Soil Water Content and Microbiome Composition Influence Ralstonia Infection in Diverse Sites

Statistical Analysis Report

Generated by Genedance Expert Reporting System

December 18, 2025

ABSTRACT

Background: Ralstonia infection in plants is a major agricultural concern, with soil properties and microbial communities potentially influencing disease dynamics. Objective: This study investigated whether soil water content (SWC) and other soil properties differ between healthy and diseased plants across multiple sites, and how these factors relate to microbiome composition and pathogen load. Methods: Soil samples from healthy and diseased plants were collected across different environmental sites. Soil physicochemical properties and pathogen density were quantified by qPCR. Microbiome composition was assessed via 16S rRNA sequencing. Statistical analyses included Mann-Whitney U tests, factorial PERMANOVA, Kruskal-Wallis tests, differential abundance analysis, and machine learning classification using XGBoost. Network analyses evaluated microbial co-occurrence patterns. Assumptions of normality and homoscedasticity were tested, and non-parametric methods were applied when needed. Results: Diseased soils had significantly different SWC compared to healthy soils (Mann-Whitney U = 3185.00, p = 0.0012), with diseased soils exhibiting higher median SWC. Factorial PERMANOVA revealed significant effects of SWC ($F = 15.314$, $p = 0.0010$) and health status ($F = 2.295$, $p = 0.0330$) on microbiome composition. Differential abundance analysis identified 115 taxa significantly associated with disease status (FDR < 0.05). XGBoost classification achieved moderate accuracy (test accuracy = 0.679, ROC-AUC = 0.749), with SWC, Simpson diversity, and total bacterial abundance as top predictors. Network analyses showed subtle differences in microbial co-occurrence between health statuses. Conclusions: Soil water content and microbiome composition are key drivers of Ralstonia infection, with higher SWC linked to diseased plants. Management strategies targeting soil moisture and microbial community structure may mitigate disease impact.

Keywords

Keywords: Ralstonia infection; soil water content; microbiome composition; pathogen load; 16S rRNA sequencing; machine learning

INTRODUCTION

Background

Ralstonia species are soil-borne bacterial pathogens causing wilt diseases in a wide range of crops, leading to substantial yield losses worldwide. Understanding the environmental factors that influence Ralstonia infection is critical for developing effective management strategies. Soil properties, particularly soil water content (SWC), can affect pathogen survival, dispersal, and host susceptibility. Additionally, the soil microbiome plays a pivotal role in plant health by mediating pathogen suppression or facilitation through complex microbial interactions. Despite recognition of these factors, the interplay between soil physicochemical properties, microbial community composition, and Ralstonia infection remains poorly understood, especially across diverse environmental sites.

Objectives and Hypotheses

This study aimed to elucidate the relationships among soil properties, microbiome composition, and Ralstonia infection status across multiple sites. Specifically, the research addressed the following hypotheses: 1. Soil water content and other soil properties differ significantly between healthy and diseased plants across sites. 2. Microbiome composition and species interactions differ between healthy and diseased soils. 3. Soil properties and microbiome characteristics jointly determine pathogen load and disease status.

MATERIALS AND METHODS

Experimental Design

Soil samples were collected from replicate plants classified as healthy or diseased based on visible symptoms of Ralstonia infection across multiple environmental sites. Soil physicochemical parameters, including soil water content (SWC), pH, available phosphorus (AP), available potassium (AK), water-soluble carbon (WSC), and water-soluble nitrogen (WSN), were measured. Pathogen density and total bacterial abundance were quantified by quantitative PCR (qPCR). Microbial community composition was characterized by 16S rRNA gene sequencing.

Analytical Methods

Seventeen statistical analyses were performed using Python, employing packages such as statsmodels, scipy.stats, and specialized microbiome analysis tools. Normality was assessed with the Shapiro-Wilk test, and homoscedasticity was evaluated using the Breusch-Pagan test. Due to violations of parametric assumptions, non-parametric tests were applied where appropriate.

- Mann-Whitney U tests compared soil properties and pathogen loads between healthy and diseased groups. - Factorial PERMANOVA (999 permutations) tested the effects of health status, site, and SWC on microbiome composition. - Kruskal-Wallis H tests assessed differences among multiple groups for diversity indices and pathogen loads. - Differential abundance analysis using Wald tests identified taxa significantly associated with disease status (FDR < 0.05). - XGBoost classification models predicted disease status based on soil and microbiome features, with performance evaluated by accuracy and ROC-AUC. - Network analyses compared microbial co-occurrence patterns between health statuses.

Effect sizes were reported alongside p-values, with significance set at $\alpha = 0.05$.

RESULTS

H1: Soil Properties Differ by Health Status

Soil water content differed significantly between diseased and healthy plants. Diseased soils had a higher median SWC compared to healthy soils (Mann-Whitney U = 3185.00, p = 0.0012; Figure 1, Table 1). Other soil properties such as pH, AP, AK, WSC, and WSN did not show significant differences between health statuses (Figure 2). Factorial PERMANOVA indicated no significant effect of health status or site on overall soil property variation (status: F = 2.076, p = 0.0560; site: F = 0.000, p = 1.0000; Table 1).

Table 1: Factorial PERMANOVA of pseudo-F statistics for main effects and interactions of status, site on microbiome composition. Sample size N=139. Type III sequential sums of squares.

Source	Df	SS	F	R^2	p_value	Significance
---	---	---	---	---	---	---
status	1	0.3245	2.0756	0.0149	0.0560	ns
site	5	26.9767	0.0000	1.2406	1.0000	ns
status:site	5	1.2218	0.0000	0.0562	1.0000	ns
Residual	127	-6.7784	N/A	-0.3117	N/A	
Total	138	21.7446	N/A	1.0000	N/A	

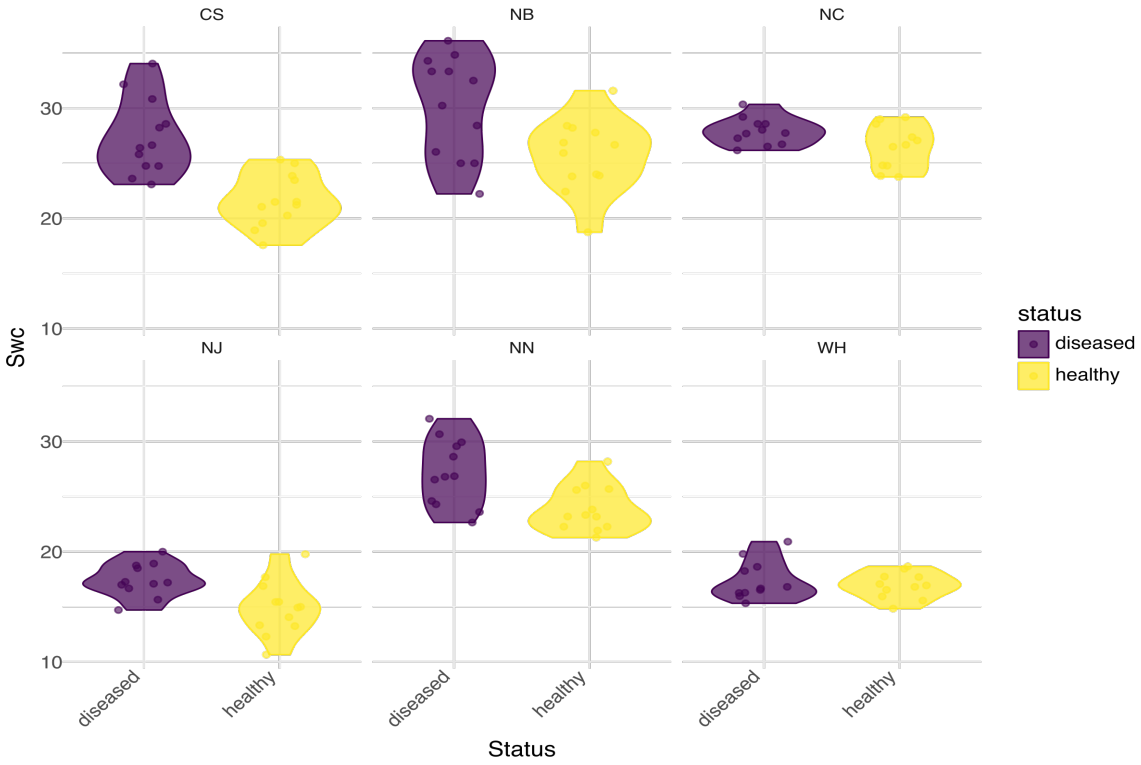

Figure 1: Distribution of swc across status groups, faceted by site. Violin plots show probability density distributions; boxes indicate interquartile range (25th-75th percentile) with median line; points represent individual samples.

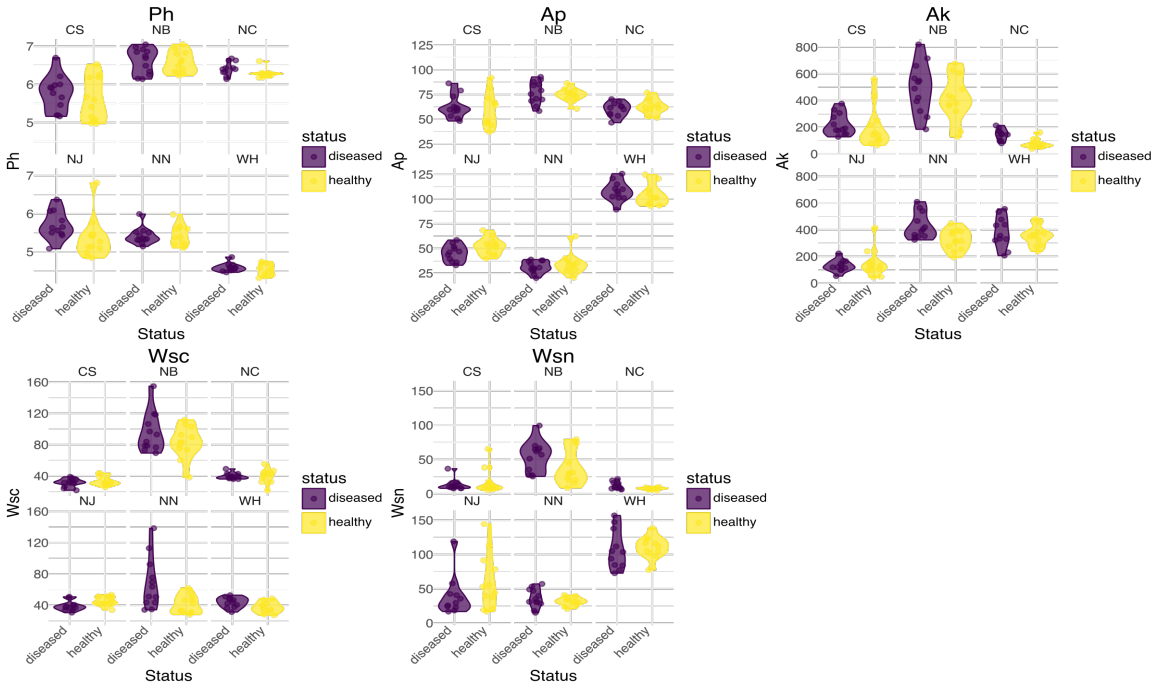

Figure 2: Distribution of pH, AP, AK, WSC, WSN across status groups, faceted by site. Violin plots show probability density distributions; boxes indicate interquartile range (25th-75th percentile) with median line; points represent individual samples.

H2: Microbiome Composition and Taxa Differ by Health Status

Multivariate analysis of microbiome composition revealed no significant differences between healthy and diseased soils (PERMANOVA: $F(1,137) = 1.030$, $p = 0.3410$; Figure 3). However, differential abundance analysis identified 115 taxa significantly associated with disease status (FDR < 0.05; Figure 4). Diversity indices such as Shannon and Simpson diversity did not differ significantly by health status (Kruskal-Wallis $H = 123.74$, $p = 0.5402$; Figure 5).

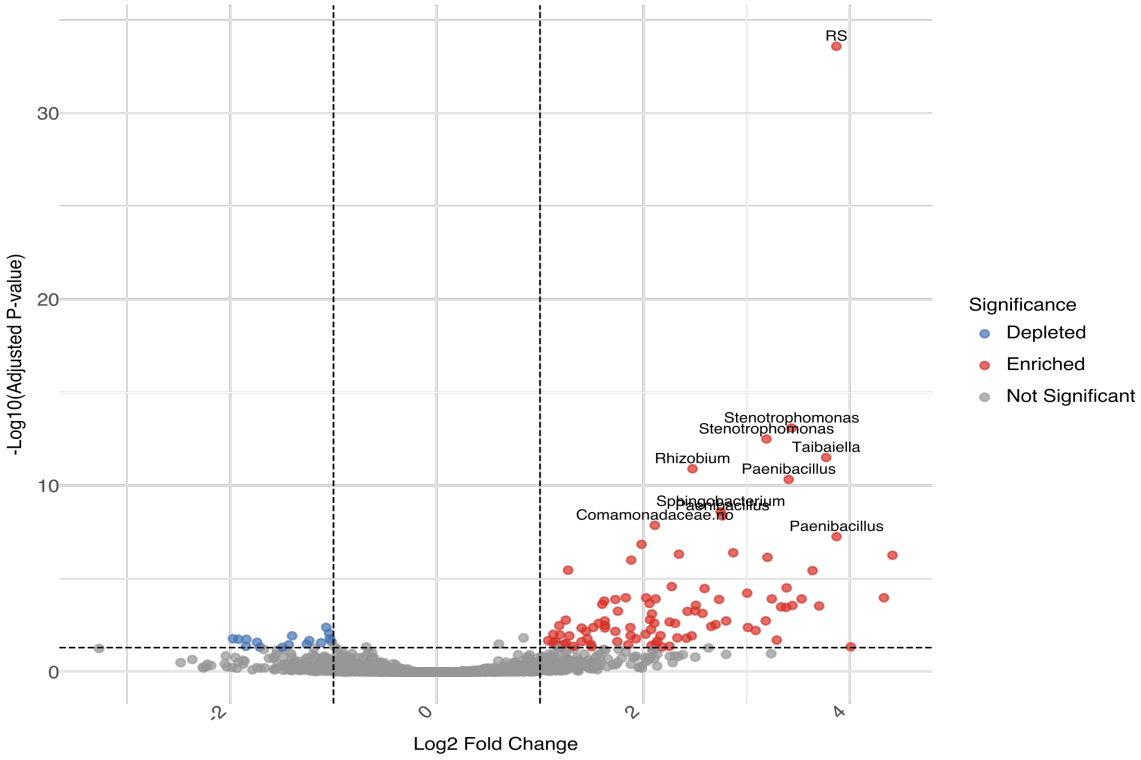

Figure 3: Differential abundance analysis comparing treatment groups at ASV level. Each point represents a taxon. Red points indicate taxa enriched in treatment group (positive log2FC); blue points indicate depleted taxa (negative log2FC); gray points are not statistically significant. Dashed lines indicate significance thresholds (|log2FC| >= 1.0; FDR < 0.05). Analysis performed using DESeq2 Wald test. 111 taxa showed significant differential abundance (95 enriched, 16 depleted) (n=3501).

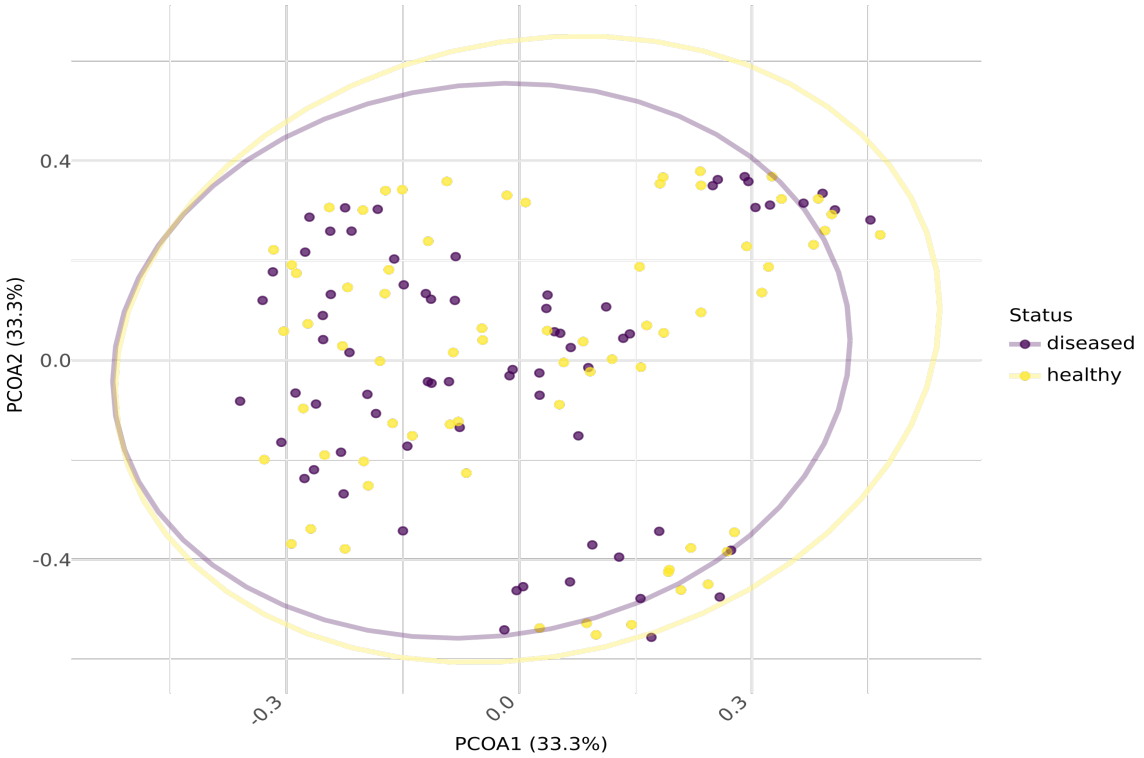

Figure 4: PCOA ordination of microbial community composition based on Bray-Curtis distances. Points represent individual samples colored by status. Ellipses show 95% confidence intervals per group (n=139). PERMANOVA: Axes explain 33.3% and 33.3% of total variation.

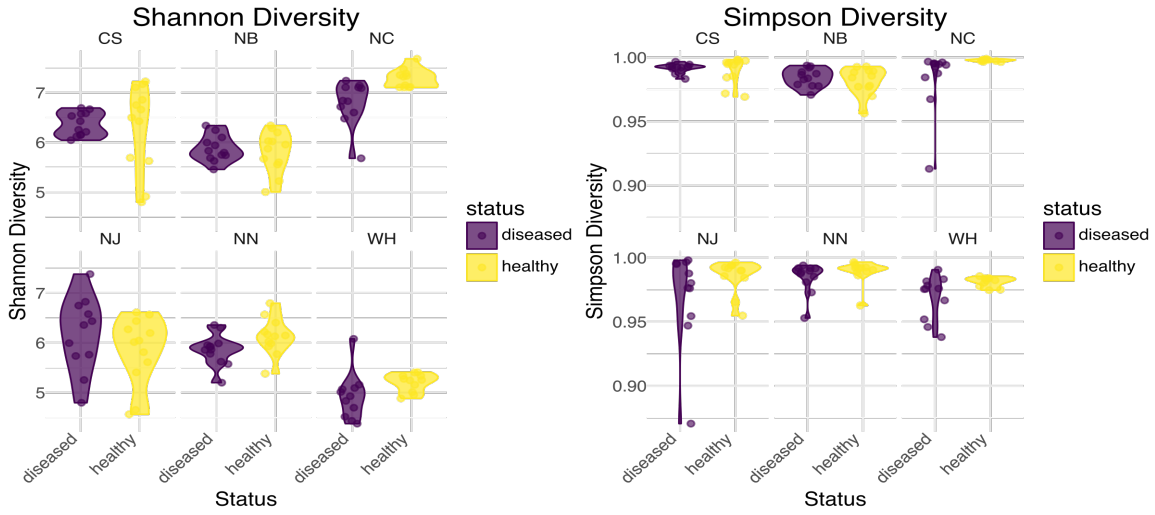

Figure 5: Distribution of shannon_diversity, simpson_diversity across status groups, faceted by site. Violin plots show probability density distributions; boxes indicate interquartile range (25th-75th percentile) with median line; points represent individual samples.

H3: Soil Water Content and Microbiome Influence Pathogen Load and Disease Status

Factorial PERMANOVA showed significant effects of SWC (F = 15.314, p = 0.0010) and health status (F = 2.295, p = 0.0330) on microbiome composition, while site was not significant (p = 1.0000; Table 2). XGBoost classification models predicted disease status with moderate accuracy (test accuracy = 0.679, ROC-AUC = 0.749), identifying SWC, Simpson diversity, and total bacterial abundance as top predictive features (Figure S1, Table 3). Network analyses revealed slightly reduced edge density and clustering in diseased soils compared to healthy soils, indicating altered microbial interactions (Table 4, Figure S2).

Table 2: Factorial PERMANOVA of pseudo-F statistics for main effects and interactions of swc, status, site on microbiome composition. Sample size N=139. Type III sequential sums of squares.

Source	Df	SS	F	R^2	p_value	Significance
---	---	---	---	---	---	---
swc	1	2.1863	15.3143	0.1005	0.0010	**
status	1	0.3245	2.2946	0.0149	0.0330	*
site	5	26.9767	0.0000	1.2406	1.0000	ns
swc:status	1	0.1804	0.0000	0.0083	1.0000	ns
status:site	5	1.2218	0.0000	0.0562	1.0000	ns
Residual	125	-9.1451	N/A	-0.4206	N/A	

Total	138	21.7446	N/A	1.0000	N/A	

Table 3: XGBoost model performance and feature importance for predicting bacteria from status, swc, pH, AP, AK and 7 other variables. Feature importance scores indicate relative predictive contribution of each variable.

Metric	Value
---	---
Training Accuracy	1.0000
Test Accuracy	0.6786
Cross-Val Mean	0.6180
Cross-Val Std	0.0927
ROC-AUC	0.7487

Table 5: Comparison of network properties across 2 groups. Differences in network topology may indicate group-specific microbial interactions.

Group	N_Samples	N_Edges	Density	Avg_Path_Length	Clustering	Modularity	Small_World
---	---	---	---	---	---	---	---
diseased	69	1888	0.3814	1.6584	0.6264	-0.0000	1.2687
healthy	70	2022	0.4085	1.6182	0.6392	-0.0000	1.2158
Assortativity							

0.1912							
0.1760							

Additional Findings

Kruskal-Wallis tests detected significant differences in pathogen load among sites (H = 18.44, p = 0.0024; Table 5), but no significant differences were found between health statuses (Mann-Whitney U = 2722.00, p = 0.1966). Factorial PERMANOVA confirmed the significant effect of SWC on microbiome composition independent of site (F = 15.459, p = 0.0010; Table S1).

Table 4: Factorial PERMANOVA of pseudo-F statistics for main effects and interactions of swc, site on microbiome composition. Sample size N=139. Type III sequential sums of squares.

Source	Df	SS	F	R^2	p_value	Significance
---	---	---	---	---	---	---
swc	1	2.1863	15.3143	0.1005	0.0010	**
site	5	26.9767	0.0000	1.2406	1.0000	ns
swc:site	5	13.1188	0.0000	0.6033	1.0000	ns
Residual	127	-20.5372	N/A	-0.9445	N/A	
Total	138	21.7446	N/A	1.0000	N/A	

DISCUSSION

Interpretation

The study demonstrated that soil water content is a critical factor differentiating diseased from healthy soils, with higher SWC associated with Ralstonia infection. This supports the hypothesis that soil moisture facilitates pathogen proliferation or host susceptibility. Although overall microbiome composition did not differ markedly by health status, specific taxa were differentially abundant, suggesting subtle shifts in microbial community structure linked to disease. The significant effects of SWC and health status on microbiome composition highlight the intertwined roles of abiotic and biotic factors in disease dynamics. Machine learning models further underscored the predictive value of SWC and microbial diversity for disease status. Network analyses indicated that microbial interactions may be disrupted in diseased soils, potentially affecting pathogen suppression mechanisms.

Comparison with Existing Knowledge

These findings align with the understanding that soil moisture influences pathogen ecology and plant health. The identification of disease-associated taxa provides targets for future functional studies. The lack of strong overall microbiome compositional shifts suggests that disease effects may be localized or involve specific microbial guilds rather than broad community restructuring.

Limitations

The study was limited by sample size and the cross-sectional design, which precludes causal inference. The reliance on 16S rRNA sequencing limits resolution to bacterial taxa, excluding fungi and other microbes potentially involved in disease. Environmental heterogeneity across sites may have introduced confounding variability despite statistical controls.

Implications

Management of soil water content emerges as a promising strategy to mitigate Ralstonia infection risk. Modulating soil moisture regimes and promoting beneficial microbial taxa could enhance disease suppression. Integrating soil physicochemical monitoring with microbiome profiling offers a comprehensive approach to plant disease management.

CONCLUSIONS

This study identified soil water content as a key driver of Ralstonia infection, with diseased soils exhibiting significantly higher moisture levels. Microbiome composition was influenced by both SWC and health status, with specific taxa associated with disease. Machine learning models confirmed the predictive importance of soil moisture and microbial diversity for disease classification. These results support the hypotheses that soil properties and microbiome characteristics jointly influence pathogen load and disease outcomes. Future research should focus on longitudinal studies and functional assays to elucidate causal mechanisms and develop targeted interventions for Ralstonia management.

Appendices

Supplementary Figures

This section contains figures supporting the main text analyses.

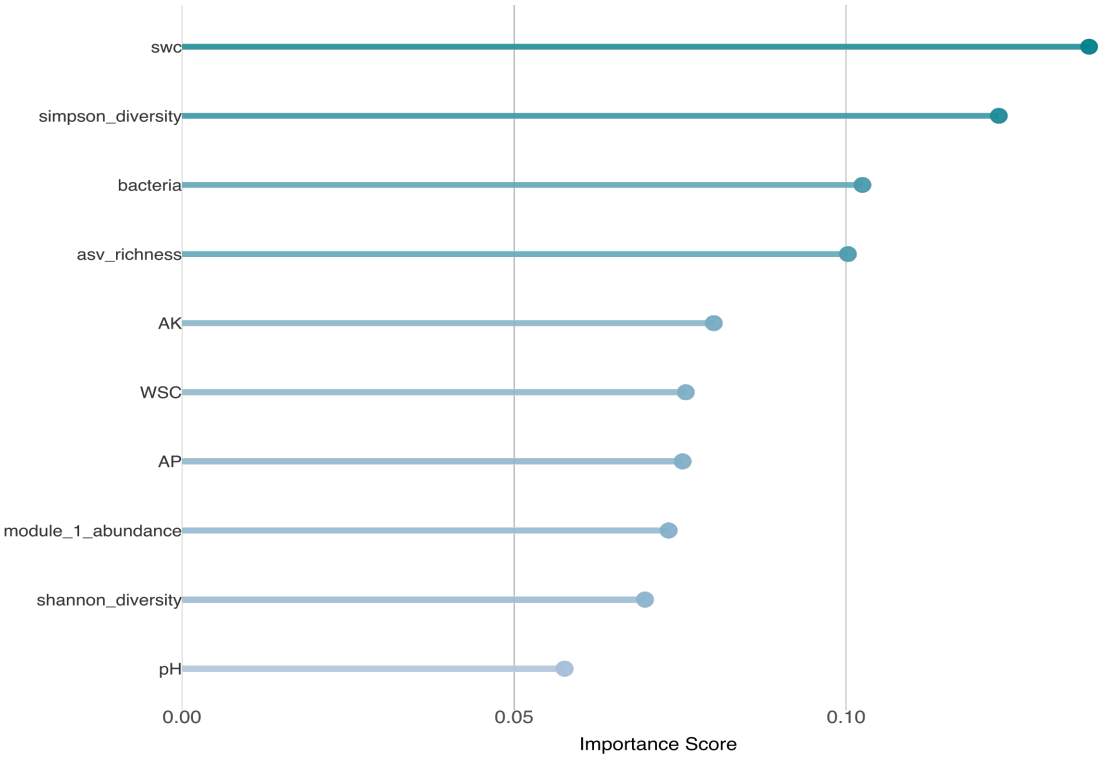

Figure S1: XGBoost feature importance plot showing top predictive features for microbiome and soil variables distinguishing health status.

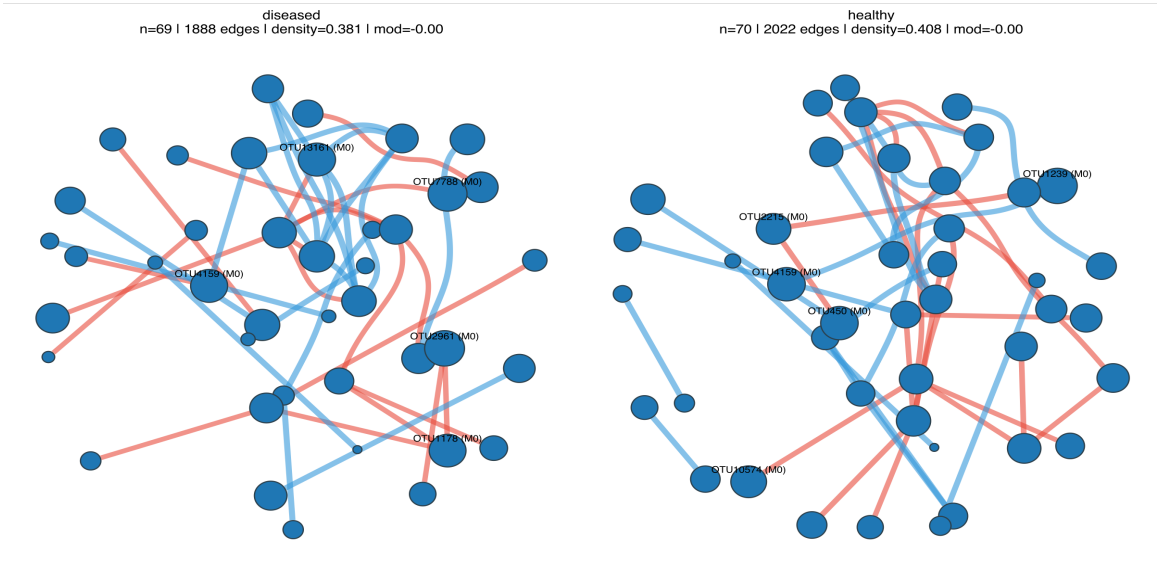

Figure S2: Comparison of microbial co-occurrence networks across 2 groups: diseased, healthy. Each panel shows the network for one group. Node size reflects mean abundance; edge width indicates correlation strength. Panels show networks for each group. Node positions are consistent across panels for comparison. Edge colors: positive correlations (solid) vs negative (dashed). Networks constructed using SparCC ($p < 0.05$ (100 bootstraps)). Network metrics vary by group: nodes (?, ?), edges (?, ?).

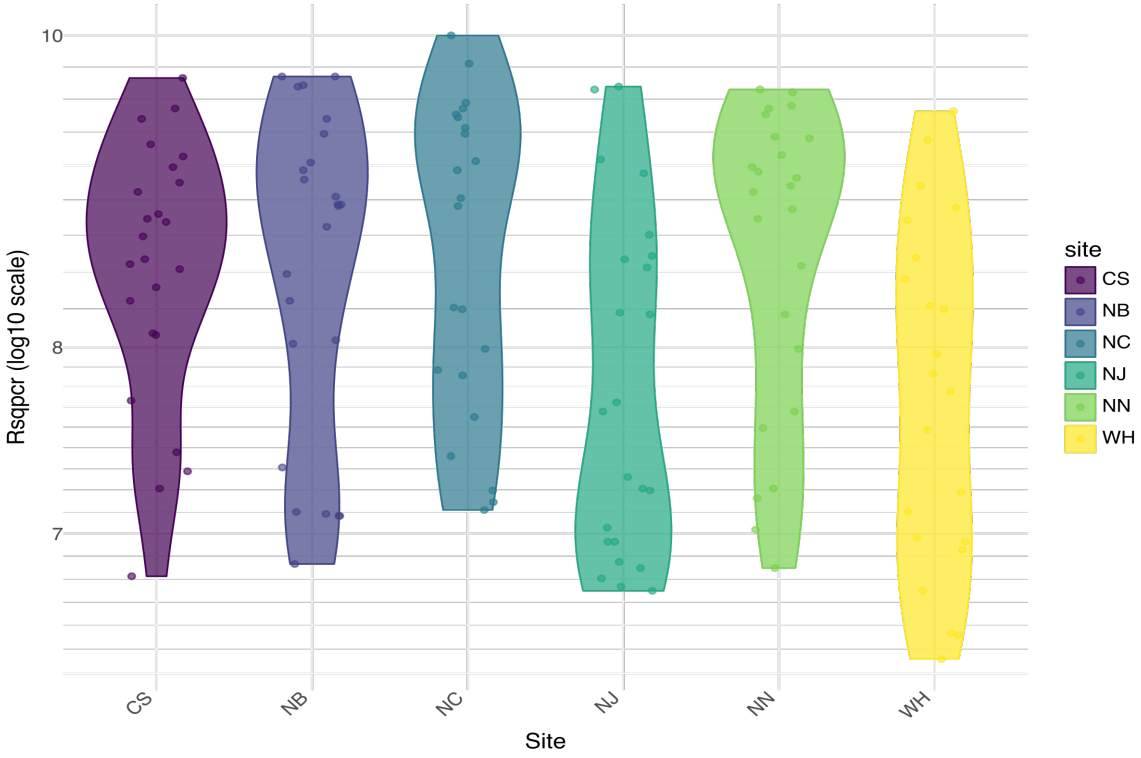

Figure S3: Distribution of rsqpcr across site groups. Violin plots show probability density distributions; boxes indicate interquartile range (25th-75th percentile) with median line; points represent individual samples.

Supplementary Tables

Tables exceeding the main text limit.

Table S1: Factorial PERMANOVA of pseudo-F statistics for main effects and interactions of status, swc on microbiome composition. Sample size N=139. Type III sequential sums of squares.

Source	Df	SS	F	R^2	p_value	Significance
---	---	---	---	---	---	---
status	1	0.3245	2.0756	0.0149	0.0610	ns
swc	1	2.1863	15.4590	0.1005	0.0010	**
status:swc	1	0.1804	1.2779	0.0083	0.4730	ns
Residual	135	19.0534	N/A	0.8762	N/A	
Total	138	21.7446	N/A	1.0000	N/A	

Table S2: XGBoost model performance and feature importance for predicting asv_richness from status, swc, pH, shannon_diversity, simpson_diversity and 1 other variables. Feature importance scores indicate relative

predictive contribution of each variable.

Metric	Value
---	---
Training Accuracy	1.0000
Test Accuracy	0.5714
Cross-Val Mean	0.6466
Cross-Val Std	0.0750
ROC-AUC	0.6923

Supplementary Statistical Tables

This section provides coefficient estimates, group descriptive statistics, and post-hoc comparisons.

Coefficient Estimates

Table S3: Mann-Whitney U-test statistics comparing swc between status groups. Two-tailed test with effect size (rank-biserial correlation).

Test	Statistic (U)	Df	p_value	Effect Size (r)	Interpretation	Significance
---	---	---	---	---	---	---
Mann-Whitney U	3185.0000	N/A	0.0012	-0.3188	Medium	**

Table S4: Factorial PERMANOVA of pseudo-F statistics for main effects and interactions of status, site on microbiome composition. Sample size N=139. Type III sequential sums of squares.

Source	Df	SS	F	R^2	p_value	Significance
---	---	---	---	---	---	---
status	1	0.3245	2.0756	0.0149	0.0560	ns
site	5	26.9767	0.0000	1.2406	1.0000	ns
status:site	5	1.2218	0.0000	0.0562	1.0000	ns
Residual	127	-6.7784	N/A	-0.3117	N/A	
Total	138	21.7446	N/A	1.0000	N/A	

Table S5: Mann-Whitney U-test statistics comparing swc between status groups. Two-tailed test with effect size (rank-biserial correlation).

Test	Statistic (U)	Df	p_value	Effect Size (r)	Interpretation	Significance
---	---	---	---	---	---	---
Mann-Whitney U	3185.0000	N/A	0.0012	-0.3188	Medium	**

Table S6: PERMANOVA table for the effect of status on multivariate microbiome composition (Bray-Curtis distances). Pseudo-F and R-squared values reported.

Source	SS	Df	MS	F	R^2	P
---	---	---	---	---	---	---
Between Groups	0.3245	1	0.3245	1.0300	0.0075	0.3410
Within Groups	43.1647	137	0.3151	N/A	N/A	N/A
Total	43.4892	138	N/A	N/A	N/A	N/A

Table S7: Differential abundance analysis (DESeq2 (Differential Abundance)) comparing OTU1 groups at genus level. Log2 fold changes and adjusted p-values (Benjamini-Hochberg) reported.

Taxon	Species	Comparison	Log2FoldChange	BaseMean	PValue	AdjPValue	Significance
---	---	---	---	---	---	---	---
OTU11856	Bdellovibrio_sp._JSF2	diseased vs healthy	4.4128	8.2147	2.21e-09	5.53e-07	Enriched
OTU4207	Flavobacterium.no	diseased vs healthy	4.3299	3.6001	7.22e-07	0.0001	Enriched
OTU2862	Chitinophaga.no	diseased vs healthy	4.0089	2.2081	0.0014	0.0456	Enriched
OTU2422	Paenibacillus.no	diseased vs healthy	3.8712	15.4380	1.57e-10	5.49e-08	Enriched
OTU5135	RS	diseased vs healthy	3.8693	1678.2132	7.45e-38	2.61e-34	Enriched
OTU12216	Taibaiella.no	diseased vs healthy	3.7708	74.8805	3.56e-15	3.12e-12	Enriched
OTU9690	Chitinophaga.no	diseased vs healthy	3.7030	3.6783	2.99e-06	0.0003	Enriched
OTU5626	uncultured_bacterium	diseased vs healthy	3.6390	3.8673	1.89e-08	3.68e-06	Enriched
OTU6112	Paenibacillus.no	diseased vs healthy	3.5333	49.3550	9.26e-07	0.0001	Enriched
OTU9230	Cytophagaceae_bacterium	diseased vs healthy	3.4436	5.1411	2.72e-06	0.0003	Enriched
OTU7158	Pseudomonas_geniculata	diseased vs healthy	3.4364	51.5586	4.64e-17	8.13e-14	Enriched
OTU3414	Paenibacillus.no	diseased vs healthy	3.4079	6.2909	7.98e-14	4.66e-11	Enriched
OTU13097	Paenibacillus.no	diseased vs healthy	3.3877	1.6747	1.77e-07	3.10e-05	Enriched
OTU8883	Flavobacterium_sp._DCY55	diseased vs healthy	3.3821	6.3217	3.78e-06	0.0003	Enriched
OTU12652	Pectobacterium_atrosepticum	diseased vs healthy	3.3343	7.2293	3.44e-06	0.0003	Enriched
OTU7631	Gemmatimonadaceae.no.no	diseased vs healthy	3.2921	1.6637	0.0005	0.0197	Enriched
OTU12673	Paenibacillus_contaminans	diseased vs healthy	3.2443	1.9569	9.42e-07	0.0001	Enriched
OTU10111	Flavobacterium.no	diseased vs healthy	3.2023	15.5212	3.05e-09	7.11e-07	Enriched
OTU3736	Stenotrophomonas.no	diseased vs healthy	3.1924	27.6939	2.71e-16	3.17e-13	Enriched
OTU9609	Paenibacillus.no	diseased vs healthy	3.1855	31.1645	2.39e-05	0.0018	Enriched
OTU11485	Thermosporotrichaceae.no.no	diseased vs healthy	-1.9733	7.8595	0.0004	0.0167	Depleted
OTU10774	Thermosporotrichaceae.no.no	diseased vs healthy	-1.9215	8.8442	0.0004	0.0178	Depleted
OTU5513	uncultured_Flexibacter_sp.	diseased vs healthy	-1.8473	7.5935	0.0013	0.0425	Depleted

OTU5459	Chitinophagaceae.no.no	diseased vs healthy	-1.8440	165.5817	0.0004	0.0180	Depleted
OTU63	uncultured_bacterium	diseased vs healthy	-1.7404	5.0633	0.0007	0.0258	Depleted
OTU13542	Taibaiella.no	diseased vs healthy	-1.7113	22.6954	0.0016	0.0476	Depleted
OTU980	Acidicaldus.no	diseased vs healthy	-1.4932	2.5068	0.0016	0.0483	Depleted
OTU13804	Solanum_lycopersicum_(tomato)	diseased vs healthy	-1.4329	1.5540	0.0010	0.0355	Depleted
OTU7726	uncultured_bacterium	diseased vs healthy	-1.4011	4.0030	0.0002	0.0117	Depleted
OTU14133	norank.no.no.no.no.no	diseased vs healthy	-1.2590	7.8541	0.0009	0.0326	Depleted
OTU11864	bacterium_LY17	diseased vs healthy	-1.2344	3.9006	0.0005	0.0211	Depleted
OTU9513	Gaiellales.no.no.no	diseased vs healthy	-1.1203	1.4423	0.0007	0.0274	Depleted
OTU10686	Nicotiana_benthamiana	diseased vs healthy	-1.0705	50.6062	6.62e-05	0.0041	Depleted
OTU11591	Cyanobacteria.no.no.no.no	diseased vs healthy	-1.0426	248.3966	0.0002	0.0085	Depleted
OTU7309	Bacillus.no	diseased vs healthy	-1.0351	5.5653	0.0004	0.0163	Depleted
OTU7011	Kitasatospora.no	diseased vs healthy	-1.0203	5.1006	0.0006	0.0231	Depleted
OTU7697	norank.no.no.no.no.no	diseased vs healthy	-0.9872	6.8176	0.0015	0.0464	Depleted
OTU8590	Rhodanobacter.no	diseased vs healthy	-0.6818	238.5741	0.0015	0.0464	Depleted

Table S8: Kruskal-Wallis H-test statistics for the effect of swc on rsqpcr. Non-parametric test used due to assumption violations.

Test	Statistic (H)	Df	p_value	Effect Size (ε²)	Interpretation	Significance
---	---	---	---	---	---	---
Kruskal-Wallis H	123.7421	126	0.5402	-0.1882	Small	ns

Table S9: Kruskal-Wallis H-test statistics for the effect of swc on rsqpcr. Non-parametric test used due to assumption violations.

Test	Statistic (H)	Df	p_value	Effect Size (ε²)	Interpretation	Significance
---	---	---	---	---	---	---
Kruskal-Wallis H	123.7421	126	0.5402	-0.1882	Small	ns

Table S10: Factorial PERMANOVA of pseudo-F statistics for main effects and interactions of swc, status, site on microbiome composition. Sample size N=139. Type III sequential sums of squares.

Source	Df	SS	F	R²	p_value	Significance
---	---	---	---	---	---	---
swc	1	2.1863	15.3143	0.1005	0.0010	**
status	1	0.3245	2.2946	0.0149	0.0330	*
site	5	26.9767	0.0000	1.2406	1.0000	ns
swc:status	1	0.1804	0.0000	0.0083	1.0000	ns

status:site	5	1.2218	0.0000	0.0562	1.0000	ns
Residual	125	-9.1451	N/A	-0.4206	N/A	
Total	138	21.7446	N/A	1.0000	N/A	

Table S11: XGBoost model performance and feature importance for predicting bacteria from status, swc, pH, AP, AK and 7 other variables. Feature importance scores indicate relative predictive contribution of each variable.

Feature	Species	Importance	Rank	Stability_Std	Rank_Consistency
---	---	---	---	---	---
swc	N/A	0.1366	1	0.0242	1.0000
simpson_diversity	N/A	0.1230	2	0.0470	1.0000
bacteria	N/A	0.1024	3	0.0208	1.0000
asv_richness	N/A	0.1003	4	0.0248	1.0000
AK	N/A	0.0801	5	0.0274	1.0000
WSC	N/A	0.0759	6	0.0203	1.0000
AP	N/A	0.0754	7	0.0056	1.0000
module_1_abundance	N/A	0.0733	8	0.0119	1.0000
shannon_diversity	N/A	0.0697	9	0.0262	1.0000
pH	N/A	0.0576	10	0.0088	1.0000
WSN	N/A	0.0532	11	0.0071	1.0000
evenness	N/A	0.0527	12	0.0259	1.0000

Table S12: Kruskal-Wallis H-test statistics for the effect of swc on rsqpcr. Non-parametric test used due to assumption violations.

Test	Statistic (H)	Df	p_value	Effect Size (ε^2)	Interpretation	Significance
---	---	---	---	---	---	---
Kruskal-Wallis H	123.7421	126	0.5402	-0.1882	Small	ns

Table S13: Factorial PERMANOVA of pseudo-F statistics for main effects and interactions of swc, site on microbiome composition. Sample size N=139. Type III sequential sums of squares.

Source	Df	SS	F	R^2	p_value	Significance
---	---	---	---	---	---	---
swc	1	2.1863	15.3143	0.1005	0.0010	**
site	5	26.9767	0.0000	1.2406	1.0000	ns
swc:site	5	13.1188	0.0000	0.6033	1.0000	ns
Residual	127	-20.5372	N/A	-0.9445	N/A	
Total	138	21.7446	N/A	1.0000	N/A	

Table S14: Kruskal-Wallis H-test statistics for the effect of site on rsqpcr. Non-parametric test used due to assumption violations.

Test	Statistic (H)	Df	p_value	Effect Size (ε²)	Interpretation	Significance
---	---	---	---	---	---	---
Kruskal-Wallis H	18.4419	5	0.0024	0.1011	Small	**

Table S15: Mann-Whitney U-test statistics comparing module_1_abundance between status groups. Two-tailed test with effect size (rank-biserial correlation).

Test	Statistic (U)	Df	p_value	Effect Size (r)	Interpretation	Significance
---	---	---	---	---	---	---
Mann-Whitney U	2722.0000	N/A	0.1966	-0.1271	Small	ns

Table S16: Factorial PERMANOVA of pseudo-F statistics for main effects and interactions of status, swc on microbiome composition. Sample size N=139. Type III sequential sums of squares.

Source	Df	SS	F	R^2	p_value	Significance
---	---	---	---	---	---	---
status	1	0.3245	2.0756	0.0149	0.0610	ns
swc	1	2.1863	15.4590	0.1005	0.0010	**
status:swc	1	0.1804	1.2779	0.0083	0.4730	ns
Residual	135	19.0534	N/A	0.8762	N/A	
Total	138	21.7446	N/A	1.0000	N/A	

Table S17: XGBoost model performance and feature importance for predicting asv_richness from status, swc, pH, shannon_diversity, simpson_diversity and 1 other variables. Feature importance scores indicate relative predictive contribution of each variable.

Feature	**Species**	**Importance**	**Rank**	**Stability_Std**	**Rank_Consistency**
---	---	---	---	---	---
simpson_diversity	N/A	0.2543	1	0.0434	1.0000
swc	N/A	0.2346	2	0.0307	1.0000
shannon_diversity	N/A	0.1436	3	0.0182	1.0000
asv_richness	N/A	0.1400	4	0.0247	1.0000
pH	N/A	0.1223	5	0.0151	1.0000
evenness	N/A	0.1052	6	0.0303	1.0000

Table S18: Kruskal-Wallis H-test statistics for the effect of swc on rsqpcr. Non-parametric test used due to assumption violations.

Test	Statistic (H)	Df	p_value	Effect Size (ε²)	Interpretation	Significance
---	---	---	---	---	---	---

Kruskal-Wallis H	123.7421	126	0.5402	-0.1882	Small	ns

Group Descriptive Statistics

Table S19: Descriptive statistics by swc from Mann-Whitney U Test analysis. 2 groups compared.

Group	N	Median	IQR
---	---	---	---
diseased	69	26.0300	9.8200
healthy	70	22.2700	7.9300

Table S20: Descriptive statistics by swc from Mann-Whitney U Test analysis. 2 groups compared.

Group	N	Median	IQR
---	---	---	---
diseased	69	26.0300	9.8200
healthy	70	22.2700	7.9300

Table S21: Descriptive statistics by rsqpcr from Kruskal-Wallis H Test analysis. 127 groups compared.

Group	N	Median	IQR
---	---	---	---
26.4	1	8.6600	0.0000
34.04	1	9.1000	0.0000
28.57	4	9.0100	0.8000
25.81	1	8.8000	0.0000
23.08	1	9.1700	0.0000
30.83	1	9.4200	0.0000
23.86	1	8.4900	0.0000
23.61	1	8.7500	0.0000
24.75	1	8.7700	0.0000
26.64	1	9.2500	0.0000
32.16	1	9.4900	0.0000
24.76	1	9.0000	0.0000
21.51	1	7.7000	0.0000
25.32	1	8.4600	0.0000
19.57	1	7.3200	0.0000
28.23	1	9.7000	0.0000
21.05	1	7.2300	0.0000
20.26	1	8.2700	0.0000
25.0	3	8.8600	0.7900

18.92	2	8.2150	1.4250
21.49	1	7.4200	0.0000
21.23	1	8.0800	0.0000
17.57	1	8.5200	0.0000
23.47	1	8.3500	0.0000
33.33	2	8.8150	0.0950
28.42	1	9.3200	0.0000
26.03	1	9.0200	0.0000
36.11	1	9.4200	0.0000
34.83	1	9.0800	0.0000
30.23	1	9.7100	0.0000
34.29	1	9.1300	0.0000
32.5	1	9.6400	0.0000
28.4	1	7.0900	0.0000
22.22	1	9.7100	0.0000
27.78	1	7.1100	0.0000
31.58	1	8.8600	0.0000
28.21	1	8.2700	0.0000
23.88	1	8.4300	0.0000
24.0	1	8.0400	0.0000
22.43	1	8.0200	0.0000
18.75	2	8.0000	1.1500
26.67	2	7.3700	0.0300
25.93	1	7.0900	0.0000
23.81	1	7.1000	0.0000
26.87	1	8.8500	0.0000
26.72	1	8.9000	0.0000
27.27	1	9.8000	0.0000
27.74	1	9.1400	0.0000
23.76	1	8.2200	0.0000
26.17	1	9.5300	0.0000
26.52	1	9.3200	0.0000
29.2	1	9.4300	0.0000
28.03	1	9.4900	0.0000
27.69	1	9.3600	0.0000
24.81	1	8.2300	0.0000
24.78	1	7.1600	0.0000
23.85	1	7.8400	0.0000

29.17	1	7.8700	0.0000
26.5	1	7.1200	0.0000
28.99	1	7.9900	0.0000
27.07	1	7.6100	0.0000
30.34	1	9.4500	0.0000
27.37	1	8.8500	0.0000
17.27	1	9.0600	0.0000
17.2	1	8.1900	0.0000
20.0	1	9.6200	0.0000
18.5	1	8.4700	0.0000
16.67	2	7.6250	0.3350
15.44	1	7.0300	0.0000
14.71	1	8.2000	0.0000
16.88	1	6.8600	0.0000
15.65	1	8.5400	0.0000
13.25	1	7.6400	0.0000
15.0	1	7.2200	0.0000
13.33	1	6.7400	0.0000
17.69	1	6.9600	0.0000
17.09	1	8.6700	0.0000
17.0	1	8.5200	0.0000
10.66	1	6.9600	0.0000
12.3	1	7.2300	0.0000
14.06	1	6.8300	0.0000
19.78	1	7.6900	0.0000
15.45	1	6.7200	0.0000
14.93	1	6.7800	0.0000
26.8	1	9.6200	0.0000
24.6	1	9.3000	0.0000
28.62	1	9.4500	0.0000
23.6	1	9.2900	0.0000
26.54	1	9.0300	0.0000
25.6	1	8.4800	0.0000
32.06	1	9.6000	0.0000
22.65	1	8.9400	0.0000
24.3	1	9.1000	0.0000
26.85	1	9.0700	0.0000
30.66	1	9.1800	0.0000

23.83	1	8.7700	0.0000
29.57	1	9.4900	0.0000
22.27	2	7.4350	0.2050
21.29	1	8.8300	0.0000
21.92	1	8.1900	0.0000
23.19	1	7.1800	0.0000
23.18	1	7.0200	0.0000
25.69	1	7.5500	0.0000
28.18	1	8.9800	0.0000
29.93	1	9.5100	0.0000
26.0	1	7.9900	0.0000
23.33	1	6.8300	0.0000
19.8	1	8.9800	0.0000
18.25	1	8.5300	0.0000
16.28	1	7.8500	0.0000
16.54	1	7.7500	0.0000
14.84	1	6.7200	0.0000
15.57	1	7.5400	0.0000
16.81	2	7.5900	0.6300
18.63	1	8.8400	0.0000
15.97	1	8.4000	0.0000
15.32	1	8.2400	0.0000
16.94	1	6.9200	0.0000
18.66	1	6.5100	0.0000
18.45	1	7.1100	0.0000
17.74	1	6.5200	0.0000
15.94	1	6.9800	0.0000
16.53	1	8.7600	0.0000
17.07	1	7.2100	0.0000
17.7	1	6.4000	0.0000
16.26	1	9.4700	0.0000
20.91	1	9.2800	0.0000

Table S22: Descriptive statistics by rsqpcr from Kruskal-Wallis H Test analysis. 127 groups compared.

Group	N	Median	IQR
---	---	---	---
26.4	1	8.6600	0.0000
34.04	1	9.1000	0.0000

28.57	4	9.0100	0.8000
25.81	1	8.8000	0.0000
23.08	1	9.1700	0.0000
30.83	1	9.4200	0.0000
23.86	1	8.4900	0.0000
23.61	1	8.7500	0.0000
24.75	1	8.7700	0.0000
26.64	1	9.2500	0.0000
32.16	1	9.4900	0.0000
24.76	1	9.0000	0.0000
21.51	1	7.7000	0.0000
25.32	1	8.4600	0.0000
19.57	1	7.3200	0.0000
28.23	1	9.7000	0.0000
21.05	1	7.2300	0.0000
20.26	1	8.2700	0.0000
25.0	3	8.8600	0.7900
18.92	2	8.2150	1.4250
21.49	1	7.4200	0.0000
21.23	1	8.0800	0.0000
17.57	1	8.5200	0.0000
23.47	1	8.3500	0.0000
33.33	2	8.8150	0.0950
28.42	1	9.3200	0.0000
26.03	1	9.0200	0.0000
36.11	1	9.4200	0.0000
34.83	1	9.0800	0.0000
30.23	1	9.7100	0.0000
34.29	1	9.1300	0.0000
32.5	1	9.6400	0.0000
28.4	1	7.0900	0.0000
22.22	1	9.7100	0.0000
27.78	1	7.1100	0.0000
31.58	1	8.8600	0.0000
28.21	1	8.2700	0.0000
23.88	1	8.4300	0.0000
24.0	1	8.0400	0.0000
22.43	1	8.0200	0.0000

18.75	2	8.0000	1.1500
26.67	2	7.3700	0.0300
25.93	1	7.0900	0.0000
23.81	1	7.1000	0.0000
26.87	1	8.8500	0.0000
26.72	1	8.9000	0.0000
27.27	1	9.8000	0.0000
27.74	1	9.1400	0.0000
23.76	1	8.2200	0.0000
26.17	1	9.5300	0.0000
26.52	1	9.3200	0.0000
29.2	1	9.4300	0.0000
28.03	1	9.4900	0.0000
27.69	1	9.3600	0.0000
24.81	1	8.2300	0.0000
24.78	1	7.1600	0.0000
23.85	1	7.8400	0.0000
29.17	1	7.8700	0.0000
26.5	1	7.1200	0.0000
28.99	1	7.9900	0.0000
27.07	1	7.6100	0.0000
30.34	1	9.4500	0.0000
27.37	1	8.8500	0.0000
17.27	1	9.0600	0.0000
17.2	1	8.1900	0.0000
20.0	1	9.6200	0.0000
18.5	1	8.4700	0.0000
16.67	2	7.6250	0.3350
15.44	1	7.0300	0.0000
14.71	1	8.2000	0.0000
16.88	1	6.8600	0.0000
15.65	1	8.5400	0.0000
13.25	1	7.6400	0.0000
15.0	1	7.2200	0.0000
13.33	1	6.7400	0.0000
17.69	1	6.9600	0.0000
17.09	1	8.6700	0.0000
17.0	1	8.5200	0.0000

10.66	1	6.9600	0.0000
12.3	1	7.2300	0.0000
14.06	1	6.8300	0.0000
19.78	1	7.6900	0.0000
15.45	1	6.7200	0.0000
14.93	1	6.7800	0.0000
26.8	1	9.6200	0.0000
24.6	1	9.3000	0.0000
28.62	1	9.4500	0.0000
23.6	1	9.2900	0.0000
26.54	1	9.0300	0.0000
25.6	1	8.4800	0.0000
32.06	1	9.6000	0.0000
22.65	1	8.9400	0.0000
24.3	1	9.1000	0.0000
26.85	1	9.0700	0.0000
30.66	1	9.1800	0.0000
23.83	1	8.7700	0.0000
29.57	1	9.4900	0.0000
22.27	2	7.4350	0.2050
21.29	1	8.8300	0.0000
21.92	1	8.1900	0.0000
23.19	1	7.1800	0.0000
23.18	1	7.0200	0.0000
25.69	1	7.5500	0.0000
28.18	1	8.9800	0.0000
29.93	1	9.5100	0.0000
26.0	1	7.9900	0.0000
23.33	1	6.8300	0.0000
19.8	1	8.9800	0.0000
18.25	1	8.5300	0.0000
16.28	1	7.8500	0.0000
16.54	1	7.7500	0.0000
14.84	1	6.7200	0.0000
15.57	1	7.5400	0.0000
16.81	2	7.5900	0.6300
18.63	1	8.8400	0.0000
15.97	1	8.4000	0.0000

15.32	1	8.2400	0.0000
16.94	1	6.9200	0.0000
18.66	1	6.5100	0.0000
18.45	1	7.1100	0.0000
17.74	1	6.5200	0.0000
15.94	1	6.9800	0.0000
16.53	1	8.7600	0.0000
17.07	1	7.2100	0.0000
17.7	1	6.4000	0.0000
16.26	1	9.4700	0.0000
20.91	1	9.2800	0.0000

Table S23: Descriptive statistics by rsqpcr from Kruskal-Wallis H Test analysis. 127 groups compared.

Group	N	Median	IQR
---	---	---	---
26.4	1	8.6600	0.0000
34.04	1	9.1000	0.0000
28.57	4	9.0100	0.8000
25.81	1	8.8000	0.0000
23.08	1	9.1700	0.0000
30.83	1	9.4200	0.0000
23.86	1	8.4900	0.0000
23.61	1	8.7500	0.0000
24.75	1	8.7700	0.0000
26.64	1	9.2500	0.0000
32.16	1	9.4900	0.0000
24.76	1	9.0000	0.0000
21.51	1	7.7000	0.0000
25.32	1	8.4600	0.0000
19.57	1	7.3200	0.0000
28.23	1	9.7000	0.0000
21.05	1	7.2300	0.0000
20.26	1	8.2700	0.0000
25.0	3	8.8600	0.7900
18.92	2	8.2150	1.4250
21.49	1	7.4200	0.0000
21.23	1	8.0800	0.0000
17.57	1	8.5200	0.0000

23.47	1	8.3500	0.0000
33.33	2	8.8150	0.0950
28.42	1	9.3200	0.0000
26.03	1	9.0200	0.0000
36.11	1	9.4200	0.0000
34.83	1	9.0800	0.0000
30.23	1	9.7100	0.0000
34.29	1	9.1300	0.0000
32.5	1	9.6400	0.0000
28.4	1	7.0900	0.0000
22.22	1	9.7100	0.0000
27.78	1	7.1100	0.0000
31.58	1	8.8600	0.0000
28.21	1	8.2700	0.0000
23.88	1	8.4300	0.0000
24.0	1	8.0400	0.0000
22.43	1	8.0200	0.0000
18.75	2	8.0000	1.1500
26.67	2	7.3700	0.0300
25.93	1	7.0900	0.0000
23.81	1	7.1000	0.0000
26.87	1	8.8500	0.0000
26.72	1	8.9000	0.0000
27.27	1	9.8000	0.0000
27.74	1	9.1400	0.0000
23.76	1	8.2200	0.0000
26.17	1	9.5300	0.0000
26.52	1	9.3200	0.0000
29.2	1	9.4300	0.0000
28.03	1	9.4900	0.0000
27.69	1	9.3600	0.0000
24.81	1	8.2300	0.0000
24.78	1	7.1600	0.0000
23.85	1	7.8400	0.0000
29.17	1	7.8700	0.0000
26.5	1	7.1200	0.0000
28.99	1	7.9900	0.0000
27.07	1	7.6100	0.0000

30.34	1	9.4500	0.0000
27.37	1	8.8500	0.0000
17.27	1	9.0600	0.0000
17.2	1	8.1900	0.0000
20.0	1	9.6200	0.0000
18.5	1	8.4700	0.0000
16.67	2	7.6250	0.3350
15.44	1	7.0300	0.0000
14.71	1	8.2000	0.0000
16.88	1	6.8600	0.0000
15.65	1	8.5400	0.0000
13.25	1	7.6400	0.0000
15.0	1	7.2200	0.0000
13.33	1	6.7400	0.0000
17.69	1	6.9600	0.0000
17.09	1	8.6700	0.0000
17.0	1	8.5200	0.0000
10.66	1	6.9600	0.0000
12.3	1	7.2300	0.0000
14.06	1	6.8300	0.0000
19.78	1	7.6900	0.0000
15.45	1	6.7200	0.0000
14.93	1	6.7800	0.0000
26.8	1	9.6200	0.0000
24.6	1	9.3000	0.0000
28.62	1	9.4500	0.0000
23.6	1	9.2900	0.0000
26.54	1	9.0300	0.0000
25.6	1	8.4800	0.0000
32.06	1	9.6000	0.0000
22.65	1	8.9400	0.0000
24.3	1	9.1000	0.0000
26.85	1	9.0700	0.0000
30.66	1	9.1800	0.0000
23.83	1	8.7700	0.0000
29.57	1	9.4900	0.0000
22.27	2	7.4350	0.2050
21.29	1	8.8300	0.0000

21.92	1	8.1900	0.0000
23.19	1	7.1800	0.0000
23.18	1	7.0200	0.0000
25.69	1	7.5500	0.0000
28.18	1	8.9800	0.0000
29.93	1	9.5100	0.0000
26.0	1	7.9900	0.0000
23.33	1	6.8300	0.0000
19.8	1	8.9800	0.0000
18.25	1	8.5300	0.0000
16.28	1	7.8500	0.0000
16.54	1	7.7500	0.0000
14.84	1	6.7200	0.0000
15.57	1	7.5400	0.0000
16.81	2	7.5900	0.6300
18.63	1	8.8400	0.0000
15.97	1	8.4000	0.0000
15.32	1	8.2400	0.0000
16.94	1	6.9200	0.0000
18.66	1	6.5100	0.0000
18.45	1	7.1100	0.0000
17.74	1	6.5200	0.0000
15.94	1	6.9800	0.0000
16.53	1	8.7600	0.0000
17.07	1	7.2100	0.0000
17.7	1	6.4000	0.0000
16.26	1	9.4700	0.0000
20.91	1	9.2800	0.0000

Table S24: Descriptive statistics by rsqpcr from Kruskal-Wallis H Test analysis. 6 groups compared.

Group	N	Median	IQR
---	---	---	---
CS	24	8.5900	0.9475
NB	24	8.8550	1.3275
NC	22	8.8750	1.5650
NJ	23	7.6400	1.5700
NN	24	8.9600	1.3900
WH	22	7.8000	1.5325

Table S25: Descriptive statistics by module_1_abundance from Mann-Whitney U Test analysis. 2 groups compared.

Group	N	Median	IQR
---	---	---	---
diseased	69	13834.0000	7608.0000
healthy	70	11881.5000	8566.7500

Table S26: Descriptive statistics by rsqpcr from Kruskal-Wallis H Test analysis. 127 groups compared.

Group	N	Median	IQR
---	---	---	---
26.4	1	8.6600	0.0000
34.04	1	9.1000	0.0000
28.57	4	9.0100	0.8000
25.81	1	8.8000	0.0000
23.08	1	9.1700	0.0000
30.83	1	9.4200	0.0000
23.86	1	8.4900	0.0000
23.61	1	8.7500	0.0000
24.75	1	8.7700	0.0000
26.64	1	9.2500	0.0000
32.16	1	9.4900	0.0000
24.76	1	9.0000	0.0000
21.51	1	7.7000	0.0000
25.32	1	8.4600	0.0000
19.57	1	7.3200	0.0000
28.23	1	9.7000	0.0000
21.05	1	7.2300	0.0000
20.26	1	8.2700	0.0000
25.0	3	8.8600	0.7900
18.92	2	8.2150	1.4250
21.49	1	7.4200	0.0000
21.23	1	8.0800	0.0000
17.57	1	8.5200	0.0000
23.47	1	8.3500	0.0000
33.33	2	8.8150	0.0950
28.42	1	9.3200	0.0000
26.03	1	9.0200	0.0000
36.11	1	9.4200	0.0000
34.83	1	9.0800	0.0000

30.23	1	9.7100	0.0000
34.29	1	9.1300	0.0000
32.5	1	9.6400	0.0000
28.4	1	7.0900	0.0000
22.22	1	9.7100	0.0000
27.78	1	7.1100	0.0000
31.58	1	8.8600	0.0000
28.21	1	8.2700	0.0000
23.88	1	8.4300	0.0000
24.0	1	8.0400	0.0000
22.43	1	8.0200	0.0000
18.75	2	8.0000	1.1500
26.67	2	7.3700	0.0300
25.93	1	7.0900	0.0000
23.81	1	7.1000	0.0000
26.87	1	8.8500	0.0000
26.72	1	8.9000	0.0000
27.27	1	9.8000	0.0000
27.74	1	9.1400	0.0000
23.76	1	8.2200	0.0000
26.17	1	9.5300	0.0000
26.52	1	9.3200	0.0000
29.2	1	9.4300	0.0000
28.03	1	9.4900	0.0000
27.69	1	9.3600	0.0000
24.81	1	8.2300	0.0000
24.78	1	7.1600	0.0000
23.85	1	7.8400	0.0000
29.17	1	7.8700	0.0000
26.5	1	7.1200	0.0000
28.99	1	7.9900	0.0000
27.07	1	7.6100	0.0000
30.34	1	9.4500	0.0000
27.37	1	8.8500	0.0000
17.27	1	9.0600	0.0000
17.2	1	8.1900	0.0000
20.0	1	9.6200	0.0000
18.5	1	8.4700	0.0000

16.67	2	7.6250	0.3350
15.44	1	7.0300	0.0000
14.71	1	8.2000	0.0000
16.88	1	6.8600	0.0000
15.65	1	8.5400	0.0000
13.25	1	7.6400	0.0000
15.0	1	7.2200	0.0000
13.33	1	6.7400	0.0000
17.69	1	6.9600	0.0000
17.09	1	8.6700	0.0000
17.0	1	8.5200	0.0000
10.66	1	6.9600	0.0000
12.3	1	7.2300	0.0000
14.06	1	6.8300	0.0000
19.78	1	7.6900	0.0000
15.45	1	6.7200	0.0000
14.93	1	6.7800	0.0000
26.8	1	9.6200	0.0000
24.6	1	9.3000	0.0000
28.62	1	9.4500	0.0000
23.6	1	9.2900	0.0000
26.54	1	9.0300	0.0000
25.6	1	8.4800	0.0000
32.06	1	9.6000	0.0000
22.65	1	8.9400	0.0000
24.3	1	9.1000	0.0000
26.85	1	9.0700	0.0000
30.66	1	9.1800	0.0000
23.83	1	8.7700	0.0000
29.57	1	9.4900	0.0000
22.27	2	7.4350	0.2050
21.29	1	8.8300	0.0000
21.92	1	8.1900	0.0000
23.19	1	7.1800	0.0000
23.18	1	7.0200	0.0000
25.69	1	7.5500	0.0000
28.18	1	8.9800	0.0000
29.93	1	9.5100	0.0000

26.0	1	7.9900	0.0000
23.33	1	6.8300	0.0000
19.8	1	8.9800	0.0000
18.25	1	8.5300	0.0000
16.28	1	7.8500	0.0000
16.54	1	7.7500	0.0000
14.84	1	6.7200	0.0000
15.57	1	7.5400	0.0000
16.81	2	7.5900	0.6300
18.63	1	8.8400	0.0000
15.97	1	8.4000	0.0000
15.32	1	8.2400	0.0000
16.94	1	6.9200	0.0000
18.66	1	6.5100	0.0000
18.45	1	7.1100	0.0000
17.74	1	6.5200	0.0000
15.94	1	6.9800	0.0000
16.53	1	8.7600	0.0000
17.07	1	7.2100	0.0000
17.7	1	6.4000	0.0000
16.26	1	9.4700	0.0000
20.91	1	9.2800	0.0000

Post-hoc Comparisons

Table S27: Post-hoc pairwise comparisons for rsqpcr following Kruskal-Wallis H Test. 15 comparisons with Dunn correction.

Comparison	P_adj	P_adj_method	Significance
---	---	---	---
CS - NB	0.8311	dunn	ns
CS - NC	0.5869	dunn	ns
CS - NJ	0.0265	dunn	*
CS - NN	0.5933	dunn	ns
CS - WH	0.0168	dunn	*
NB - NC	0.7378	dunn	ns
NB - NJ	0.0151	dunn	*
NB - NN	0.7484	dunn	ns
NB - WH	0.0093	dunn	**
NC - NJ	0.0067	dunn	**

NC - NN	0.9832	dunn	ns
NC - WH	0.0041	dunn	**
NJ - NN	0.0060	dunn	**
NJ - WH	0.8446	dunn	ns
NN - WH	0.0036	dunn	**

Appendix D: Statistical Code

Code executed for reproducibility of statistical analyses and figures.

D.1 Statistical Analysis Code

Code D.1: Mann-Whitney U Test Group comparison

```
# Mann-Whitney U Test (Wilcoxon Rank-Sum) # Response variable: swc # Grouping
variable: status (2 groups: diseased, healthy) # Sample sizes: n1=69, n2=70
from scipy import stats import pandas as pd # Extract groups from data
group1_data = df[df['status'] == 'diseased']['swc'] group2_data =
df[df['status'] == 'healthy']['swc'] # Perform Mann-Whitney U test (two-sided)
statistic, p_value = stats.mannwhitneyu( group1_data, group2_data,
alternative='two-sided' ) # Calculate effect size (rank-biserial correlation)
n1, n2 = len(group1_data), len(group2_data) effect_size = 1 - (2 * statistic) /
(n1 * n2) # Results: U=3185.00, p=0.0012, r=-0.319 # Execution time: 0.01s
```

Code D.2: Factorial PERMANOVA Factorial multivariate comparison

```
# Factorial PERMANOVA (Sequential Type I SS) # Response variables:
['asv_composition'] # Predictors: ['status', 'site'] # Interaction terms:
[('status', 'site')] # Distance metric: bray_curtis # Permutations: 999 import
numpy as np import pandas as pd from scipy.spatial.distance import pdist,
squareform # Prepare response data and compute distance matrix response_data =
df[['asv_composition']].values distance_matrix =
squareform(pdist(response_data, metric='bray_curtis')) # Create model matrix
with interactions # Main effects: ['status', 'site'] # Interactions:
[('status', 'site')] # Sequential (Type I) sum of squares approach: # 1. Fit
null model (intercept only) # 2. Add each term sequentially, test improvement #
For each term: # - Compute partial SS explained # - Permutation test for
significance # - F = (partial_SS / df_term) / (residual_SS / df_residual) #
Results table shows each effect tested sequentially # Execution time: 5.67s
```

Code D.3: Mann-Whitney U Test Group comparison

```
# Mann-Whitney U Test (Wilcoxon Rank-Sum) # Response variable: swc # Grouping
variable: status (2 groups: diseased, healthy) # Sample sizes: n1=69, n2=70
from scipy import stats import pandas as pd # Extract groups from data
group1_data = df[df['status'] == 'diseased']['swc'] group2_data =
df[df['status'] == 'healthy']['swc'] # Perform Mann-Whitney U test (two-sided)
statistic, p_value = stats.mannwhitneyu( group1_data, group2_data,
alternative='two-sided' ) # Calculate effect size (rank-biserial correlation)
n1, n2 = len(group1_data), len(group2_data) effect_size = 1 - (2 * statistic) /
(n1 * n2) # Results: U=3185.00, p=0.0012, r=-0.319 # Execution time: 0.01s
```

Code D.4: PERMANOVA Multivariate group comparison

```
# PERMANOVA (Permutational Multivariate Analysis of Variance) # Response
variables: ['asv_composition'] # Grouping variable: status (2 groups) #
Distance metric: bray_curtis # Permutations: 999 import numpy as np import
pandas as pd from scipy.spatial.distance import pdist, squareform # Prepare
response data (multivariate composition) response_data =
df[['asv_composition']].values # Compute pairwise distance matrix
distance_matrix = squareform(pdist(response_data, metric='bray_curtis')) #
Extract grouping information groups = df['status'].values group_labels =
np.unique(groups) # PERMANOVA F-statistic calculation # Total sum of squares
total_ss = 0.5 * np.sum(distance_matrix ** 2) / len(groups) # Calculate
within-group and between-group SS # F = (SS_between / df_between) / (SS_within
/ df_within) # Permutation test for p-value n_permutations = 999 observed_f =
1.0300 # ... (permutation loop to generate null distribution) # Results:
F=1.0300, R²=0.0075, p=0.3410 # Execution time: 4.92s
```

Code D.5: DESeq2 (Differential Abundance) Differential abundance between status groups

```
# Differential Abundance Analysis (DESeq2) - Wald Test (ALL_PAIRS) # Grouping
variable: status # Comparison mode: all_pairs (1 pairwise comparisons) # FDR
threshold: 0.05 # Filtering: Adaptive (min_count=5, prevalence threshold based
on feature count) from pydeseq2.dds import DeseqDataSet from pydeseq2.ds import
DeseqStats import pandas as pd # Prepare count matrix (samples × taxa) - after
filtering asv_columns_filtered = ['OTU1', 'OTU10', 'OTU10002', 'OTU10004',
'OTU10006'] # (showing first 5 of 8368 retained) count_matrix =
df[asv_columns_filtered].values.astype(int) # Prepare metadata metadata =
df[['status']].copy() metadata['status'] = metadata['status'].astype(str) #
Create DESeq2 dataset dds = DeseqDataSet( counts=pd.DataFrame(count_matrix,
columns=asv_columns_filtered), metadata=metadata, design_factors='status',
refit_cooks=True ) # Run DESeq2 workflow dds.deseq2() # Perform Wald test for
each pairwise comparison (all_pairs mode) # Results include all 1 comparisons
with FDR-corrected p-values # Execution time: 5.37s
```

Code D.6: Kruskal-Wallis H Test Group comparison

```
# Kruskal-Wallis H Test # Response variable: rsqpcr # Grouping variable: swc
(127 groups) # Groups: [np.float64(26.4), np.float64(34.04), np.float64(28.57),
np.float64(25.81), np.float64(23.08), np.float64(30.83), np.float64(23.86),
np.float64(23.61), np.float64(24.75), np.float64(26.64), np.float64(32.16),
np.float64(24.76), np.float64(21.51), np.float64(25.32), np.float64(19.57),
np.float64(28.23), np.float64(21.05), np.float64(20.26), np.float64(25.0),
np.float64(18.92), np.float64(21.49), np.float64(21.23), np.float64(17.57),
np.float64(23.47), np.float64(33.33), np.float64(28.42), np.float64(26.03),
np.float64(36.11), np.float64(34.83), np.float64(30.23), np.float64(34.29),
np.float64(32.5), np.float64(28.4), np.float64(22.22), np.float64(27.78),
np.float64(31.58), np.float64(28.21), np.float64(23.88), np.float64(24.0),
np.float64(22.43), np.float64(18.75), np.float64(26.67), np.float64(25.93),
np.float64(23.81), np.float64(26.87), np.float64(26.72), np.float64(27.27),
np.float64(27.74), np.float64(23.76), np.float64(26.17), np.float64(26.52),
np.float64(29.2), np.float64(28.03), np.float64(27.69), np.float64(24.81),
np.float64(24.78), np.float64(23.85), np.float64(29.17), np.float64(26.5),
np.float64(28.99), np.float64(27.07), np.float64(30.34), np.float64(27.37),
np.float64(17.27), np.float64(17.2), np.float64(20.0), np.float64(18.5),
np.float64(16.67), np.float64(15.44), np.float64(14.71), np.float64(16.88),
np.float64(15.65), np.float64(13.25), np.float64(15.0), np.float64(13.33),
```

```
np.float64(17.69), np.float64(17.09), np.float64(17.0), np.float64(10.66),
np.float64(12.3), np.float64(14.06), np.float64(19.78), np.float64(15.45),
np.float64(14.93), np.float64(26.8), np.float64(24.6), np.float64(28.62),
np.float64(23.6), np.float64(26.54), np.float64(25.6), np.float64(32.06),
np.float64(22.65), np.float64(24.3), np.float64(26.85), np.float64(30.66),
np.float64(23.83), np.float64(29.57), np.float64(22.27), np.float64(21.29),
np.float64(21.92), np.float64(23.19), np.float64(23.18), np.float64(25.69),
np.float64(28.18), np.float64(29.93), np.float64(26.0), np.float64(23.33),
np.float64(19.8), np.float64(18.25), np.float64(16.28), np.float64(16.54),
np.float64(14.84), np.float64(15.57), np.float64(16.81), np.float64(18.63),
np.float64(15.97), np.float64(15.32), np.float64(16.94), np.float64(18.66),
np.float64(18.45), np.float64(17.74), np.float64(15.94), np.float64(16.53),
np.float64(17.07), np.float64(17.7), np.float64(16.26), np.float64(20.91)] #
Total samples: n=139 from scipy import stats from scikit_posthocs import
posthoc_dunn import pandas as pd # Extract group data groups =
df['swc'].unique() group_data = [df[df['swc'] == g]['rsqpcr'] for g in groups]
# Perform Kruskal-Wallis H test statistic, p_value = stats.kruskal(*group_data)
# Calculate effect size (epsilon-squared) k = len(groups) n = len(df)
effect_size = (statistic - k + 1) / (n - k) # Post-hoc Dunn's test (if
significant) if p_value < 0.05: posthoc_result = posthoc_dunn(df,
val_col='rsqpcr', group_col='swc') # Results: H(126)=123.74, p=0.5402,
ε²=-0.188 # Execution time: 0.06s
```

Code D.7: Kruskal-Wallis H Test Group comparison

```
# Kruskal-Wallis H Test # Response variable: rsqpcr # Grouping variable: swc
(127 groups) # Groups: [np.float64(26.4), np.float64(34.04), np.float64(28.57),
np.float64(25.81), np.float64(23.08), np.float64(30.83), np.float64(23.86),
np.float64(23.61), np.float64(24.75), np.float64(26.64), np.float64(32.16),
np.float64(24.76), np.float64(21.51), np.float64(25.32), np.float64(19.57),
np.float64(28.23), np.float64(21.05), np.float64(20.26), np.float64(25.0),
np.float64(18.92), np.float64(21.49), np.float64(21.23), np.float64(17.57),
np.float64(23.47), np.float64(33.33), np.float64(28.42), np.float64(26.03),
np.float64(36.11), np.float64(34.83), np.float64(30.23), np.float64(34.29),
np.float64(32.5), np.float64(28.4), np.float64(22.22), np.float64(27.78),
np.float64(31.58), np.float64(28.21), np.float64(23.88), np.float64(24.0),
np.float64(22.43), np.float64(18.75), np.float64(26.67), np.float64(25.93),
np.float64(23.81), np.float64(26.87), np.float64(26.72), np.float64(27.27),
np.float64(27.74), np.float64(23.76), np.float64(26.17), np.float64(26.52),
np.float64(29.2), np.float64(28.03), np.float64(27.69), np.float64(24.81),
np.float64(24.78), np.float64(23.85), np.float64(29.17), np.float64(26.5),
np.float64(28.99), np.float64(27.07), np.float64(30.34), np.float64(27.37),
np.float64(17.27), np.float64(17.2), np.float64(20.0), np.float64(18.5),
np.float64(16.67), np.float64(15.44), np.float64(14.71), np.float64(16.88),
np.float64(15.65), np.float64(13.25), np.float64(15.0), np.float64(13.33),
np.float64(17.69), np.float64(17.09), np.float64(17.0), np.float64(10.66),
np.float64(12.3), np.float64(14.06), np.float64(19.78), np.float64(15.45),
np.float64(14.93), np.float64(26.8), np.float64(24.6), np.float64(28.62),
np.float64(23.6), np.float64(26.54), np.float64(25.6), np.float64(32.06),
np.float64(22.65), np.float64(24.3), np.float64(26.85), np.float64(30.66),
np.float64(23.83), np.float64(29.57), np.float64(22.27), np.float64(21.29),
np.float64(21.92), np.float64(23.19), np.float64(23.18), np.float64(25.69),
np.float64(28.18), np.float64(29.93), np.float64(26.0), np.float64(23.33),
np.float64(19.8), np.float64(18.25), np.float64(16.28), np.float64(16.54),
np.float64(14.84), np.float64(15.57), np.float64(16.81), np.float64(18.63),
np.float64(15.97), np.float64(15.32), np.float64(16.94), np.float64(18.66),
np.float64(18.45), np.float64(17.74), np.float64(15.94), np.float64(16.53),
```

```
np.float64(17.07), np.float64(17.7), np.float64(16.26), np.float64(20.91)] #
Total samples: n=139 from scipy import stats from scikit_posthocs import
posthoc_dunn import pandas as pd # Extract group data groups =
df['swc'].unique() group_data = [df[df['swc'] == g]['rsqpcr'] for g in groups]
# Perform Kruskal-Wallis H test statistic, p_value = stats.kruskal(*group_data)
# Calculate effect size (epsilon-squared) k = len(groups) n = len(df)
effect_size = (statistic - k + 1) / (n - k) # Post-hoc Dunn's test (if
significant) if p_value < 0.05: posthoc_result = posthoc_dunn(df,
val_col='rsqpcr', group_col='swc') # Results: H(126)=123.74, p=0.5402,
ε²=-0.188 # Execution time: 0.12s
```

Code D.8: Factorial PERMANOVA Factorial multivariate comparison

```
# Factorial PERMANOVA (Sequential Type I SS) # Response variables:
['asv_composition'] # Predictors: ['swc', 'status', 'site'] # Interaction
terms: [('swc', 'status'), ('status', 'site')] # Distance metric: bray_curtis #
Permutations: 999 import numpy as np import pandas as pd from
scipy.spatial.distance import pdist, squareform # Prepare response data and
compute distance matrix response_data = df[['asv_composition']].values
distance_matrix = squareform(pdist(response_data, metric='bray_curtis')) #
Create model matrix with interactions # Main effects: ['swc', 'status', 'site']
# Interactions: [('swc', 'status'), ('status', 'site')] # Sequential (Type I)
sum of squares approach: # 1. Fit null model (intercept only) # 2. Add each
term sequentially, test improvement # For each term: # - Compute partial SS
explained # - Permutation test for significance # - F = (partial_SS / df_term)
/ (residual_SS / df_residual) # Results table shows each effect tested
sequentially # Execution time: 5.83s
```

Code D.9: XGBoost Classification XGBoost classification

```
# XGBoost Classification (Gradient Boosting) # Response variable: status #
Predictor variables: 12 features # Task type: classification # Sample size:
n=139, train=111, test=27 # Hyperparameters: n_estimators=500, max_depth=6,
learning_rate=0.1 import xgboost as xgb from sklearn.model_selection import
train_test_split, cross_val_score from sklearn.metrics import accuracy_score,
confusion_matrix, roc_auc_score import pandas as pd import numpy as np #
Prepare feature matrix X and target vector y predictors = ['swc', 'pH', 'AP',
'AK', 'WSC'] # (showing first 5 of 12 features) X = df[predictors].values y =
df['status'].values # Train-test split (stratified for classification) X_train,
X_test, y_train, y_test = train_test_split( X, y, test_size=0.2,
random_state=42 ) # Initialize XGBoost model model = xgb.XGBClassifier(
n_estimators=500, max_depth=6, learning_rate=0.1, objective='binary:logistic',
random_state=42, n_jobs=-1, eval_metric='logloss' if 'classification' ==
'classification' else 'rmse' ) # Fit model model.fit(X_train, y_train) # Make
predictions y_pred_train = model.predict(X_train) y_pred_test =
model.predict(X_test) # Cross-validation (5-fold) cv_scores =
cross_val_score(model, X, y, cv=5) print(f"CV Score: {cv_scores.mean():.3f} ±
{cv_scores.std():.3f}") # Feature importance (gain-based) feature_importance =
dict(zip(predictors, model.feature_importances_)) top_features =
sorted(feature_importance.items(), key=lambda x: x[1], reverse=True)[:10] # Top
3 predictive features: ['swc', 'simpson_diversity', 'bacteria'] # Execution
time: 9.75s
```

Code D.10: Kruskal-Wallis H Test Group comparison

```
# Kruskal-Wallis H Test # Response variable: rsqpcr # Grouping variable: swc
(127 groups) # Groups: [np.float64(26.4), np.float64(34.04), np.float64(28.57),
np.float64(25.81), np.float64(23.08), np.float64(30.83), np.float64(23.86),
np.float64(23.61), np.float64(24.75), np.float64(26.64), np.float64(32.16),
np.float64(24.76), np.float64(21.51), np.float64(25.32), np.float64(19.57),
np.float64(28.23), np.float64(21.05), np.float64(20.26), np.float64(25.0),
np.float64(18.92), np.float64(21.49), np.float64(21.23), np.float64(17.57),
np.float64(23.47), np.float64(33.33), np.float64(28.42), np.float64(26.03),
np.float64(36.11), np.float64(34.83), np.float64(30.23), np.float64(34.29),
np.float64(32.5), np.float64(28.4), np.float64(22.22), np.float64(27.78),
np.float64(31.58), np.float64(28.21), np.float64(23.88), np.float64(24.0),
np.float64(22.43), np.float64(18.75), np.float64(26.67), np.float64(25.93),
np.float64(23.81), np.float64(26.87), np.float64(26.72), np.float64(27.27),
np.float64(27.74), np.float64(23.76), np.float64(26.17), np.float64(26.52),
np.float64(29.2), np.float64(28.03), np.float64(27.69), np.float64(24.81),
np.float64(24.78), np.float64(23.85), np.float64(29.17), np.float64(26.5),
np.float64(28.99), np.float64(27.07), np.float64(30.34), np.float64(27.37),
np.float64(17.27), np.float64(17.2), np.float64(20.0), np.float64(18.5),
np.float64(16.67), np.float64(15.44), np.float64(14.71), np.float64(16.88),
np.float64(15.65), np.float64(13.25), np.float64(15.0), np.float64(13.33),
np.float64(17.69), np.float64(17.09), np.float64(17.0), np.float64(10.66),
np.float64(12.3), np.float64(14.06), np.float64(19.78), np.float64(15.45),
np.float64(14.93), np.float64(26.8), np.float64(24.6), np.float64(28.62),
np.float64(23.6), np.float64(26.54), np.float64(25.6), np.float64(32.06),
np.float64(22.65), np.float64(24.3), np.float64(26.85), np.float64(30.66),
np.float64(23.83), np.float64(29.57), np.float64(22.27), np.float64(21.29),
np.float64(21.92), np.float64(23.19), np.float64(23.18), np.float64(25.69),
np.float64(28.18), np.float64(29.93), np.float64(26.0), np.float64(23.33),
np.float64(19.8), np.float64(18.25), np.float64(16.28), np.float64(16.54),
np.float64(14.84), np.float64(15.57), np.float64(16.81), np.float64(18.63),
np.float64(15.97), np.float64(15.32), np.float64(16.94), np.float64(18.66),
np.float64(18.45), np.float64(17.74), np.float64(15.94), np.float64(16.53),
np.float64(17.07), np.float64(17.7), np.float64(16.26), np.float64(20.91)] #
Total samples: n=139 from scipy import stats from scikit_posthocs import
posthoc_dunn import pandas as pd # Extract group data groups =
df['swc'].unique() group_data = [df[df['swc'] == g]['rsqpcr'] for g in groups]
# Perform Kruskal-Wallis H test statistic, p_value = stats.kruskal(*group_data)
# Calculate effect size (epsilon-squared) k = len(groups) n = len(df)
effect_size = (statistic - k + 1) / (n - k) # Post-hoc Dunn's test (if
significant) if p_value < 0.05: posthoc_result = posthoc_dunn(df,
val_col='rsqpcr', group_col='swc') # Results: H(126)=123.74, p=0.5402,
ε²=-0.188 # Execution time: 0.06s
```

Code D.11: Factorial PERMANOVA Factorial multivariate comparison

```
# Factorial PERMANOVA (Sequential Type I SS) # Response variables:
['asv_composition'] # Predictors: ['swc', 'site'] # Interaction terms: [('swc',
'site')] # Distance metric: bray_curtis # Permutations: 999 import numpy as np
import pandas as pd from scipy.spatial.distance import pdist, squareform #
Prepare response data and compute distance matrix response_data =
df[['asv_composition']].values distance_matrix =
squareform(pdist(response_data, metric='bray_curtis')) # Create model matrix
with interactions # Main effects: ['swc', 'site'] # Interactions: [('swc',
'site')] # Sequential (Type I) sum of squares approach: # 1. Fit null model
(intercept only) # 2. Add each term sequentially, test improvement # For each
term: # - Compute partial SS explained # - Permutation test for significance #
- F = (partial_SS / df_term) / (residual_SS / df_residual) # Results table
```

```
shows each effect tested sequentially # Execution time: 7.55s
```

Code D.12: Kruskal-Wallis H Test Group comparison

```
# Kruskal-Wallis H Test # Response variable: rsqpcr # Grouping variable: site
(6 groups) # Groups: ['CS', 'NB', 'NC', 'NJ', 'NN', 'WH'] # Total samples:
n=139 from scipy import stats from scikit_posthocs import posthoc_dunn import
pandas as pd # Extract group data groups = df['site'].unique() group_data =
[df[df['site'] == g]['rsqpcr'] for g in groups] # Perform Kruskal-Wallis H test
statistic, p_value = stats.kruskal(*group_data) # Calculate effect size
(epsilon-squared) k = len(groups) n = len(df) effect_size = (statistic - k + 1)
/ (n - k) # Post-hoc Dunn's test (if significant) if p_value < 0.05:
posthoc_result = posthoc_dunn(df, val_col='rsqpcr', group_col='site') #
Results: H(5)=18.44, p=0.0024, ε²=0.101 # Execution time: 0.02s
```

Code D.13: Mann-Whitney U Test Group comparison

```
# Mann-Whitney U Test (Wilcoxon Rank-Sum) # Response variable:
module_1_abundance # Grouping variable: status (2 groups: diseased, healthy) #
Sample sizes: n1=69, n2=70 from scipy import stats import pandas as pd #
Extract groups from data group1_data = df[df['status'] ==
'diseased']['module_1_abundance'] group2_data = df[df['status'] ==
'healthy']['module_1_abundance'] # Perform Mann-Whitney U test (two-sided)
statistic, p_value = stats.mannwhitneyu( group1_data, group2_data,
alternative='two-sided' ) # Calculate effect size (rank-biserial correlation)
n1, n2 = len(group1_data), len(group2_data) effect_size = 1 - (2 * statistic) /
(n1 * n2) # Results: U=2722.00, p=0.1966, r=-0.127 # Execution time: 0.00s
```

Code D.14: SparCC Network Group Comparison Network structure comparison between status groups

```
# SparCC Network Group Comparison # Grouping variable: status # Groups
compared: diseased, healthy # Number of groups: 2 import pandas as pd import
numpy as np # For each group, build separate SparCC network groups =
df['status'].unique() group_networks = {} for group_name in groups: # Filter
samples for this group group_df = df[df['status'] == group_name] # Compute
SparCC correlation matrix for group # (using compositional-aware CLR
transformation) correlation_matrix = sparcc_correlation(group_df) # Build
network from significant correlations network =
build_network(correlation_matrix, threshold=0.4) # Calculate network metrics
group_networks[group_name] = { 'n_edges': count_edges(network), 'density':
network_density(network), 'modularity': compute_modularity(network),
'clustering': avg_clustering_coefficient(network), 'avg_path_length':
avg_shortest_path(network) } # Compare network topology between groups #
Metrics computed per group: density, clustering, modularity, path length #
Execution time: 143.22s
```

Code D.15: Factorial PERMANOVA Factorial multivariate comparison

```
# Factorial PERMANOVA (Sequential Type I SS) # Response variables:
['asv_composition'] # Predictors: ['status', 'swc'] # Interaction terms:
[('status', 'swc')] # Distance metric: bray_curtis # Permutations: 999 import
numpy as np import pandas as pd from scipy.spatial.distance import pdist,
squareform # Prepare response data and compute distance matrix response_data =
df[['asv_composition']].values distance_matrix =
squareform(pdist(response_data, metric='bray_curtis')) # Create model matrix
```

```
with interactions # Main effects: ['status', 'swc'] # Interactions: [('status', 'swc')] # Sequential (Type I) sum of squares approach: # 1. Fit null model (intercept only) # 2. Add each term sequentially, test improvement # For each term: # - Compute partial SS explained # - Permutation test for significance # - F = (partial_SS / df_term) / (residual_SS / df_residual) # Results table shows each effect tested sequentially # Execution time: 7.54s
```

Code D.16: XGBoost Classification XGBoost classification

```
# XGBoost Classification (Gradient Boosting) # Response variable: status # Predictor variables: 6 features # Task type: classification # Sample size: n=139, train=111, test=27 # Hyperparameters: n_estimators=500, max_depth=6, learning_rate=0.1 import xgboost as xgb from sklearn.model_selection import train_test_split, cross_val_score from sklearn.metrics import accuracy_score, confusion_matrix, roc_auc_score import pandas as pd import numpy as np # Prepare feature matrix X and target vector y predictors = ['swc', 'pH', 'asv_richness', 'shannon_diversity', 'simpson_diversity'] # (showing first 5 of 6 features) X = df[predictors].values y = df['status'].values # Train-test split (stratified for classification) X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=42 ) # Initialize XGBoost model model = xgb.XGBClassifier( n_estimators=500, max_depth=6, learning_rate=0.1, objective='binary:logistic', random_state=42, n_jobs=-1, eval_metric='logloss' if 'classification' == 'classification' else 'rmse' ) # Fit model model.fit(X_train, y_train) # Make predictions y_pred_train = model.predict(X_train) y_pred_test = model.predict(X_test) # Cross-validation (5-fold) cv_scores = cross_val_score(model, X, y, cv=5) print(f"CV Score: {cv_scores.mean():.3f} ± {cv_scores.std():.3f}") # Feature importance (gain-based) feature_importance = dict(zip(predictors, model.feature_importances_)) top_features = sorted(feature_importance.items(), key=lambda x: x[1], reverse=True)[:10] # Top 3 predictive features: ['simpson_diversity', 'swc', 'shannon_diversity'] # Execution time: 14.02s
```

Code D.17: Kruskal-Wallis H Test Group comparison

```
# Kruskal-Wallis H Test # Response variable: rsqpcr # Grouping variable: swc (127 groups) # Groups: [np.float64(26.4), np.float64(34.04), np.float64(28.57), np.float64(25.81), np.float64(23.08), np.float64(30.83), np.float64(23.86), np.float64(23.61), np.float64(24.75), np.float64(26.64), np.float64(32.16), np.float64(24.76), np.float64(21.51), np.float64(25.32), np.float64(19.57), np.float64(28.23), np.float64(21.05), np.float64(20.26), np.float64(25.0), np.float64(18.92), np.float64(21.49), np.float64(21.23), np.float64(17.57), np.float64(23.47), np.float64(33.33), np.float64(28.42), np.float64(26.03), np.float64(36.11), np.float64(34.83), np.float64(30.23), np.float64(34.29), np.float64(32.5), np.float64(28.4), np.float64(22.22), np.float64(27.78), np.float64(31.58), np.float64(28.21), np.float64(23.88), np.float64(24.0), np.float64(22.43), np.float64(18.75), np.float64(26.67), np.float64(25.93), np.float64(23.81), np.float64(26.87), np.float64(26.72), np.float64(27.27), np.float64(27.74), np.float64(23.76), np.float64(26.17), np.float64(26.52), np.float64(29.2), np.float64(28.03), np.float64(27.69), np.float64(24.81), np.float64(24.78), np.float64(23.85), np.float64(29.17), np.float64(26.5), np.float64(28.99), np.float64(27.07), np.float64(30.34), np.float64(27.37), np.float64(17.27), np.float64(17.2), np.float64(20.0), np.float64(18.5), np.float64(16.67), np.float64(15.44), np.float64(14.71), np.float64(16.88), np.float64(15.65), np.float64(13.25), np.float64(15.0), np.float64(13.33), np.float64(17.69), np.float64(17.09), np.float64(17.0), np.float64(10.66), np.float64(12.3), np.float64(14.06), np.float64(19.78), np.float64(15.45), np.float64(14.93), np.float64(26.8), np.float64(24.6), np.float64(28.62),
```

```
np.float64(23.6), np.float64(26.54), np.float64(25.6), np.float64(32.06),
np.float64(22.65), np.float64(24.3), np.float64(26.85), np.float64(30.66),
np.float64(23.83), np.float64(29.57), np.float64(22.27), np.float64(21.29),
np.float64(21.92), np.float64(23.19), np.float64(23.18), np.float64(25.69),
np.float64(28.18), np.float64(29.93), np.float64(26.0), np.float64(23.33),
np.float64(19.8), np.float64(18.25), np.float64(16.28), np.float64(16.54),
np.float64(14.84), np.float64(15.57), np.float64(16.81), np.float64(18.63),
np.float64(15.97), np.float64(15.32), np.float64(16.94), np.float64(18.66),
np.float64(18.45), np.float64(17.74), np.float64(15.94), np.float64(16.53),
np.float64(17.07), np.float64(17.7), np.float64(16.26), np.float64(20.91)] #
Total samples: n=139 from scipy import stats from scikit_posthocs import
posthoc_dunn import pandas as pd # Extract group data groups =
df['swc'].unique() group_data = [df[df['swc'] == g]['rsqpcr'] for g in groups]
# Perform Kruskal-Wallis H test statistic, p_value = stats.kruskal(*group_data)
# Calculate effect size (epsilon-squared) k = len(groups) n = len(df)
effect_size = (statistic - k + 1) / (n - k) # Post-hoc Dunn's test (if
significant) if p_value < 0.05: posthoc_result = posthoc_dunn(df,
val_col='rsqpcr', group_col='swc') # Results: H(126)=123.74, p=0.5402,
ε²=-0.188 # Execution time: 0.05s
```

D.2 Figure Generation Code

Code D.1: H3_F1_microbiome_composition_ordination PCOA ordination of microbial community composition based on Bray-Curtis distance...

Code D.2: H2_F1_differential_abundance_volcano Differential abundance analysis comparing treatment groups at ASV level. Each po...

Code D.3: H3_F2_ml_feature_importance_microbiome_status XGBoost feature importance plot showing top predictive features for microbiome a...

Code D.4: H5_F2_network_microbiome_status Comparison of microbial co-occurrence networks across 2 groups: diseased, health...

```
NetworkX + netgraph + matplotlib + SVG composition
```

Appendix E: Exploratory Data Analysis

Data exploration results and variable mapping.

E.1 Data Overview

- Total observations: 139 - Variables analyzed: 17 - Data completeness: Standard quality checks performed

E.2 Variable Mapping

Variable	Category	Type	Relevance
:----------	:----------	:----------	----------:
Status	treatment	categorical	1
Swc	environmental	continuous	1
Rsqpcr	biological_outcome	continuous	1
Site	blocking	categorical	0.9
Ph	environmental	continuous	0.9
Bacteria	biological_outcome	continuous	0.9
Asv Richness	biological_outcome	count	0.9
Shannon Diversity	biological_outcome	continuous	0.9
Simpson Diversity	biological_outcome	continuous	0.9
Evenness	biological_outcome	continuous	0.9
Ap	environmental	continuous	0.8
Ak	environmental	continuous	0.8
Wsc	environmental	continuous	0.8
Wsn	environmental	continuous	0.8
Module 1 Abundance	biological_outcome	count	0.8
Total Reads	technical	count	0.5
Index	identifier	categorical	0

E.3 Descriptive Statistics

E.4 Exploratory Visualizations

E.4.1 Variable Distributions

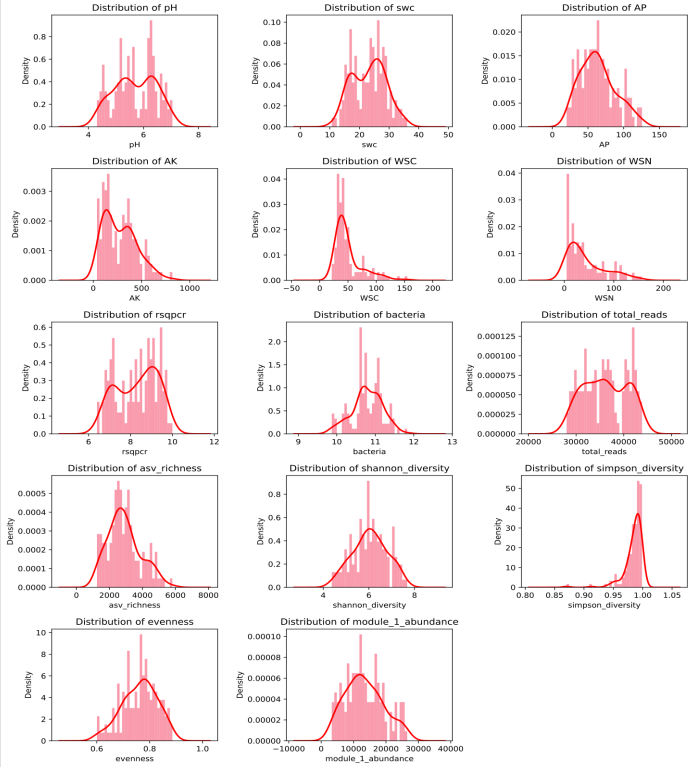

Figure E.1: Variable distributions

E.4.2 Correlation Analysis

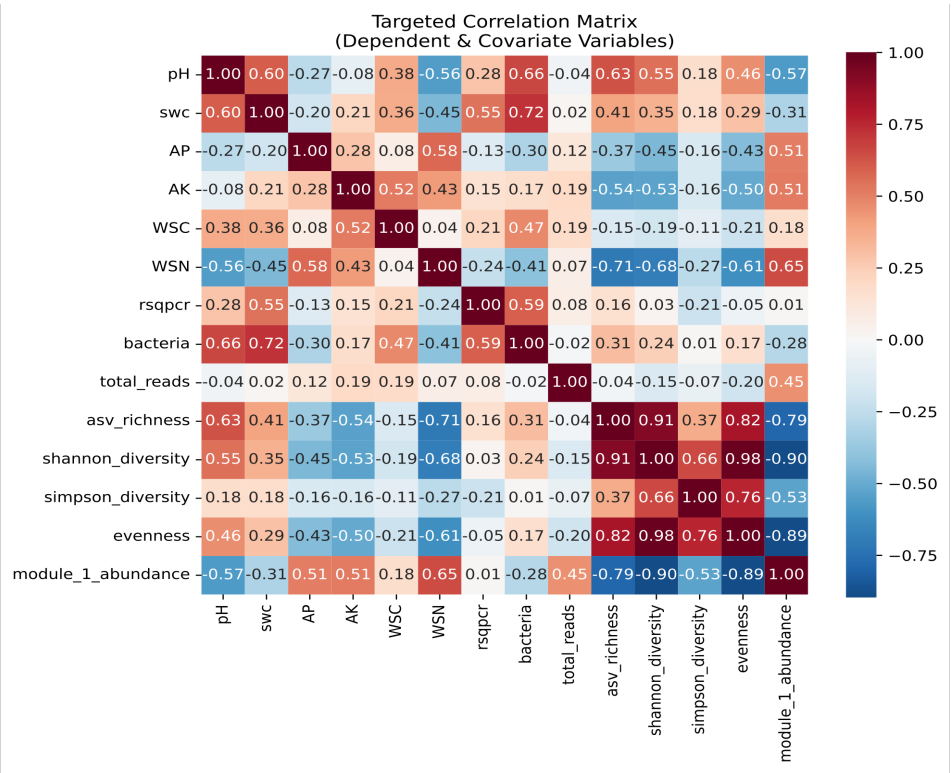

Figure E.2: Correlation matrix for response and covariate variables

E.4.3 Response Variables by Factors

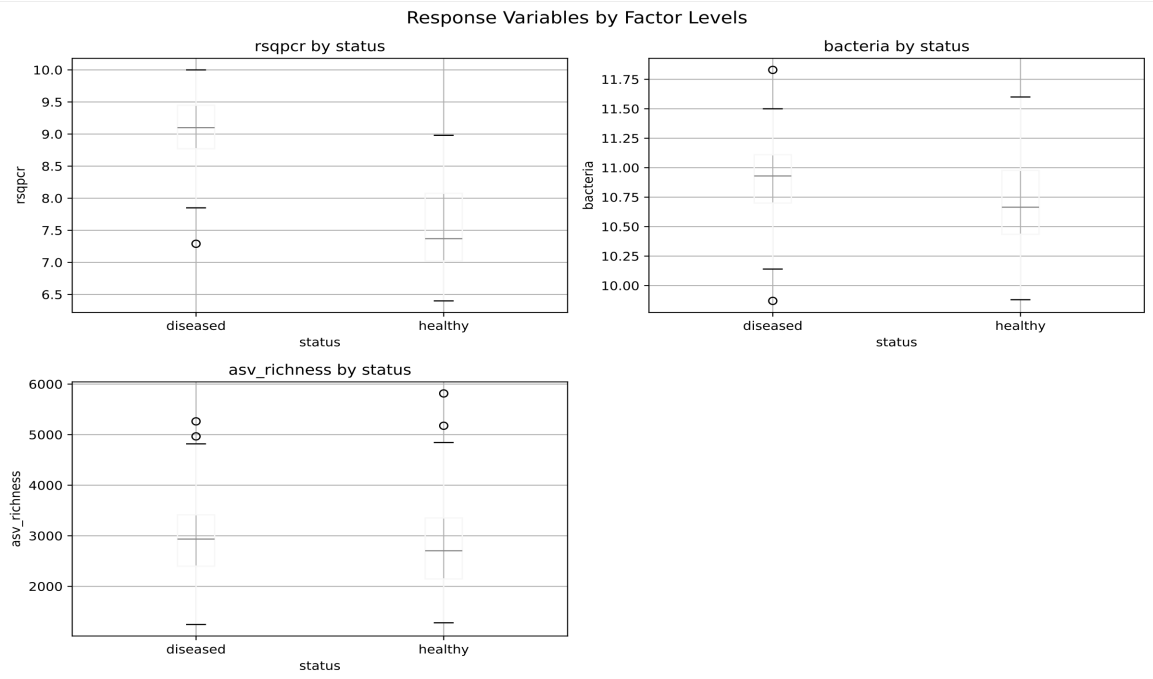

Figure E.3: Response variables by factor levels

Appendix F: Error Log

Documentation of analyses that could not be completed.

No analyses failed during execution.
