rigorous framework with permutation-based significance testing; and (4) it visualizes both sample-environment relationships and SNP-environment associations simultaneously (Legendre *et al.* 2012). These properties make RDA particularly well-suited for natural environments, where environmental gradients are often multivariate and interactive.

The goals of this study are to 1) Determine which environmental factors drive putative adaptive genetic variation among *Bradyrhizobium* populations 2) Test whether genetic variants show consistent patterns of environmental association across different *Bradyrhizobium* reference genomes 3) To characterize how environmental factors influence genomic variation in symbiosis-specific genes 4) Identify any distinct environmental response patterns through clustering of outlier loci.

Methods

Study Design and Environmental Sampling

To examine adaptive genomic diversity among *Bradyrhizobium* species, previous work undertook a systematic sampling procedure to maximize soil and climate variability of sampling sites (Dinnage *et al.* 2018; Simonsen *et al.* 2019). In brief, strains that symbiotically associate with an endemic species, *Acacia acuminata*, which only grows in Southwest Australia region, were targeted. Distribution data for *A. acuminata* were obtained from the Atlas of Living

Australia, while mean annual precipitation and temperature data were obtained from BIOCLIM datasets (Booth *et al.* 2014). A systematic algorithmic selection procedure identified twenty sampling sites that maximized climatic diversity while minimizing the often confounding co-variation between temperature and precipitation gradients, and controlling for spatial autocorrelation in precipitation patterns. Three soil samples were obtained from each site and brought back to the lab for rhizobial isolation through inoculation trapping techniques on *A. acuminata* seedlings. This selection method maximizes sampling variation attributed to climate and soil chemistry, while minimizing genetic variation attributed to specialization to specific *Acacia* host species. Consequently, results regarding adaptation to abiotic stresses such as drought, salinity, and acidity are limited to rhizobia that retain symbiotic competence.

Isolates of a single ubiquitous rhizobia species, *Bradyrhizobium diazoefficiens,* that occurred abundantly in all soil samples were chosen for Illumina sequencing, generating a sequenced cultured collection of 380 isolates (19 sequenced isolates per site). The final dataset comprised 375 *Bradyrhizobium diazoefficiens* isolates after quality filtering (5 contaminated samples were removed from an initial set of 380), (Simonsen *et al.* 2019).

Environmental metadata were collected for each sampling location, including climate variables from BIOCLIM datasets and soil factors measured directly at sampling sites. Climate variables spanned a range of annual mean temperatures from 14.8°C to 18.2°C (maximum: 18.2°C, minimum: 14.8°C) and annual precipitation from 315 mm to 625 mm (maximum: 625 mm, minimum: 315 mm). Soil chemistry measurements included gravel content (maximum: 65%, minimum: 5%), ammonia concentration (maximum: 28.3 ppm, minimum: 2.1 ppm), nitrate concentration (maximum: 15.7 ppm, minimum: 0.8 ppm), potassium (maximum: 185 ppm, minimum: 42 ppm), carbon content (maximum: 2.8%, minimum: 0.4%), electrical conductivity (maximum: 0.35 dS/m, minimum: 0.08 dS/m), pH (maximum: 7.2, minimum: 4.9), copper concentration (maximum: 3.2 ppm, minimum: 0.3 ppm), zinc concentration (maximum: 2.8 ppm, minimum: 0.2 ppm), and magnesium concentration (maximum: 245 ppm, minimum: 68 ppm).

Reference Genome Preparation and Read Alignment

Paired-end reads from all 375 *Bradyrhizobium diazoefficiens* isolates were mapped against the multiple reference genomes to reduce bias inherent to single-reference analysis. To enhance genomic resolution and variability of single nucleotide variation, four fully annotated reference genomes from strains isolated from *Acacia acuminata* in southwestern Australia were used. These genomes included Genome A (GCA_016616425.1, strain 41.2), Genome B (GCA_016616235.1, strain 38.8), Genome C (GCA_016616885.1, strain 36.1), and Genome D (GCA_016599855.1, strain 65.7) (Chia & Simonsen 2021).

Each reference assembly was indexed using BWA-MEM version 0.7.17 (Li 2013), an accurate alignment algorithm optimized for high-throughput short-read sequencing data. Paired-end reads were aligned to each reference assembly independently using the `bwa mem` command.

SAM format alignment outputs were converted to compressed BAM format using SAMtools version 1.10 (Danecek *et al.* 2021), sorted by genomic coordinates, and indexed. To ensure high-confidence variant calling, reads with mapping quality scores below 10 were filtered out. This filtration step removes low-confidence alignments likely derived from repetitive or paralogous regions common in bacterial genomes.

Variant Calling

Variants were called from each aligned isolate using FreeBayes version 1.3.2 (Garrison & Marth 2012), a haplotype-based variant caller designed for populations with limited recombination. FreeBayes was run with the following parameters: --ploidy 1 to reflect the haploid nature of the *Bradyrhizobium* genome, --min-alternate-count 2 to require at least two reads supporting the alternate allele, --min-alternate-fraction 0.005 to retain low-frequency variants, and --haplotype-length 0 to disable local haplotype assembly and call each site independently.

Resulting per-isolate Variant Call Format (VCF) files were compressed using bgzip and indexed using tabix (Li 2011) for efficient data retrieval. Sample names were standardized across VCF files using bcftools reheader. Individual VCF files for each reference genome were merged into multi-sample VCFs using bcftools merge with the --force-samples option, producing a single combined variant dataset per reference genome covering all 375 isolates.

SNPs were selectively retained from merged VCF files, excluding indels and multi-nucleotide polymorphisms to focus analysis on biallelic single-nucleotide variants. Variant filtering was performed using VCFtools version 0.1.17 (Danecek *et al.* 2011). Filters applied included a minimum minor allele frequency of 0.001 (to exclude extremely rare variants) and a maximum missing data threshold of 5% (--max-missing 0.95), ensuring that loci were not missing in more than 5% of samples.

Biallelic Filtering

After minor allele frequency and missing data filtering, SNPs were further refined using a fuzzy biallelic filtering approach designed to eliminate residual multiallelic noise. Because most population genomic tools are optimized for biallelic SNPs, and triallelic variants are susceptible to miscalling in bacterial short-read data, a biallelic filtering strategy was implemented to retain informative variation while minimizing noise.

For each SNP, alleles were recoded based on frequency: the most frequent allele was coded as 0 (major), the second-most frequent as 1 (minor), and all remaining alleles provisionally coded as missing (NA). A SNP was retained only if the count of the third-most frequent allele was less than 80% of the second-most frequent allele's count. Sites exceeding this threshold were excluded entirely. For retained loci, all third alleles and ambiguous calls were recoded as NA. This approach preserved a large number of usable SNPs with consistent biallelic structure while tolerating minor background noise inherent to microbial sequencing.

Missing Genotype Imputation

After fuzzy biallelic filtering, remaining missing genotypes were imputed using a K-nearest neighbors (KNN) algorithm with k = 5, implemented in the VIM package for R (Kowarik & Templ 2016). The genotype matrix was transposed such that isolates occupied rows and SNPs

occupied columns. A pairwise Euclidean distance matrix between isolates was constructed based on SNP profiles.

For each missing genotype, the five isolates with highest genotypic similarity that possessed non-missing values at the focal locus were identified. Missing values (both rare alleles and true missing data) were imputed as the most frequent allele among these five nearest neighbors. This approach leverages the highly clonal population structure characteristic of *Bradyrhizobium*, where genetically similar isolates are expected to share high genomic identity. KNN imputation preserves overall genetic structure while eliminating biases from non-random missingness. After imputation, the matrix contained no missing values, making it suitable for constrained ordination methods and downstream variance analyses.

Environmental Data Preprocessing

Environmental variables underwent preprocessing to meet RDA assumptions and improve interpretability. High correlation among environmental variables can cause multicollinearity issues in RDA. Hierarchical clustering based on correlation distance was performed to identify redundant variables (see Supporting Information, Section S1). Variables with pairwise correlations $|r| > 0.7$ were removed to create a reduced set free from severe multicollinearity, using an hierarchical clustering plot as a guide (Figure S1). The remaining variables were mostly only weakly correlated (Figure S2).

Final environmental variable set (13 variables):

Climate variables: - Annual mean temperature (Env_ann_temp): 14.8°C to 18.2°C - Annual precipitation (Env_ann_rain): 315 mm to 625 mm

Soil variables: - Soil gravel content: 5% to 65% - Soil ammonia concentration: 2.1 ppm to 28.3 ppm - Soil nitrate concentration: 0.8 ppm to 15.7 ppm - Soil potassium: 42 ppm to 185 ppm - Soil carbon content: 0.4% to 2.8% - Electrical conductivity/salinity proxy (Soil_conductivity): 0.08 dS/m to 0.35 dS/m - Soil pH: 4.9 to 7.2 - Soil copper concentration: 0.3 ppm to 3.2 ppm - Soil zinc concentration: 0.2 ppm to 2.8 ppm - Soil magnesium concentration: 68 ppm to 245 ppm

To meet RDA assumptions of normality and homoscedasticity, transformations were applied using the tidymodels framework: (1) log transformation of soil conductivity as log(conductivity + 0.01) to avoid log(0); (2) Yeo-Johnson transformation to improve normality and homogeneity of variance, which appropriately handles negative values and zeros (see Supporting Information, Section S1 for mathematical details); and (3) normalization (centering and scaling) to achieve mean = 0 and standard deviation = 1 for all variables.

Redundancy Analysis (RDA)

RDA was performed using the vegan package in R. RDA performs multivariate linear regression where the genotype matrix (Y) is explained by environmental predictors (X), yielding fitted values representing genetic variation explained by environmental variables. Principal Component Analysis (PCA) is then performed on the fitted matrix to obtain canonical axes, which are linear combinations of environmental variables that explain maximal constrained variance.

The 13 environmental variables were selected based on their potential to influence *Bradyrhizobium* physiology, genome structure, and community-level adaptation. Environmental metadata were matched to genomic samples using sample identifiers, and samples present in genotype data but lacking environmental data were excluded from RDA.

To determine which RDA axes represent significant genotype-environment associations, permutation-based ANOVA was performed using an adaptive permutation scheme with cutoff = 0.1. Axes with $p \leq 0.01$ were retained as "significant," reducing false positives while retaining axes with strong environmental signals. Additional by-margin permutation tests assessed the independent contribution of each environmental variable.

Outlier Detection

For each SNP, loadings on significant RDA axes ($p \leq 0.01$) were extracted. These loadings represent the contribution of each SNP to genotype-environment associations captured by each axis. To identify outlier SNPs in multivariate loading space, Mahalanobis distance was used, which accounts for correlation structure among RDA axes and is more statistically rigorous than examining individual axes independently (see Supporting Information, Section S2 for mathematical details). Under the null hypothesis of neutral evolution, the squared Mahalanobis distance follows a chi-squared distribution. P-values were calculated and a threshold of $p \leq 0.05$ was applied.

Clustering of Outlier SNPs

Outlier SNPs sharing similar environmental associations likely respond to the same environmental pressures, may be functionally related, or could represent linked variants in the same genomic region. Clustering allows identification of groups of SNPs with shared environmental drivers, dimensionality reduction for interpretation, and testing for functional enrichment within clusters.

A key methodological challenge is that SNP allele coding (0/1 for reference/alternate) is arbitrary. Two SNPs with opposite RDA loadings may represent the same biological pattern if alleles are coded differently. A hemisphere folding transformation was developed to address this issue (see Supporting Information, Section S3). Before clustering, loading vectors were unit-normalized and then "folded" such that vectors and their negatives map to the same point. This transformation is mathematically equivalent to clustering in projective space and ensures that biologically similar SNPs cluster together regardless of arbitrary allele coding.

Von Mises-Fisher (vMF) mixture models were applied for clustering unit vectors. Unlike Euclidean clustering methods such as k-means, vMF is the appropriate distribution for directional data naturally residing on a unit sphere. The concentration parameter κ has direct interpretation, with higher values indicating tighter clusters (see Supporting Information, Section S3 for mathematical details).

Bayesian Information Criterion (BIC) was evaluated across a range of cluster numbers ($K = 2$ to $K = 40$). Optimal K was selected using the elbow method at 95% of total BIC improvement. LOESS smoothing (span = 0.3) was applied to the BIC curve to reduce noise from convergence failures. A consensus approach took the median of optimal K across all four runs, yielding $K = 10$ clusters (see Supporting Information, Section S3 for mathematical details).

After independent clustering in each run, cluster labels are arbitrary. An optimization algorithm was used to match clusters across reference genomes based on similarity in environmental projections and concentration parameters (see Supporting Information, Section S4 for details). This ensures consistent cluster coloring across plots and enables cross-run comparisons.

COG Functional Enrichment Analysis

To statistically test whether specific COG functional categories were significantly enriched among outlier loci, gene-level enrichment analysis was performed using the hypergeometric test. This approach tests whether the proportion of outlier genes in a given functional category exceeds expectation by chance. The statistical background for enrichment testing consisted of all genes containing at least one SNP in the filtered dataset. This represents the set of genes that could have been detected as outliers given the sampling design and sequencing coverage, providing an appropriate null expectation for enrichment. COG annotations can assign multiple functional categories to a single gene (e.g., "EH" indicating both amino acid metabolism and coenzyme transport). Multi-category annotations were split into individual letters and tested independently, allowing genes to contribute to multiple functional categories. For each of 26 COG functional categories, a 2×2 contingency table was constructed comparing the number of outlier versus background genes within that category versus all other categories. Analysis was performed independently for each of the four reference genomes.

Symbiosis Gene Enrichment Analysis

To test whether genes involved in nitrogen-fixing symbiosis were specifically associated with environmental adaptation, an enrichment analysis focused on symbiosis-related gene families was performed. A reference list of 164 symbiosis genes was compiled from the scientific literature, including genes involved in nodulation signaling (nod/noe/nol family, 49 genes), nitrogen fixation (nif family, 20 genes), oxygen response during symbiosis (fix family, 19 genes), bacterial secretion systems (T3SS/T4SS, 48 genes), and effector proteins (nop family, 28 genes). All SNPs in the comprehensive dataset were annotated with symbiosis gene status, according to "Symbiosis_gene" (binary: 1 = symbiosis gene, 0 = non-symbiosis gene), and "Symbiosis_gene_family" (gene family name or "non-symbiotic" for non-symbiosis genes).

Enrichment testing followed the same gene-level approach as COG enrichment: genes were classified as outlier genes if they contained at least one outlier SNP ($p \leq 0.05$), and the background consisted of all genes with at least one SNP in the filtered dataset. For each reference genome, a hypergeometric test was performed comparing the proportion of outlier genes among symbiosis genes versus non-symbiosis genes.

Visualization

RDA biplots visualize sample scores (positions of isolates in environmental space), environmental vectors (directions of environmental gradients), and cluster coloring (samples colored by predominant cluster membership). Manhattan plots were generated in both linear and circular formats showing genomic position versus Mahalanobis distance, with points colored by cluster assignment. Circular parallel coordinates plots showing environmental associations for each cluster were created to visualize which environmental variables drive each cluster using ggplot2 (Wickham 2016). See Supplementary Section S5 for details.

Results

Sample and SNP Summary Statistics

A total of 375 *Bradyrhizobium* isolates were analyzed after quality filtering Variant calling and filtering across four reference genomes yielded 1,859 outlier SNPs using a Mahalanobis distance threshold of $p \leq 0.05$ (Table 2).

Table 2. Number of outlier SNPs detected per reference genome using Mahalanobis distance with $p \leq 0.05$ threshold.

Genome Run	Outlier SNPs
Genome A	579
Genome B	552
Genome C	255
Genome D	473

Redundancy Analysis Results

RDA models showed that environmental variables explained approximately 9-10% of total genomic variation across all runs (Table 3). All models were statistically significant (F-statistics > 6.3, $p = 0.001$ by 999 permutations).

Table 3. Variance partitioning in RDA models. Environmental variables explain approximately 9-10% of total genomic variation across all runs.

Run	Total Inertia	Constrained	% Explained
Genome A	191.87	19.84	10.34
Genome B	201.30	19.43	9.65
Genome C	86.98	8.04	9.24
Genome D	144.22	13.43	9.31

Significant RDA Axes

Permutation-based ANOVA (999 permutations, adaptive cutoff = 0.1) identified exactly four significant axes ($p \leq 0.01$) in each run (Table 4). The first axis (RDA1) consistently explained the largest proportion of constrained variance (4.54-12.03 units), with F-statistics ranging from 20.70 to 27.42. Subsequent axes (RDA2-RDA4) showed progressively decreasing variance and F-statistics but remained highly significant. These four significant axes were used for outlier detection and clustering analyses.

Table 4. Significance testing of RDA axes by permutation ANOVA (999 permutations). Permutation tests were halted after p-values exceeded 0.1 for computational efficiency (indicated by "-"). The four significant axes ($p \leq 0.01$) per run were used for outlier detection and subsequent clustering analyses.

Run	Axis	Variance	F	p-value
Genome A	RDA1	12.034	27.42	<0.001
Genome A	RDA2	5.788	13.22	<0.001
Genome A	RDA3	4.243	9.72	<0.001
Genome A	RDA4	2.765	6.35	<0.001
Genome A	RDA5	1.789	4.12	0.02
Genome A	RDA6	1.280	2.96	0.14
Genome A	RDA7	1.045	2.38	-
Genome B	RDA1	11.549	25.05	<0.001
Genome B	RDA2	5.690	12.38	<0.001
Genome B	RDA3	4.393	9.58	<0.001
Genome B	RDA4	3.317	7.25	<0.001
Genome B	RDA5	1.880	4.12	0.03
Genome B	RDA6	1.504	3.31	0.10
Genome B	RDA7	1.220	2.65	-
Genome C	RDA1	4.540	22.33	<0.001
Genome C	RDA2	2.357	11.63	<0.001
Genome C	RDA3	1.671	8.27	<0.001
Genome C	RDA4	1.107	5.49	0.006
Genome C	RDA5	0.681	3.39	0.12
Genome C	RDA6	0.624	3.07	-
Genome C	RDA7	0.540	2.66	-
Genome D	RDA1	6.943	20.70	<0.001
Genome D	RDA2	4.168	12.46	<0.001
Genome D	RDA3	3.334	10.00	<0.001
Genome D	RDA4	2.324	6.99	<0.001
Genome D	RDA5	1.248	3.76	0.03
Genome D	RDA6	1.043	3.15	0.07
Genome D	RDA7	0.802	2.43	0.21

Environmental Variable Contributions

By-margin permutation tests (999 permutations) assessed the independent contribution of each environmental variable while accounting for all other variables (Table 5). Climate variables (annual rainfall and temperature) consistently explained the largest variance across runs (mean variance: 3.99 and 2.83 respectively), followed by soil carbon (2.04) and gravel content (1.95). All soil nitrogen variables (ammonia, nitrate) and metal variables (zinc, copper, magnesium)

showed significant associations ($p \leq 0.01$) in all four runs. Soil pH showed the weakest associations (mean variance: 0.65, rank 12/12), with no runs reaching $p \leq 0.01$ significance.

Table 5. Environmental variable contributions across all four runs, ranked by mean variance explained. By-margin permutation tests (999 permutations) assessed each variable's independent contribution while accounting for all other variables.

Variable	Mean Variance	Mean F	Min p-value	Runs Significant ($p \leq 0.01$)
Climate:Rain	3.988	10.81	<0.001	4
Climate:Temp	2.826	7.69	<0.001	4
Soil:C	2.037	5.56	<0.001	4
Soil:Gravel	1.948	5.24	<0.001	4
Soil:NO_3	1.295	3.65	<0.001	4
Soil:Zn	1.275	3.54	<0.001	4
Soil:Cu	1.206	3.24	<0.001	4
Soil:Mg	1.137	3.09	<0.001	4
Soil:NH_3	1.062	2.93	<0.001	4
Soil:Salt	1.023	2.85	0.002	4
Soil:K	0.816	2.30	0.003	3
Soil:pH	0.649	1.79	0.026	0

RDA Biplots

RDA biplots show the relationship between isolate genotypes, environmental gradients, and cluster assignments (Figure 1). Samples are positioned based on genotype projections onto environmental space, with environmental vectors indicating direction and strength of correlations. Clusters show varying degrees of environmental specialization, with some clusters tightly grouped and others broadly dispersed across environmental space.

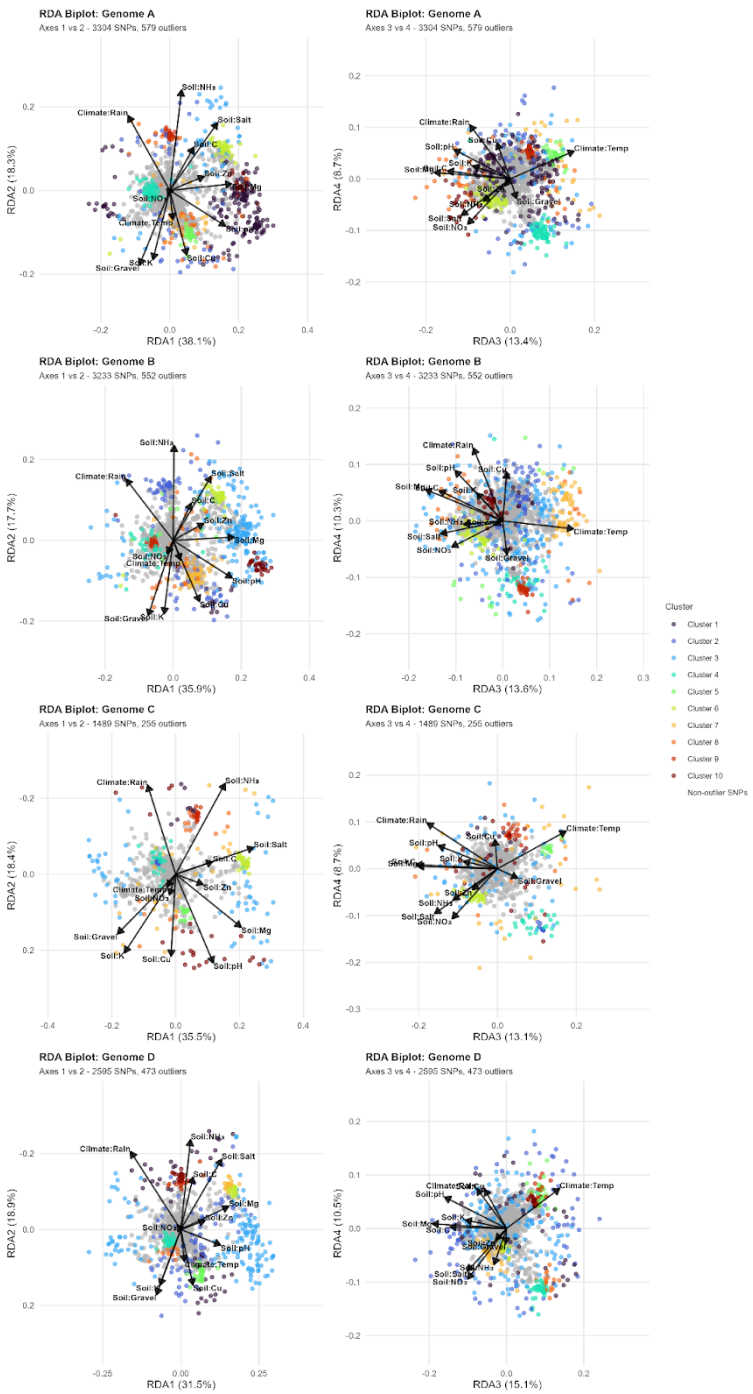

Figure 1. *RDA biplots for all four runs showing RDA axes 1-2 (left panels) and axes 3-4 (right panels). Points represent individual isolates colored by their predominant cluster membership (10 clusters total). Arrows show environmental variable loadings. Clear separation of samples along environmental gradients is visible, with climate variables (temperature, rainfall) often driving RDA1 and soil chemistry variables (pH, nutrients, salinity) driving RDA2. Some clusters show strong environmental associations with tight grouping in environmental space.*

Genomic Distribution of Outlier SNPs

Circular Manhattan plots (Figure 3) display the genome-wide distribution of outlier SNPs colored by cluster assignment. Outliers are distributed across the genome rather than concentrated in single regions. The radial pattern shows Mahalanobis distance, with outliers extending beyond the significance threshold circle.

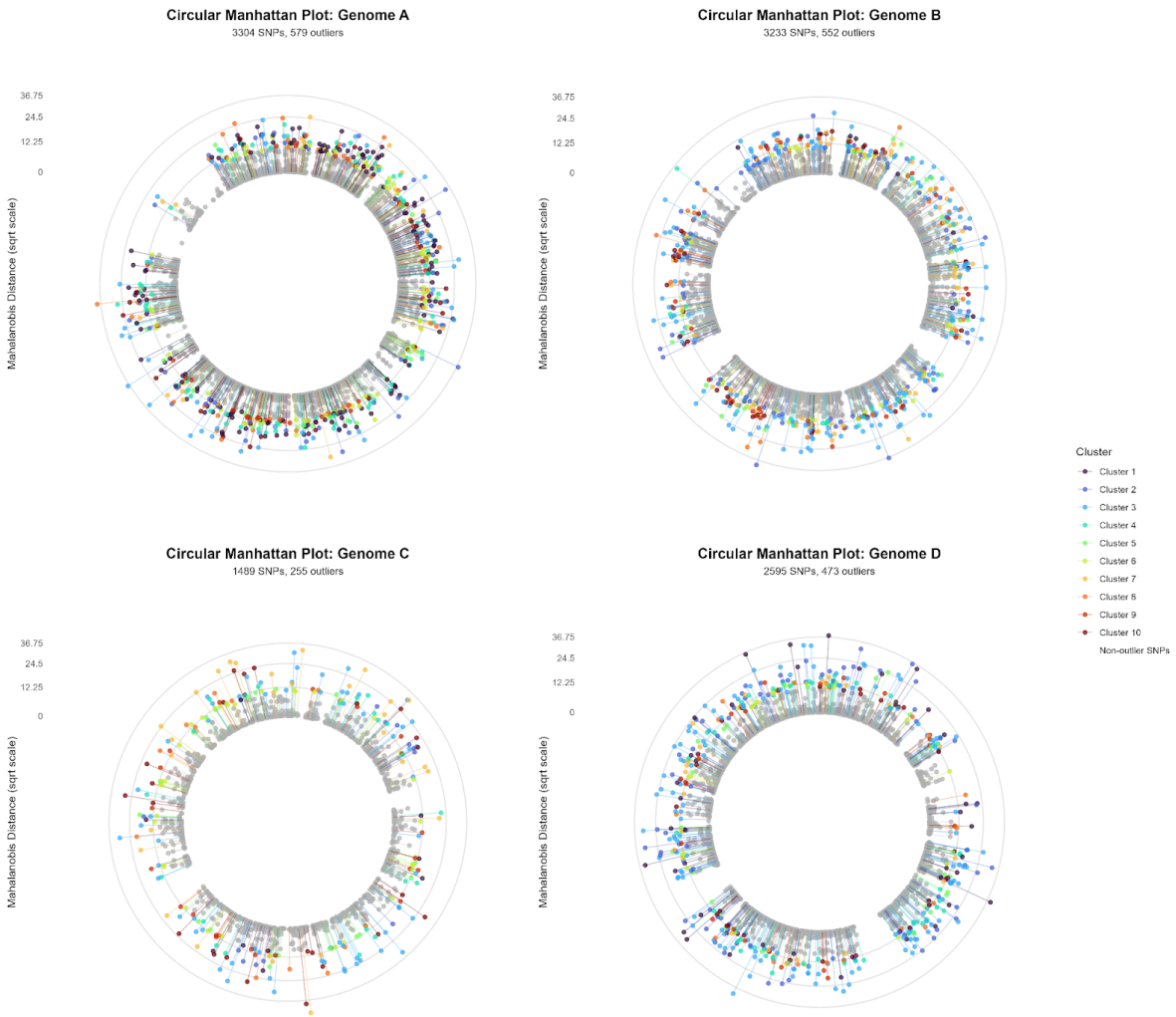

Figure 3. *Circular Manhattan plots for all four runs. Radial distance represents Mahalanobis distance, with outliers extending beyond the significance threshold circle (innermost dashed circle). Points are colored by cluster assignment (10 clusters). Non-outlier SNPs shown in grey. Outliers are distributed across the genome, indicating polygenic architecture of environmental adaptation.*

Clustering of Outlier SNPs

BIC-based optimization using the 95% improvement threshold yielded a consensus of K = 10 clusters across all runs (Figure 4). This balances model complexity with biological interpretability.

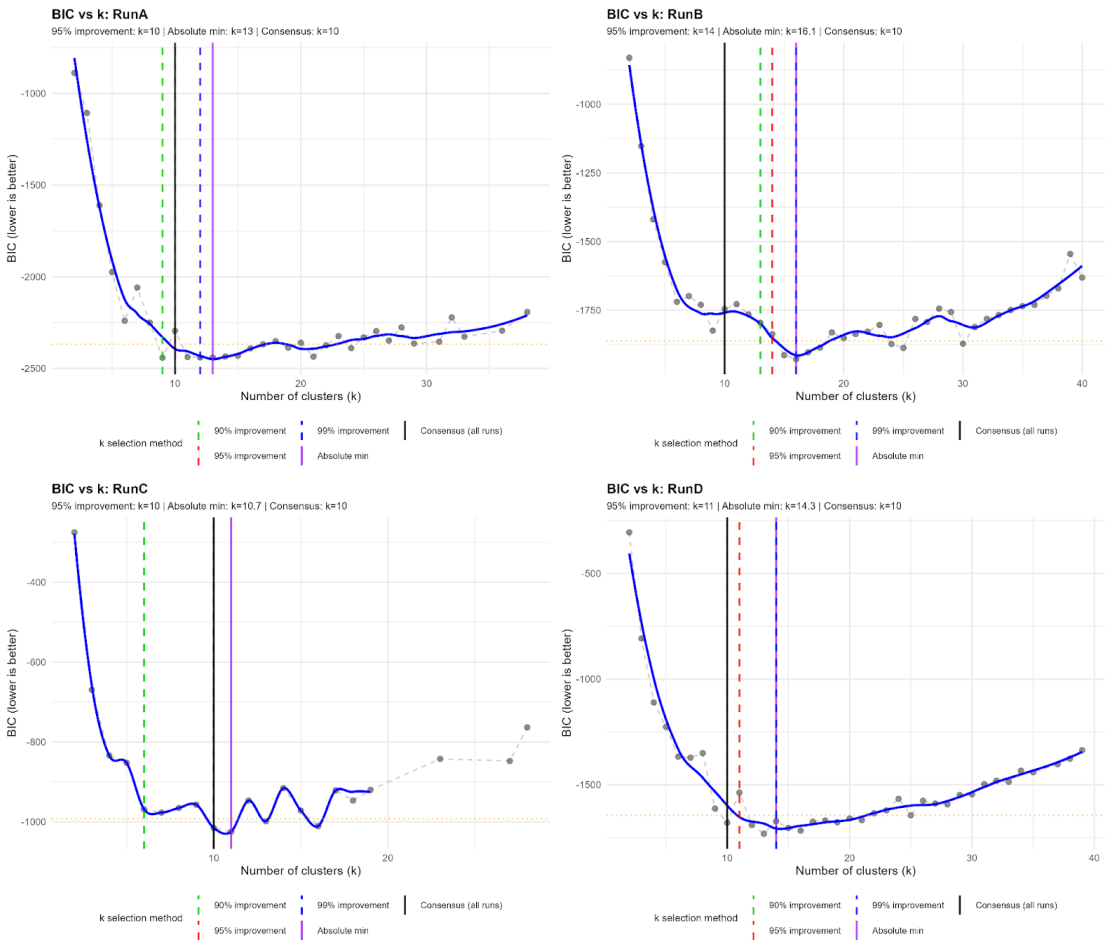

Figure 4. *BIC curves for cluster number selection. Grey points show raw BIC values across different numbers of clusters (K), blue line shows LOESS-smoothed curve. Vertical lines indicate different selection criteria: 90% improvement (dashed), 95% improvement (solid), 99% improvement (dotted), and absolute minimum (dash-dot). The 95% improvement threshold was selected as optimal, yielding a consensus K = 10 clusters across runs.*

Cluster Environmental Profiles

Circular parallel coordinates plots (Figure 5) display environmental associations for each of the 10 clusters. Each cluster shows a distinct environmental profile, with individual SNP projections (colored lines) radiating from the cluster center (thick black line). The spread of SNP lines within each cluster visualizes heterogeneity of environmental associations. Clusters show diverse environmental response patterns. Some clusters associate strongly with climate variables (temperature, rainfall), others with soil chemistry (pH, nutrients), and others with soil physical

properties (gravel content). Some clusters display mixed associations across multiple environmental axes.

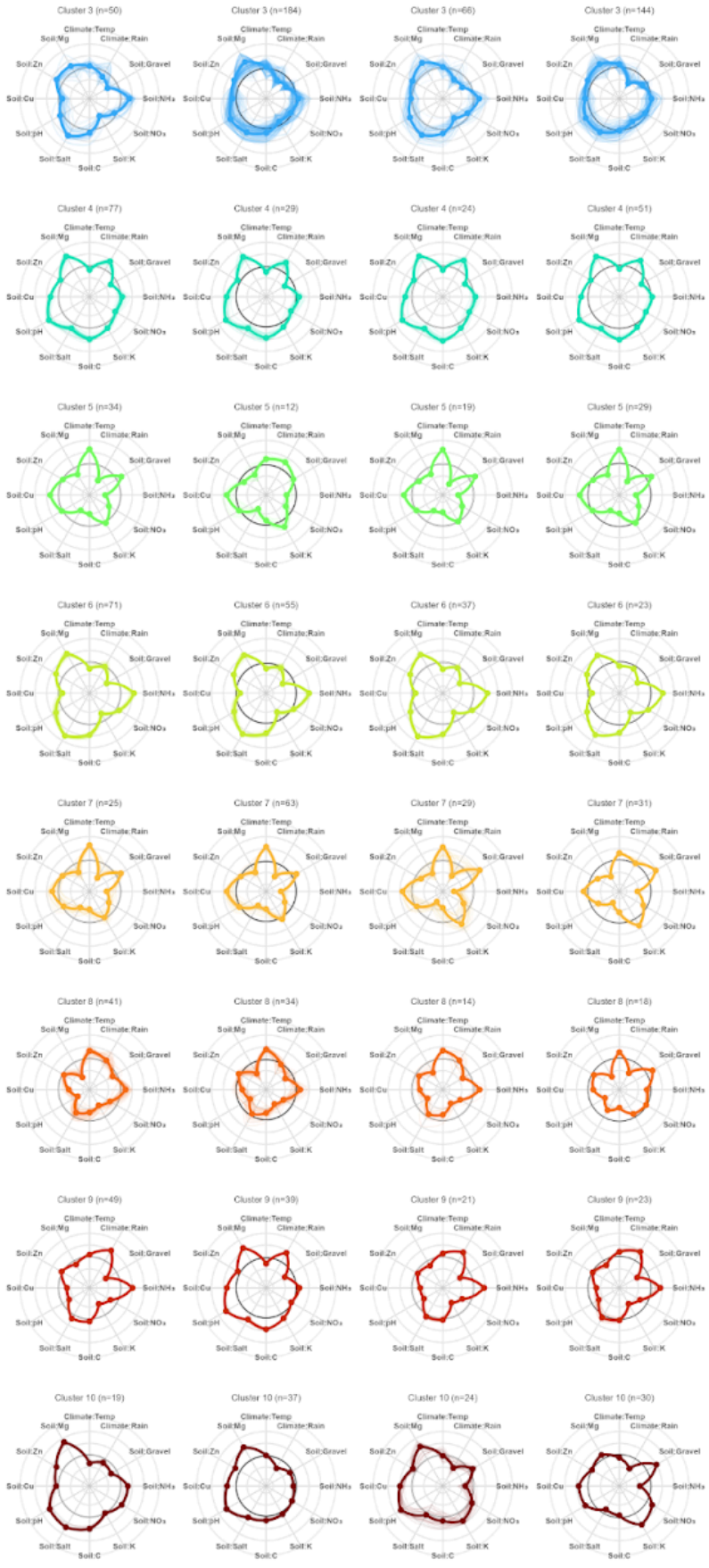
Cluster 3 (n=50)
Cluster 3 (n=184)
Cluster 3 (n=66)
Cluster 3 (n=144)
Cluster 4 (n=77)
Cluster 4 (n=29)
Cluster 4 (n=24)
Cluster 4 (n=51)
Cluster 5 (n=34)
Cluster 5 (n=12)
Cluster 5 (n=19)
Cluster 5 (n=29)
Cluster 6 (n=71)
Cluster 6 (n=55)
Cluster 6 (n=37)
Cluster 6 (n=23)
Cluster 7 (n=25)
Cluster 7 (n=63)
Cluster 7 (n=29)
Cluster 7 (n=31)
Cluster 8 (n=41)
Cluster 8 (n=34)
Cluster 8 (n=14)
Cluster 8 (n=18)
Cluster 9 (n=49)
Cluster 9 (n=39)
Cluster 9 (n=21)
Cluster 9 (n=23)
Cluster 10 (n=19)
Cluster 10 (n=37)
Cluster 10 (n=24)
Cluster 10 (n=30)
Climate:Temp
Climate:Rain
Soil:Gravel
Soil:NH₄
Soil:NO₃
Soil:K
Soil:C
Soil:Salt
Soil:pH
Soil:Cu
Soil:Zn
Soil:Mg

Figure 5. *Environmental profiles for each of the 10 clusters across all runs. Each subplot shows individual SNP projections from one cluster as colored lines radiating from center, with the cluster center shown as a thick black line. Radial distance indicates strength of association with each environmental variable (labeled around the circle). Positive values (outward) indicate positive correlation; negative values (inward) indicate negative correlation. The spread of individual SNP lines visualizes heterogeneity within each cluster.*

Functional Annotation of Outlier Loci

The majority of outlier SNPs (94.9%, 1,767/1,862) map to coding sequences, with 73.0% (1,359) having assigned gene names. Outlier SNPs across all four runs map to 1,055 unique genes spanning diverse functional categories (Table 7; see Table S1 for full category list). The most represented categories include unknown function (328 genes), amino acid metabolism (230 genes), signal transduction (182 genes), ion transport (142 genes), and carbohydrate metabolism (136 genes).

Table 7. Top 10 COG functional categories among outlier genes. Counts represent unique genes containing at least one outlier SNP.

Rank	COG	Category	Count	Key Genes
1	S	Unknown function	328	Many adaptive genes lack annotation
2	E	Amino acid metabolism	230	Metabolic adaptation to C/N
3	T	Signal transduction	182	regB, regM, tacA, chvG, feuQ
4	P	Ion transport	142	modAB (Mo), nikAB (Ni), corB (Co/Mg)
5	G	Carbohydrate metabolism	136	Carbon source utilization
6	C	Energy production	135	Respiratory flexibility
7	M	Cell envelope	130	Membrane adaptation, EPS
8	I	Lipid metabolism	117	Membrane composition
9	K	Transcription	100	Regulatory adaptation
10	H	Coenzyme metabolism	90	Cofactor biosynthesis

COG Functional Enrichment

Hypergeometric enrichment analysis tested whether specific COG categories were over-represented among outlier genes (genes containing ≥1 outlier SNP) relative to the background distribution of all genes with detectable SNPs.

Statistical testing across 26 COG analysis revealed one category with significant enrichment at FDR < 0.01: signal transduction mechanisms (COG T) in Genome A showed 1.58-fold enrichment (51 outlier genes, FDR = 0.0026). One additional category showed suggestive enrichment at 0.05 < FDR < 0.15 (Table 13): signal transduction mechanisms in Genome B (44 outlier genes, 1.41-fold enrichment, FDR = 0.129).

Other functional categories showing modest representation among outliers include amino acid metabolism (COG E), ion transport (COG P), and energy production (COG C), though these did not reach statistical significance after multiple testing corrections. The high count of unknown function genes (COG S, 328 genes) among outliers likely reflects the large fraction of unannotated genes in bacterial genomes rather than selective enrichment, as this category showed no statistical enrichment (fold enrichment ~1.0 in all genomes).

Symbiosis Gene Enrichment

Hypergeometric enrichment analysis tested whether 164 symbiosis-related genes (nod, nif, fix, T3SS/T4SS, nop families) were over-represented among outlier genes. All four genomes showed fold enrichment below 1.0 (Table 14), indicating depletion rather than enrichment. Genome D showed marginal significance for depletion (p = 0.075, fold enrichment 0.38×).

Table 14. Symbiosis gene enrichment test results. All four genomes show fold enrichment below 1.0, indicating depletion rather than enrichment of outliers in symbiosis genes. Note: P-values calculated using two-tailed hypergeometric test, testing the appropriate tail based on direction of deviation (enrichment vs. depletion).

Genome	Total Genes	Symbiosis Genes (%)	Outlier Genes	Symbiosis Outliers (%)	Fold Enrichment	P-value	Significant
Genome A	1,561	17 (1.1%)	386	3 (17.6%)	0.71×	0.362	No
Genome B	1,478	13 (0.9%)	358	3 (23.1%)	0.95×	0.610	No
Genome C	868	17 (2.0%)	174	1 (5.9%)	0.29×	0.115	No
Genome D	1,316	21 (1.6%)	327	2 (9.5%)	0.38×	0.075	Marginal

Breaking down symbiosis genes by functional family (Figure 7) reveals heterogeneity in outlier rates. Regulatory genes (exoR, regB: 2/2 genes, 100% with outliers) and exopolysaccharide genes (exo family: 3/3 genes, 100% with outliers) show complete representation among outliers. In contrast, Type IV secretion (T4SS: 0/8 genes) and nodulation outer proteins (nol family: 0/2 genes) show complete conservation.

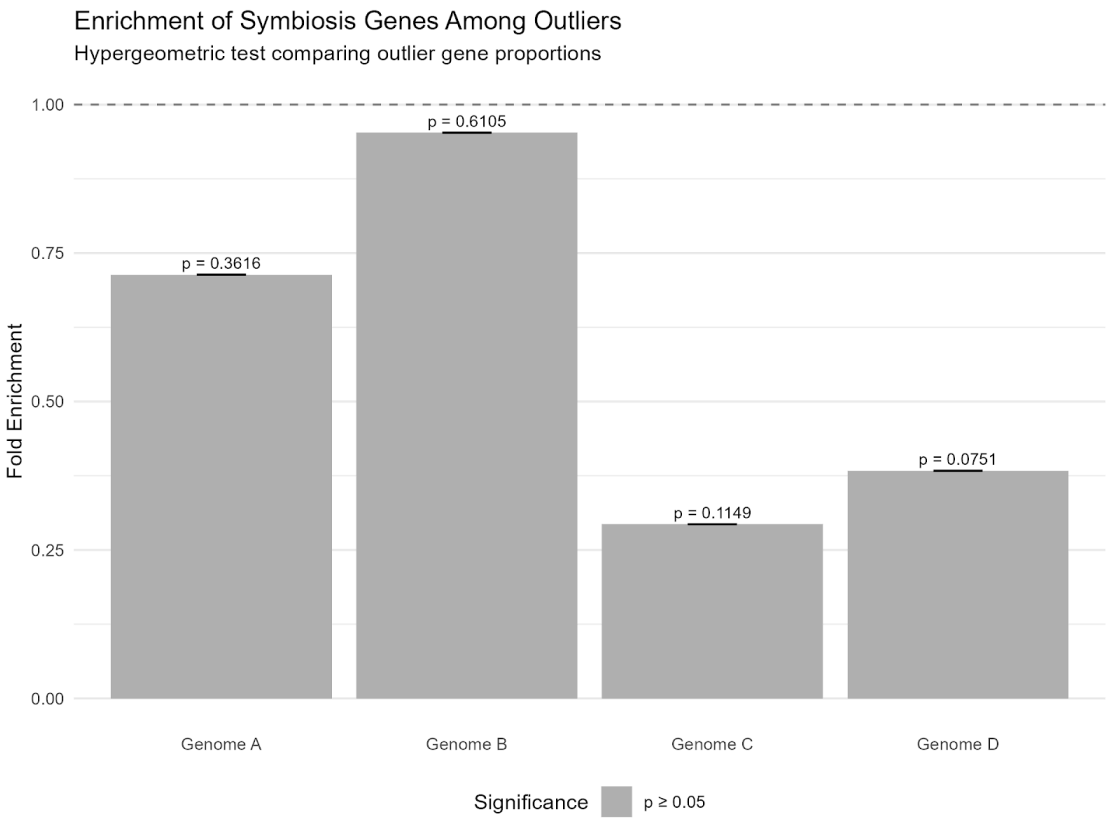

Figure 7. *Symbiosis gene enrichment across four reference genomes. Bar heights represent fold enrichment relative to the neutral expectation (dashed line at 1.0). All genomes show values below 1.0, indicating fewer outlier SNPs in symbiosis genes than expected by chance. P-values (displayed above bars) show that Genome D approaches statistical significance (p = 0.075), providing evidence that depletion in core symbiosis genes reflects genuine purifying selection rather than random chance.*

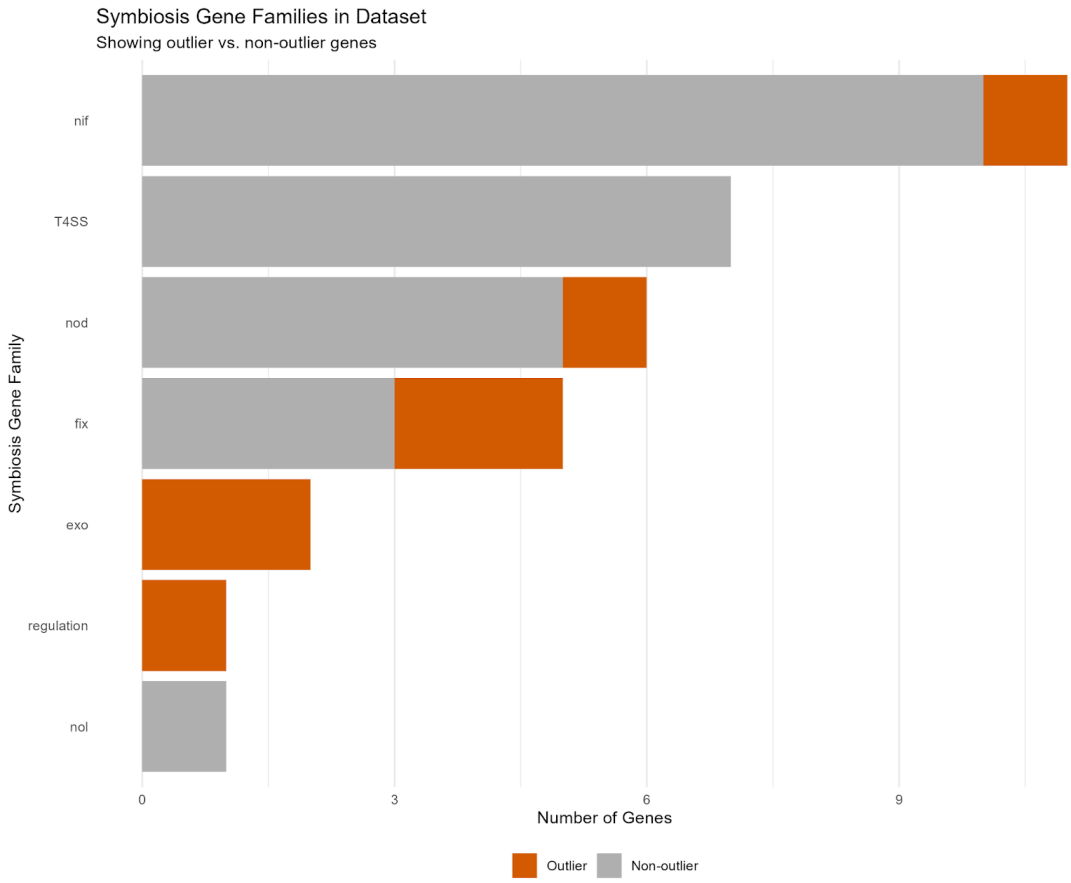

Figure 8. *Symbiosis gene family breakdown. Stacked bar plot shows number of genes per family, with orange indicating genes containing outlier SNPs and grey indicating non-outlier genes. The nif family is most abundant in our dataset (~10 genes), while exo and regulation families show 100% outlier rates despite small sample sizes. T4SS genes show complete conservation (no outliers), consistent with purifying selection on secretion machinery.*

Discussion

Australia harbors exceptional legume diversity, with thousands of native legume species that dominate many Australian landscapes (Barrett *et al.* 2015). These legumes thrive on the continent's nutrient-poor soils primarily through symbiotic associations with nitrogen-fixing rhizobia, among which *Bradyrhizobium* species predominate (Lafay & Burdon 1998). *Acacia* species form highly specialized root-nodule symbioses with *Bradyrhizobium*, enabling survival under conditions of nutrient limitation (Barrett *et al.* 2015; Lafay & Burdon 1998). Australia faces severe climate change pressures, including increasing temperatures, intensified heatwaves, and declining rainfall, with trends especially acute in southwestern Australia – a globally recognized biodiversity hotspot (Australian Academy of Science 2021). In this context, *Bradyrhizobium* populations alternating between free-living soil stages and symbiotic stages may

face strong selective pressures from both climate variability and soil chemistry gradients. Southwestern Australia provides a natural laboratory for studying these selective pressures, with pronounced environmental gradients in both climate and soil properties occurring within relatively short geographic distances. This study employed redundancy analysis (RDA) to identify genomic loci in *Bradyrhizobium* populations that show statistical associations with environmental gradients. Using 375 isolates analyzed across four independent reference genomes, we detected 1,859 outlier SNPs distributed across 1,628 unique genes.

Climate Variability as the Primary Driver

Climate variables emerge as the dominant environmental factors associated with genomic variation in *Bradyrhizobium* populations. By-margin permutation tests revealed that annual rainfall (mean variance = 3.99) and annual temperature (mean variance = 2.83) ranked first and second among all 13 environmental variables tested (Table 5). These climate effects substantially exceeded those of soil chemistry variables, with rainfall explaining nearly twice the variance of the strongest soil factor (carbon = 2.04). This result was not context dependent on the reference genome, strengthening the generality of the result.

Local soil chemistry is considered a more proximate factor, while larger scale climate factors are considered more distal and so their effects may also be mediated through proximate effects on the residing microbial community members (Gupta & Tiedje 2024). Based on this, our initial prediction was that soil chemistry was more likely to have a stronger effect in structuring putative adaptive variation. Previous analysis revealed no distinct population structure emerging at site level or higher scales, indicating high rates of genetic mixing that are indicative of low dispersal barriers to *Bradyrhizobium* (Simonsen et al. 2019), further strengthening our preliminary expectation on a larger effect of local soil chemistry factors. However, our results show climate variables as having a strong signal in all detected outlier loci. Given that there was a weak correlation between climate and soil chemistry variables, this provides a strong case on *Bradyrhizobium*'s capacity to adapt to climate variation. In southwestern Australia-biodiversity hotspot experiencing severe climate change including increasing temperatures, intensified heatwaves, and declining rainfall (Australian Academy of Science 2021)- our results strongly suggest that these large-scale climate gradients create persistent selective pressures on free-living soil bacteria during the non-symbiotic phase of the *Bradyrhizobium* life cycle.

Soil Chemistry: Important but Secondary

Soil chemistry variables also show significant associations with genomic variation, though their effects are generally weaker than climate factors. Among soil variables, soil carbon and gravel content show the strongest associations (mean variance 2.04 and 1.95 respectively, Table 5), followed by nitrate (1.30). Soil carbon likely reflects overall soil organic matter content influencing microbial habitat quality and nutrient availability during the free-living phase. Gravel content affects soil physical structure, water retention, and root accessibility – factors that influence both bacterial survival and host plant root distribution.

Nitrate (ranking 5th overall among 12 variables) shows moderate but consistent associations across all four reference genomes. This is biologically coherent: high soil nitrogen reduces plant dependence on bacterial nitrogen fixation, potentially altering the selective advantage of maintaining costly symbiotic machinery. However, this effect is substantially weaker than climate variables (nitrate variance = 1.30 vs. rainfall variance = 3.99).

Soil pH, despite its well-documented importance for rhizobial biology (Bordeleau & Prévost 1994), and having large general influence in structuring microbial communities (Gupta & Tiedje 2024), actually showed the weakest associations among all variables tested (mean variance 0.65, rank 12/12, Table 5), with no significant effects at $p \leq 0.01$ in any genome. This unexpected finding may reflect: (1) the relatively narrow pH range in the sampling area (most southwestern Australian soils are acidic to neutral) limiting selective differentiation, (2) pH effects operating primarily through indirect mechanisms – metal availability, nutrient solubility – already captured by other variables in the model, or (3) pH tolerance having evolved earlier in *Bradyrhizobium* evolutionary history such that standing genetic variation in pH response is now limited.

Functional Enrichment: Signal Transduction Mechanisms

To test whether specific functional categories were over-represented among outlier loci, we performed hypergeometric enrichment analysis on COG functional categories. Signal transduction mechanisms (COG category T) showed statistically significant enrichment in Genome A (FDR = 0.0026, 1.58-fold enrichment), with suggestive enrichment in Genome B (FDR = 0.129, 1.41-fold enrichment). This category includes two-component regulatory systems (e.g., regB/regA for oxygen response, fixL/fixJ for nitrogen fixation regulation, chvG/chvI for acid tolerance) and chemotaxis machinery.

The enrichment of signal transduction systems is consistent with expectations of adaptive variation along environmental variation: these genes represent the interface between genotype and environment, where genetic variation directly translates into differential environmental responses (Carroll 2005). Unlike structural genes with conserved biochemical functions, regulatory genes can evolve sensitivity thresholds and signaling outputs without disrupting core cellular processes, making them prime targets for adaptation to environmental heterogeneity. In the context of climate variability – the dominant selective force identified here – signal transduction systems allow bacteria to sense and respond to fluctuating conditions during free-living soil survival (Biswas *et al.* 2007; Bridges *et al.* 2022).

Symbiosis Genes: Purifying Selection and Host Plant Buffering

Given the ecological importance of nitrogen-fixing symbiosis, we specifically tested whether symbiosis-related genes showed enrichment among outliers, which would indicate if environmental factors act as putative selective agents on symbiosis genes. Symbiosis genes as a class showed consistent depletion across all four reference genomes (fold enrichment 0.29-0.95×). The near-significant depletion in Genome D (62% fewer outliers than expected) and strong depletion in Genome C (71% fewer outliers) indicate this pattern reflects genuine selective constraint. This depletion provides statistical evidence that some symbiosis genes, such as nod (nodulation), nif (nitrogen fixation), and fix (oxygen response) genes are under purifying selection, maintaining highly conserved sequences essential for the nitrogen-fixing mutualism. Mutations in these genes are more likely to be deleterious than adaptive, as loss of nodulation or nitrogen fixation capacity would severely disadvantage isolates in legume-dominated plant communities. Despite SNPs occurring within nifHDK genes, the genes which encode the subunits for the nitrogen-fixing enzyme nitrogenase, these did not show any significant association with the environment (as evident by a lack of outlier loci in these regions). This finding is consistent with theoretical predictions that genes providing obligate fitness benefits experience strong purifying selection. The strong depletion of outliers in symbiosis genes,

combined with the enrichment of signal transduction mechanisms among outliers, supports a model where environmental heterogeneity shapes the free-living phase of the life cycle while the symbiotic phase remains under functional constraint.

Cross-Genome Validation

The consistency of results across four independent reference genomes validates that observed patterns represent biological signals rather than assembly artifacts or reference bias. Despite distinct properties of each reference genome, all four showed: (1) similar levels of environmental variance explained (9-10%), (2) comparable numbers of outlier loci relative to genome size, (3) consistent ranking of environmental variables (climate > soil chemistry), (4) enrichment or suggestive enrichment of signal transduction mechanisms, and (5) depletion of outliers in core symbiosis genes.

The use of multiple genomes also revealed that statistical power varies with sample size: Genome A (503 outlier genes) showed the strongest enrichment signals, while Genome C (241 outlier genes) showed weaker statistical support despite similar biological patterns. This finding emphasizes the importance of the reference genome choice in influencing sample sizes for detecting functional enrichment in landscape genomics studies and limitations of genotype by environment association studies.

The application of von Mises-Fisher mixture models to cluster outlier loci by their environmental associations revealed that outliers group into approximately 10 distinct environmental response patterns (Figure 5). This finding suggests that adaptation involves coordinated changes in functionally related genes responding to different environmental pressures, rather than independent mutations scattered randomly across the genome. Some clusters show strong positive associations with climate variables (temperature, rainfall), others with soil chemistry (pH, nutrients), and others with soil physical properties (gravel content). This diversity of environmental profiles indicates that selection acts through multiple independent environmental gradients, each shaping different subsets of genomic variation.

Conclusions and Future Work

This study demonstrates that *Bradyrhizobium* populations show genomic signatures consistent with local adaptation to environmental gradients, with climate variability (particularly rainfall and temperature) as the dominant putative selective force. Soil chemistry factors, while significant, play a secondary role – with carbon and gravel content showing stronger effects than nitrate, and pH showing the weakest associations. The enrichment of signal transduction genes among outliers suggests that bacteria adapt to variable environments primarily through regulatory tuning. The conservation of core symbiosis genes indicates that the nitrogen-fixing mutualism itself is maintained by functional constraint, potentially buffered by host plant regulation of the nodule microenvironment.

While these results strongly suggest the pivotal role climate has in structuring adaptive variation in *Bradyrhizobium*, we also acknowledge the limit of this study, shared in common with other correlational genotype by environment association studies. Since this work focused on single nucleotide polymorphisms, we also acknowledge the importance of other genetic variation, such as structural variation (i.e. large deletions, insertions and rearrangements), as previous work on the *Bradyrhizobium* strains analyzed here also revealed highly prominent genome streamlining

and gene deletions with environmental stress (Simonsen 2022). Future work can consider experimental validation of candidate outlier loci through reciprocal transplant experiments or common garden studies, testing whether outlier genotypes confer fitness advantages in their native environments.

These findings have implications for predicting rhizobial responses to climate change in southwestern Australia and similar Mediterranean-climate regions experiencing increased temperature variability and rainfall decline. The extent to which standing genetic variation in regulatory loci can accommodate projected climate scenarios (2-4°C warming, 10-30% rainfall reduction by 2070; Australian Academy of Science, 2021) remains unclear.

Supporting Information: Please see the Supporting Information file attached separately for all supplementary information on methods and results.

Data Availability: The genome sequence data are available in GenBank under BioProject accession number PRJNA669073. Metadata is available here: 10.6084/m9.figshare.9698438. Code will be made available upon publication.

Acknowledgements: We would like to thank Dr. Janna Fierst, Dr. Paulo Olivas and Dr. Russell Dinnage for their suggestions and comments on this manuscript during its development.

5. REFERENCES

Allison, S.D. & Martiny, J.B.H. (2008). Colloquium paper: resistance, resilience, and redundancy in microbial communities. *Proc. Natl. Acad. Sci. U. S. A.*, 105 Suppl 1, 11512–11519.

Australian Academy of Science. (2021). *The risks to Australia of a 3°C warmer world*. Australian Academy of Science, Canberra.

Bardgett, R.D., Freeman, C. & Ostle, N.J. (2008). Microbial contributions to climate change through carbon cycle feedbacks. *ISME J.*, 2, 805–814.

Barrett, L.G., Bever, J.D., Bissett, A. & Thrall, P.H. (2015). Partner diversity and identity impacts on plant productivity in Acacia–rhizobial interactions. *J. Ecol.*, 103, 130–142.

Biswas, S., Van Dijck, P. & Datta, A. (2007). Environmental sensing and signal transduction pathways regulating morphopathogenic determinants of Candida albicans. *Microbiol. Mol. Biol. Rev.*, 71, 348–376.

Booth, T.H., Nix, H.A., Busby, J.R. & Hutchinson, M.F. (2014). bioclim: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Divers. Distrib.*, 20, 1–9.

Bordeleau, L.M. & Prévost, D. (1994). Nodulation and nitrogen fixation in extreme environments. *Plant Soil*, 161, 115–125.

Bridges, A.A., Prentice, J.A., Wingreen, N.S. & Bassler, B.L. (2022). Signal transduction network principles underlying bacterial collective behaviors. *Annu. Rev. Microbiol.*, 76, 235–257.

Capblancq, T. & Forester, B.R. (2021). Redundancy analysis: A Swiss Army Knife for landscape

genomics. *Methods Ecol. Evol.*, 12, 2298–2309.
Carroll, S.B. (2005). Evolution at two levels: on genes and form. *PLoS Biol.*, 3, e245.
Chia, M.-D. & Simonsen, A.K. (2021). Four Complete Genome Sequences for Bradyrhizobium sp. Strains Isolated from an Endemic Australian Acacia Legume Reveal Structural Variation. *Microbiol Resour Announc*, 10.
Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., *et al.* (2011). The variant call format and VCFtools. *Bioinformatics*, 27, 2156–2158.
Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., *et al.* (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, 10.
Denison, R.F. & Kiers, E.T. (2011). Life histories of symbiotic rhizobia and mycorrhizal fungi. *Curr. Biol.*, 21, R775–85.
Dinnage, R., Simonsen, A.K., Barrett, L.G., Cardillo, M., Raisbeck-Brown, N., Thrall, P.H., *et al.* (2018). Larger plants promote a greater diversity of symbiotic nitrogen-fixing soil bacteria associated with an Australian endemic legume: Acacia acuminata. *J. Ecol.*, 103, 130.
Fierer, N. & Jackson, R.B. (2006). The diversity and biogeography of soil bacterial communities. *Proc. Natl. Acad. Sci. U. S. A.*, 103, 626–631.
Forester, B.R., Lasky, J.R., Wagner, H.H. & Urban, D.L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. *Mol. Ecol.*, 27, 2215–2233.
Galloway, J.N., Townsend, A.R., Erisman, J.W., Bekunda, M., Cai, Z., Freney, J.R., *et al.* (2008). Transformation of the nitrogen cycle: recent trends, questions, and potential solutions. *Science*, 320, 889–892.
Garrison, E. & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]*.
Griffiths, B.S. & Philippot, L. (2013). Insights into the resistance and resilience of the soil microbial community. *FEMS Microbiol. Rev.*, 37, 112–129.
Gupta, V.V.S.R. & Tiedje, J.M. (2024). Ranking environmental and edaphic attributes driving soil microbial community structure and activity with special attention to spatial and temporal scales. *mLife*, 3, 21–41.
Kowarik, A. & Templ, M. (2016). Imputation with the R package VIM. *J. Stat. Softw.*, 74.
Lafay, B. & Burdon, J.J. (1998). Molecular diversity of rhizobia occurring on native shrubby legumes in southeastern australia. *Appl. Environ. Microbiol.*, 64, 3989–3997.
Legendre, P., Legendre, L. & Legendre, L.F. (2012). *Numerical Ecology*. Developments in Environmental Modelling. 3rd edn. Elsevier.
Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, 27, 718–719.
Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*.
Masson-Boivin, C., Giraud, E., Perret, X. & Batut, J. (2009). Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes? *Trends Microbiol.*, 17, 458–466.
Poole, P., Ramachandran, V. & Terpolilli, J. (2018). Rhizobia: from saprophytes to endosymbionts. *Nat. Rev. Microbiol.*
Rellstab, C., Gugerli, F., Eckert, A.J., Hancock, A.M. & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.*, 24, 4348–4370.

Remigi, P., Zhu, J., Young, J.P.W. & Masson-Boivin, C. (2016). Symbiosis within Symbiosis: Evolving Nitrogen-Fixing Legume Symbionts. *Trends Microbiol.*, 24, 63–75.

Rousk, J., Bååth, E., Brookes, P.C., Lauber, C.L., Lozupone, C., Caporaso, J.G., *et al.* (2010). Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J.*, 4, 1340–1351.

Savolainen, O., Lascoux, M. & Merilä, J. (2013). Ecological genomics of local adaptation. *Nat. Rev. Genet.*, 14, 807–820.

Simonsen, A.K. (2022). Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. *ISME J.*, 16, 423–434.

Simonsen, A.K., Barrett, L.G., Thrall, P.H. & Prober, S.M. (2019). Novel model-based clustering reveals ecologically differentiated bacterial genomes across a large climate gradient. *Ecol. Lett.*, 22, 2077–2086.

Wickham, H. (2016). Data Analysis. In: *Use R!* Springer International Publishing, Cham, pp. 189–201.

Zhang, Y., Ku, Y.-S., Cheung, T.-Y., Cheng, S.-S., Xin, D., Gombeau, K., *et al.* (2024). Challenges to rhizobial adaptability in a changing climate: Genetic engineering solutions for stress tolerance. *Microbiol. Res.*, 288, 127886.

Supporting Material

Section S1: Environmental Variable Transformations

Correlation Analysis and Variable Selection

High correlation between environmental variables can cause multicollinearity issues in RDA. We performed hierarchical clustering of variables based on correlation distance:

$$d_{ij} = 1 - |r_{ij}|$$

where r_{ij} is the Pearson correlation between variables i and j. Variables with high correlation (|r| > 0.7) were removed. Variables removed included soil texture (correlated with gravel), soil aluminium (correlated with pH), soil pH_h (correlated with pH_c), soil potassium_t (correlated with other nutrients), soil moisture (correlated with climate variables), soil iron, sulphur, sodium (correlated with other metals), soil manganese, phosphorus, boron (correlated with other nutrients), and soil calcium (correlated with magnesium and pH).

Yeo-Johnson Transformation

The Yeo-Johnson transformation is a generalization of the Box-Cox transformation that handles negative values and zeros. The transformation is defined as:

$$y(\lambda) = \begin{cases} \dfrac{(y+1)^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y+1) & \text{if } \lambda = 0, y \geq 0 \\ -\dfrac{(-y+1)^{2-\lambda} - 1}{2-\lambda} & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y+1) & \text{if } \lambda = 2, y < 0 \end{cases}$$

The optimal λ parameter is determined automatically for each variable using maximum likelihood estimation. After transformation, all variables were normalized (centered and scaled) to have mean = 0 and standard deviation = 1.

Section S2: Mahalanobis Distance for Outlier Detection

To identify outlier SNPs in multivariate loading space, we used Mahalanobis distance, which accounts for the correlation structure among RDA axes. The squared Mahalanobis distance is defined as:

$$D^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

where: - x: SNP's loading vector across significant RDA axes - μ: Mean loading vector across all SNPs - Σ: Covariance matrix of loadings - D^2: Squared Mahalanobis distance

Under the null hypothesis of neutral evolution, D^2 follows a chi-squared distribution with degrees of freedom equal to the number of RDA axes:

$$D^2 \sim \chi^2_{df}$$

P-values were calculated as:

$$p = P(\chi^2_{df} \geq D^2)$$

SNPs with $p \leq 0.05$ were classified as outliers.

Rationale for Mahalanobis distance: This approach is more statistically rigorous than examining individual axes independently because it accounts for correlation structure among RDA axes, provides a multivariate test statistic with known distribution, and has greater power to detect coordinated responses across multiple environmental gradients.

Section S3: Hemisphere Folding and von Mises-Fisher Mixture Models

The Arbitrary Allele Coding Problem

SNP allele coding (0/1 for reference/alternate allele) is arbitrary. Two SNPs with opposite RDA loadings (e.g., [+0.6, -0.3] vs. [-0.6, +0.3]) may actually represent the same biological pattern if their alleles are coded differently. Standard clustering methods would place these in different clusters despite identical environmental associations.

Hemisphere Folding Transformation

Before clustering, we transform loading vectors to account for arbitrary directionality:

Step 1 - Unit normalization: Scale each loading vector to unit length:

$$v_{norm} = \frac{v}{||v||}$$

Step 2 - Hemisphere folding: If the sum of vector components is negative, reflect the vector:

$$v_{folded} = \{v_{norm} \; if \sum_i v_i \geq 0 \quad -v_{norm} \; if \sum_i v_i < 0$$

This ensures that vectors v and $-v$ map to the same point, preserves relationships between environmental variables, and converts the n-sphere to a half n-sphere. This essentially removes direction information, appropriate because of the arbitrariness of the 0 or 1 coding of snps.

von Mises-Fisher Distribution

The von Mises-Fisher (vMF) distribution is the appropriate probability distribution for unit vectors on a sphere:

$$f(x|\mu,\kappa) = C_p(\kappa)exp(\kappa\mu^T x)$$

where: - x: Unit vector (SNP loading) - μ: Mean direction (cluster center), a unit vector - κ: Concentration parameter (analogous to inverse variance) - $C_p(\kappa)$: Normalization constant

The concentration parameter κ has direct interpretation: - $\kappa = 0$: Uniform distribution on sphere (no clustering) - $\kappa \to \infty$: Concentrated at point (perfect clustering) - Typical values: $\kappa = 1 - 10$ (loose clusters) to $\kappa > 100$ (tight clusters)

Mixture Model

The mixture model combines multiple vMF distributions:

$$P(x) = \sum_{k=1}^{K} \alpha_k f(x|\mu_k, \kappa_k)$$

where α_k are mixing proportions (sum to 1). Model parameters are estimated using the Expectation-Maximization (EM) algorithm as implemented in the movMF R package.

Model Selection Using BIC

The Bayesian Information Criterion (BIC) balances model fit and complexity:

$$BIC = -2logL + klogn$$

where: - L: Model likelihood - k: Number of parameters - n: Number of observations

Lower BIC indicates better model. We tested K = 2 to K = 40 and selected K using the elbow method at 95% of total BIC improvement:

$$K_{opt} = arg \min_k \left\{ k: \frac{BIC_{max} - BIC_k}{BIC_{max} - BIC_{min}} \geq 0.95 \right\}$$

LOESS smoothing (span = 0.3) was applied to reduce noise from convergence failures.

Section S4: Cluster Matching Across Runs Using the Hungarian Algorithm

The Cluster Matching Problem

After independent clustering in each run, cluster labels are arbitrary (Cluster 1 in Genome A ≠ Cluster 1 in Genome B biologically). To enable cross-run comparisons and consistent visualization, we need to optimally match clusters based on similarity.

Cluster Feature Extraction

For each cluster in each run, we extract:

1. **Environmental projection vector:** The cluster center (mean direction μ_k) projected onto the original environmental variable space
2. **Concentration parameter:** κ_k indicating cluster tightness

Distance Matrix Construction

We calculate a pairwise distance matrix between clusters from different runs:

$$D_{ij} = w_{env} \cdot d_{cosine}(env_i, env_j) + w_{\kappa} \cdot |\kappa_i - \kappa_j|$$

where: - d_{cosine}: Cosine distance between environmental projections - $w_{env} = 0.7$: Weight for environmental similarity - $w_{\kappa} = 0.3$: Weight for concentration similarity - Normalized κ values to have comparable scale

Hungarian Algorithm

The Hungarian algorithm (Kuhn-Munkres algorithm) finds the optimal 1-to-1 assignment that minimizes total cost:

Input: Distance matrix D (Genome B clusters × Genome A clusters)

Output: Assignment minimizing $\sum_i D_{i,assignment(i)}$

Procedure: 1. Designate Genome A as reference 2. For Genome B, Genome C, Genome D: - Compute distance matrix between that run's clusters and Genome A clusters - Apply Hungarian algorithm to find minimum-cost assignment - Relabel clusters to match Genome A

Benefits: - Globally optimal matching (not greedy) - Ensures 1-to-1 correspondence - Consistent cluster coloring across all plots - Enables quantitative comparison of cluster properties

Section S5: Supplementary Methods for Circular Parallel Coordinates Plots

Circular parallel coordinates plots (also called radar plots or spider plots) were used to visualize environmental associations of clusters. Each environmental variable is assigned an angular

position around a circle. For each SNP or cluster center, the value on each environmental variable is plotted as radial distance from the origin. Positive values extend outward, negative values extend inward (toward center).

Construction: 1. Arrange 13 environmental variables at equal angular intervals: $\theta_i = \frac{2\pi \cdot i}{13}$ 2. For each SNP or cluster center with loading vector v: - Project onto environmental space if needed - Plot point at angle θ_i and radius $r_i = v_i$ 3. Connect points to form a closed polygon 4. Overlay individual SNP lines (colored by cluster) and cluster centers (thick black lines)

Interpretation: - **Shape:** Overall pattern of environmental associations - **Area:** Magnitude of environmental response (larger = stronger associations) - **Asymmetry:** Differential associations with different environmental axes - **Within-cluster spread:** Heterogeneity of SNP responses within cluster

Supporting Tables

Table S1. Complete COG (Clusters of Orthologous Groups) functional category enrichment results across four *Bradyrhizobium* reference genomes.

For each genome and COG category, we report the number of outlier loci (n_outlier), background loci (n_background), fold enrichment, Fisher's exact test p-value, and false discovery rate (FDR). Enrichment was tested using Fisher's exact test comparing the proportion of outlier loci in each category to the background proportion. Categories are ordered by genome then p-value within genome.

Genome	COG	Category	outlier	background	Fold enrichment	p_value	FDR
A	T	Signal transduction mechanisms	51	129	1.58	1.20e-04	2.63e-03
A	E	Amino acid transport and metabolism	54	257	0.84	6.40e-02	4.75e-01
A	M	Cell wall/membrane/envelope biogenesis	36	117	1.23	8.70e-02	4.75e-01
A	K	Transcription	22	112	0.79	1.06e-01	4.75e-01
A	P	Inorganic ion transport and metabolism	35	116	1.21	1.14e-01	4.75e-01

Genome	COG	Category	outlier	background	Fold enrichment	p_value	FDR
A	D	Cell cycle control, cell division, chromosome partitioning	3	23	0.52	1.35e-01	4.75e-01
A	U	Intracellular trafficking, secretion, and vesicular transport	6	35	0.69	1.89e-01	4.75e-01
A	I	Lipid transport and metabolism	34	120	1.13	2.22e-01	4.75e-01
A	H	Coenzyme transport and metabolism	22	100	0.88	2.80e-01	4.75e-01
A	J	Translation, ribosomal structure and biogenesis	19	87	0.87	2.88e-01	4.75e-01
A	V	Defense mechanisms	8	40	0.80	2.97e-01	4.75e-01
A	Q	Secondary metabolites biosynthesis, transport and catabolism	21	75	1.12	3.12e-01	4.75e-01
A	F	Nucleotide transport and metabolism	10	47	0.85	3.43e-01	4.75e-01
A	G	Carbohydrate transport and metabolism	38	143	1.06	3.60e-01	4.75e-01
A	N	Cell motility	11	50	0.88	3.79e-01	4.75e-01
A	O	Posttranslational modification, protein turnover, chaperones	19	70	1.09	3.83e-01	4.75e-01

Genome	COG	Category	outlier	background	Fold enrichment	p_value	FDR
A	-	NA	30	113	1.06	3.85e-01	4.75e-01
A	C	Energy production and conversion	36	152	0.95	3.89e-01	4.75e-01
A	S	Function unknown	89	348	1.02	4.19e-01	4.85e-01
A	L	Replication, recombination and repair	17	72	0.94	4.52e-01	4.98e-01
A	B	Chromatin structure and dynamics	0	2	0.00	5.62e-01	5.62e-01
A	Z	Cytoskeleton	0	2	0.00	5.62e-01	5.62e-01
B	T	Signal transduction mechanisms	44	132	1.41	5.86e-03	1.29e-01
B	C	Energy production and conversion	22	133	0.70	2.56e-02	2.82e-01
B	L	Replication, recombination and repair	10	69	0.61	4.12e-02	3.02e-01
B	Z	Cytoskeleton	2	2	4.22	5.60e-02	3.08e-01
B	V	Defense mechanisms	11	31	1.50	9.27e-02	4.08e-01
B	H	Coenzyme transport and metabolism	18	95	0.80	1.62e-01	4.33e-01
B	U	Intracellular trafficking, secretion, and vesicular transport	8	46	0.73	2.03e-01	4.33e-01
B	D	Cell cycle control, cell division,	7	21	1.41	2.10e-01	4.33e-01

Genome	COG	Category chromosome partitioning	outlier	background	Fold enrichment	p_value	FDR
B	K	Transcription	32	119	1.14	2.27e-01	4.33e-01
B	O	Posttranslational modification, protein turnover, chaperones	17	60	1.20	2.36e-01	4.33e-01
B	J	Translation, ribosomal structure and biogenesis	25	92	1.15	2.44e-01	4.33e-01
B	N	Cell motility	8	44	0.77	2.51e-01	4.33e-01
B	E	Amino acid transport and metabolism	53	242	0.93	2.72e-01	4.33e-01
B	S	Function unknown	72	324	0.94	2.75e-01	4.33e-01
B	F	Nucleotide transport and metabolism	8	41	0.82	3.37e-01	4.84e-01
B	P	Inorganic ion transport and metabolism	31	124	1.06	3.96e-01	4.84e-01
B	M	Cell wall/membrane/envelope biogenesis	27	121	0.94	4.06e-01	4.84e-01
B	I	Lipid transport and metabolism	33	133	1.05	4.10e-01	4.84e-01
B	G	Carbohydrate transport and metabolism	32	141	0.96	4.34e-01	4.84e-01
B	-	NA	29	118	1.04	4.43e-01	4.84e-01
B	Q	Secondary metabolites	19	77	1.04	4.62e-01	4.84e-01

Genome	COG	Category	outlier	background	Fold enrichment	p_value	FDR
		biosynthesis, transport and catabolism					
B	B	Chromatin structure and dynamics	0	2	0.00	5.82e-01	5.82e-01
C	I	Lipid transport and metabolism	9	77	0.56	2.19e-02	2.39e-01
C	M	Cell wall/membrane/envelope biogenesis	20	62	1.54	2.24e-02	2.39e-01
C	V	Defense mechanisms	10	27	1.76	3.94e-02	2.39e-01
C	O	Posttranslational modification, protein turnover, chaperones	3	35	0.41	4.35e-02	2.39e-01
C	N	Cell motility	9	28	1.53	1.12e-01	4.12e-01
C	T	Signal transduction mechanisms	19	69	1.31	1.12e-01	4.12e-01
C	H	Coenzyme transport and metabolism	12	45	1.27	2.18e-01	5.60e-01
C	E	Amino acid transport and metabolism	37	157	1.12	2.26e-01	5.60e-01
C	G	Carbohydrate transport and metabolism	16	92	0.83	2.29e-01	5.60e-01
C	L	Replication, recombination and repair	11	43	1.22	2.79e-01	5.99e-01
C	P	Inorganic ion transport and metabolism	17	93	0.87	3.01e-01	5.99e-01

Genome	COG	Category	outlier	background	Fold enrichment	p_value	FDR
C	F	Nucleotide transport and metabolism	3	20	0.71	3.67e-01	5.99e-01
C	U	Intracellular trafficking, secretion, and vesicular transport	6	25	1.14	4.32e-01	5.99e-01
C	J	Translation, ribosomal structure and biogenesis	12	53	1.08	4.37e-01	5.99e-01
C	K	Transcription	12	62	0.92	4.45e-01	5.99e-01
C	D	Cell cycle control, cell division, chromosome partitioning	3	18	0.79	4.58e-01	5.99e-01
C	S	Function unknown	37	173	1.02	4.80e-01	5.99e-01
C	Q	Secondary metabolites biosynthesis, transport and catabolism	9	46	0.93	4.90e-01	5.99e-01
C	-	NA	9	45	0.95	5.21e-01	6.04e-01
C	C	Energy production and conversion	19	91	0.99	5.51e-01	6.06e-01
C	B	Chromatin structure and dynamics	0	1	0.00	7.90e-01	7.90e-01
C	Z	Cytoskeleton	0	1	0.00	7.90e-01	7.90e-01
D	V	Defense mechanisms	13	30	1.76	1.82e-02	4.01e-01
D	C	Energy production and conversion	27	141	0.78	6.74e-02	4.33e-01

Genome	COG	Category	outlier	background	Fold enrichment	p_value	FDR
D	H	Coenzyme transport and metabolism	25	80	1.27	1.03e-01	4.33e-01
D	L	Replication, recombination and repair	9	54	0.68	1.08e-01	4.33e-01
D	U	Intracellular trafficking, secretion, and vesicular transport	4	29	0.56	1.22e-01	4.33e-01
D	F	Nucleotide transport and metabolism	12	35	1.39	1.28e-01	4.33e-01
D	D	Cell cycle control, cell division, chromosome partitioning	9	25	1.46	1.38e-01	4.33e-01
D	N	Cell motility	11	34	1.31	1.94e-01	4.61e-01
D	M	Cell wall/membrane/envelope biogenesis	23	108	0.86	2.40e-01	4.61e-01
D	B	Chromatin structure and dynamics	1	1	4.06	2.46e-01	4.61e-01
D	E	Amino acid transport and metabolism	56	210	1.08	2.57e-01	4.61e-01
D	K	Transcription	21	98	0.87	2.66e-01	4.61e-01
D	Q	Secondary metabolites biosynthesis, transport and catabolism	18	63	1.16	2.72e-01	4.61e-01
D	Z	Cytoskeleton	0	4	0.00	3.22e-01	5.01e-01

Genome	COG	Category	outlier	background	Fold enrichment	p_value	FDR
D	O	Posttranslational modification, protein turnover, chaperones	13	59	0.89	3.84e-01	5.01e-01
D	I	Lipid transport and metabolism	27	104	1.05	4.11e-01	5.01e-01
D	-	NA	23	89	1.05	4.35e-01	5.01e-01
D	S	Function unknown	71	294	0.98	4.49e-01	5.01e-01
D	J	Translation, ribosomal structure and biogenesis	20	85	0.96	4.63e-01	5.01e-01
D	T	Signal transduction mechanisms	25	105	0.97	4.73e-01	5.01e-01
D	G	Carbohydrate transport and metabolism	29	115	1.02	4.78e-01	5.01e-01
D	P	Inorganic ion transport and metabolism	24	99	0.98	5.18e-01	5.18e-01

Supporting Figures

Figure S1. Hierarchical clustering of environmental variables based on absolute correlation. Variables clustering together indicate high correlation. The dendrogram height represents 1 - |correlation|.

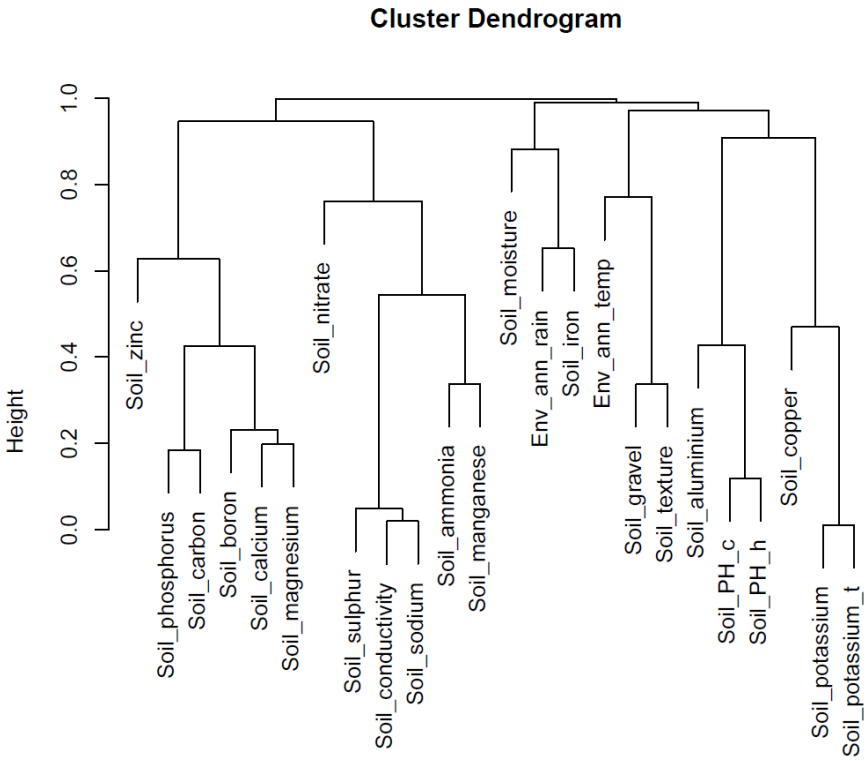

Figure S2. Pairwise relationships between transformed environmental variables. Diagonal shows density distributions, upper triangle shows Pearson correlations, lower triangle shows scatter plots. All variables are normalized (mean=0, SD=1) after Yeo-Johnson transformation.