Scientific Research Publishing

American Journal of Plant Sciences, 2025, 16(4), 494-508
https://www.scirp.org/journal/ajps
ISSN Online: 2158-2750
ISSN Print: 2158-2742

Structural Analysis of Tobacco Rhizosphere Soil Microbial Communities Based on Metagenomics and Deep Learning for Association with Disease Resistance

Jinming Lu[1*], Haibo Xiang[1#], Rubing Xu[2#], Yanyan Li[2#], Yong Yang[1#]

[1]State Key Laboratory of Biocatalysis and Enzyme Engineering, School of Life Sciences, Hubei University, Wuhan, China
[2]Tobacco Research Institute of Hubei Province, Wuhan, China
Email: #xhb2086@hubu.edu.cn, #kubingbing@163.com, #yanyanli0025@126.com, #yangyong@hubu.edu.cn

How to cite this paper: Lu, J.M., Xiang, H.B., Xu, R.B., Li, Y.Y. and Yang, Y. (2025) Structural Analysis of Tobacco Rhizosphere Soil Microbial Communities Based on Metagenomics and Deep Learning for Association with Disease Resistance. *American Journal of Plant Sciences*, **16**, 494-508.
https://doi.org/10.4236/ajps.2025.164037

Received: January 18, 2025
Accepted: April 24, 2025
Published: April 27, 2025

Abstract

The rhizosphere microbiome, often termed the plant's "second genome", plays a pivotal role in regulating plant health and disease resistance. This study integrated metagenomic sequencing and deep learning to systematically compare the composition, diversity, and functional metabolism of microbial communities in healthy (NB) and diseased (NF) tobacco rhizosphere soils. Using BGISEQ-500 sequencing, 18 soil samples were analyzed, yielding 8.63 million clean reads and 6.27 million non-redundant genes. Taxonomic profiling revealed Proteobacteria (35%), Actinobacteria (16%), Firmicutes (10%), and Bacteroidetes (7%) as dominant phyla. Significant structural disparities in α-diversity indices (e.g., Shannon and Simpson) and β-diversity (PCoA) were observed between NB and NF groups (ANOVA, $p < 0.05$). LEfSe analysis identified 181 and 240 biomarkers in NB and NF, respectively, with healthy soils enriched in Gemmatimonadetes and Sphingomonadaceae, while diseased soils were dominated by Deltaproteobacteria and Rubrobacteraceae. Functional annotation highlighted the enrichment of sulfur metabolism (ko00920), terpenoid biosynthesis (ko00900), and antibiotic synthesis (ko01130) pathways in NB, whereas NF exhibited upregulated lipopolysaccharide biosynthesis (ko00540) and flagellar assembly (ko02040). A deep learning model (H2O framework, two hidden layers) achieved perfect classification (AUC = 1.0) and identified 10 microbial biomarkers as robust indicators of soil health. These findings elucidate the linkage between rhizosphere microbiome dynamics and tobacco disease resistance, providing a methodological framework for ecolog-

[*]First author.
[#]Corresponding author.

ical disease control and precision agriculture. The study highlights the potential of microbial biomarkers in guiding sustainable agricultural practices and reducing reliance on chemical pesticides.

Keywords

Metagenomics, Rhizosphere Microbiome, Tobacco Disease Resistance, Deep Learning, Community Diversity, Microbial Biomarkers

1. Introduction

Plant rhizosphere microbial communities, often referred to as the "second genome", play an indispensable role in regulating plant growth, nutrient uptake, and disease resistance [1]. As a critical interface for plant-soil interactions, rhizosphere microbes directly or indirectly suppress pathogen invasion through mechanisms such as antibiotic secretion, resource competition, and induced systemic resistance [2]. Tobacco (Nicotiana tabacum), a globally significant economic crop, relies heavily on the composition and functionality of its rhizosphere microbiome for plant health and yield. However, frequent outbreaks of soil-borne diseases caused by pathogens such as *Ralstonia solanacearum* (bacterial wilt) and *Phytophthora parasitica* (black shank) severely threaten the sustainable cultivation of tobacco. While conventional chemical control methods offer short-term disease management, they often disrupt soil microecological balance, enhance pathogen resistance, and contribute to environmental pollution [3]. Therefore, elucidating the mechanisms linking tobacco rhizosphere microbial communities to disease resistance and developing ecological control strategies based on microbiome modulation have become pivotal for advancing sustainable agriculture.

Recent breakthroughs in metagenomic sequencing have revolutionized the comprehensive analysis of soil microbial communities. Unlike traditional culture-dependent approaches, metagenomics enables unbiased capture of total DNA directly from environmental samples, including unculturable microorganisms, through high-throughput sequencing [4]. The workflow typically involves physical disruption and chemical lysis to release microbial DNA, followed by purification via magnetic bead adsorption or column chromatography to ensure high integrity and minimal host contamination. The DNA is then fragmented, ligated with sequencing adapters, and subjected to paired-end sequencing on platforms such as Illumina NovaSeq or PacBio, generating billions of base pairs. Raw data undergo quality control (e.g., removal of low-quality reads and host-derived sequences), followed by sequence assembly (using tools like MEGAHIT or MetaSPAdes) to reconstruct microbial genomic drafts. Taxonomic and functional annotations are performed using reference databases (e.g., NCBI NR, KEGG, COG) [5]. Metagenomics further deciphers the metabolic potential of microbial communities, such as antibiotic synthesis and sulfur metabolism pathways, providing molecular insights into microbe-hostenvironment interactions [6].

In plant microbiome research, metagenomics has been successfully applied to various crop systems. For instance, Bulgarelli *et al.* (2015) compared rhizosphere microbiomes of wild and domesticated barley, revealing host genotype-dependent regulation of microbial community structure. Mendes *et al.* (2011) identified disease-suppressive Pseudomonas species in sugarcane rhizospheres [7], demonstrating their role in inhibiting fungal pathogens via 2,4-diacetylphloroglucinol synthesis. However, studies on tobacco rhizosphere microbiomes remain limited. Existing research predominantly focuses on single-pathogen virulence mechanisms, with insufficient systemic exploration of host-microbe-pathogen tripartite interactions [8]. Additionally, the high dimensionality and noise inherent in metagenomic data challenge traditional statistical methods (e.g., PCA, LEfSe), necessitating advanced data-mining tools. The study indicates that continuous tobacco cropping leads to decreased similarity in rhizobacterial communities and significant alterations in fungal community structure (e.g., enrichment of Ascomycota and Fusarium), which may be associated with increased disease susceptibility [9]. The response of tobacco rhizosphere microbiome to continuous cropping obstacles exhibits specificity. For instance, after five consecutive years of flue-cured tobacco (Nicotiana tabacum) cultivation, bacterial community similarity markedly declined, while the abundance of pathogenic fungi (e.g., Fusarium spp.) in fungal communities significantly increased, demonstrating close correlation with root rot incidence. Through metagenomic non-targeted capture of functional genes, this research pioneers in revealing the synergistic interaction between sulfur metabolism pathway (ko00920) and antibiotic biosynthesis (ko01130) in tobacco rhizosphere, addressing a critical knowledge gap in tobacco-specific metabolic network studies [10].

Deep learning, a core artificial intelligence technology, offers a novel paradigm for metagenomic data analysis due to its robust nonlinear modeling and feature extraction capabilities. For example, convolutional neural networks (CNNs) capture spatial correlations in microbial abundance matrices and have been applied to predict associations between gut microbiota and colorectal cancer [11]. Recurrent neural networks (RNNs), adept at processing temporal data, show promise in modeling microbial community dynamics [12]. Nevertheless, deep learning applications in plant rhizosphere microbiome research, particularly for biomarker screening and classification models in tobacco disease resistance, remain underexplored.

This study integrates metagenomic sequencing and deep learning to analyze microbial community composition, diversity, and functional metabolism in healthy versus diseased tobacco rhizosphere soils. Key objectives include identifying ecologically significant taxa (e.g., Streptomyces, Pseudomonas), annotating disease resistance-related metabolic pathways (e.g., antibiotic synthesis, sulfur metabolism) via KEGG and COG databases, and constructing a deep learning classification model based on microbial abundance features to screen high-discriminatory biomarkers. The proposed "metagenomics-deep learning" framework aims to establish a rapid diagnostic system for soil health assessment. By elucidating molecular linkages between microbial communities and disease resistance, this research provides theoretical foundations for ecological disease control in tobacco cultivation. Methodologically, it offers a scalable technical framework for agricultural microbiome studies, advancing precision agriculture and promoting sustainable practices such as "reducing pesticide use while enhancing efficiency". These outcomes hold significant scientific and practical implications for green agricultural development.

2. Materials and Methods

2.1. Study Site and Sampling

The study was conducted in tobacco fields located in Lichuan City, Hubei Province (latitude N, longitude W), where tobacco had been cultivated continuously for over five years. The research focused on root-associated microbial communities and their association with disease resistance using deep learning models. Healthy (labeled as JK) and diseased (labeled as FB) tobacco plants, along with their corresponding rhizosphere soil samples (labeled as NB for healthy and NF for diseased), were selected. A total of 18 soil samples (nine healthy and nine diseased) were collected using a five-point mixed sampling method. The sampling area is characterized by an average annual temperature of 16.7˚C and an average annual precipitation of 1,304 mm.

2.2. Soil DNA Extraction and Metagenomic Analysis

Soil DNA extraction and purification were performed following the protocol described by Han *et al.* (2010) [13]. Briefly, crude DNA was extracted using chloroform-isoamyl alcohol, precipitated with isopropanol, and purified using QIAquick Gel Extraction Kit buffers. DNA concentration and integrity were verified prior to metagenomic sequencing. High-throughput sequencing was conducted on the BGISEQ-500 platform (BGI, Shenzhen, China). Raw metagenomic data and metabolic pathway abundances were screened using linear discriminant analysis effect size (LEfSe), followed by enrichment analysis of microbial metabolic pathways [14].

2.3. Data Analysis

2.3.1. Data Preprocessing and Assembly

Raw sequencing data were subjected to quality control using Trimmomatic software (v3.3) to remove adapter sequences and low-quality reads (quality score <20, length <50 bp), yielding high-quality clean reads [15]. Cleaned reads were assembled into contigs using MEGAHIT (parameters: --min-contig-len 500 --preset meta-large), with sequences pooled across samples to enhance assembly completeness [16]. Assembly quality was assessed using METAQUAST, with key metrics (e.g., N50 length, longest contig) reported to ensure reliability.

2.3.2. Species Annotation and Diversity Analysis

Assembled sequences were aligned against the NCBI database (including bacterial, archaeal, fungal, viral, and protozoan genomes) using Kraken2, and species abundances were calculated using Bracken. Community composition was summarized at phylum, class, order, family, genus, and species levels, visualized via KronaTools v2.8. Alpha diversity indices (Shannon, Simpson) were computed using the VEGAN package in R, while beta diversity was assessed via principal coordinate analysis (PCoA) with 95% confidence intervals [17].

2.3.3. LEfSe Analysis

LEfSe analysis [18] was applied to identify differentially abundant taxa (LDA score >3.0, p <0.05). Kruskal-Wallis rank-sum tests were used to detect intergroup differences, followed by Wilcoxon rank-sum tests to validate taxonomic consistency. Linear discriminant analysis (LDA) in Prism9 quantified contribution values. A total of 181 and 240 biomarkers were identified in healthy (JK) and diseased (FB) groups, respectively, with Gemmatimonadetes and Deltaproteobacteria serving as core differential taxa (LDA > 4.0).

2.3.4. Functional Annotation and Enrichment Analysis

Non-redundant gene sets were functionally annotated against the KEGG database, and metabolic pathway distributions were visualized using iPath. Differential metabolic pathways between groups were identified via LEfSe and validated using Wilcoxon tests (p < 0.05).

2.4. Construction of Neural Network Model

A neural network model was developed using the H2O deep learning library, with input features derived from genus-level microbial abundance data. The model architecture included two hidden layers (128 and 64 neurons, Rectifier With Dropout activation function) and was trained for 50 epochs (70% training set, 30% test set). Key biomarkers were selected through variable importance analysis. Although the sample size (n = 18) aligns with the design of comparable studies (e.g., Bulgarelli *et al.*, 2015), future investigations should incorporate multi-regional independent datasets (e.g., flue-cured tobacco fields under different cropping systems) to enhance the generalizability of the conclusions.

3. Results

3.1. Overview of Rhizosphere Microorganisms

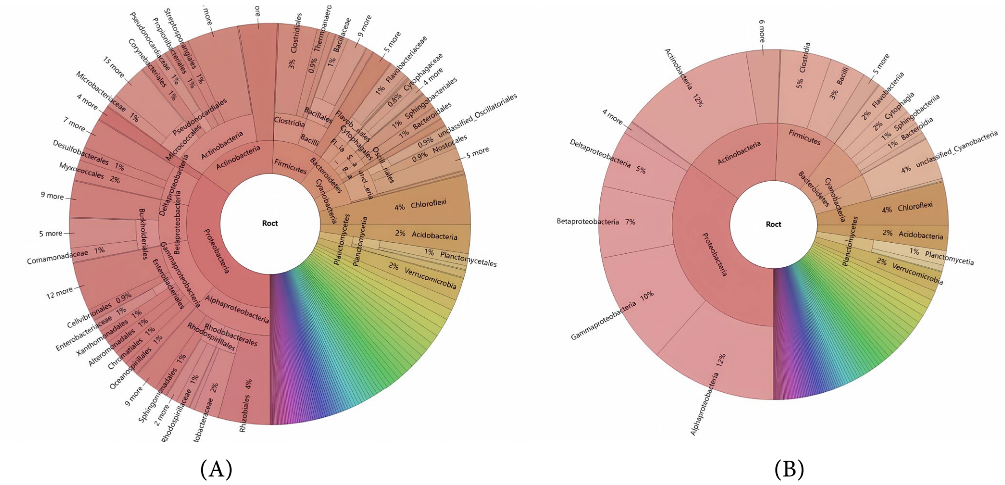
In this study, metagenomic sequencing of tobacco rhizosphere soil samples was conducted to elucidate the compositional structure of microbial communities. After sequencing and quality control, a total of 8630542.04 clean reads were retained, with an average of 7192118.4 clean reads per sample (**Table 1**). The assembly results revealed an average of 3203.72 contigs per sample, with an average N50 length of 1.78 kb. A total of 6.27 million non-redundant genes were identified, with an average open reading frame (ORF) length of 1,071 bp. These data indicate the presence of a highly diverse and abundant microbial ecosystem in the tobacco rhizosphere soil.

Hierarchical taxonomic analysis based on Krona plots demonstrated that at the class level (Figure 1(A)), the microbial community was dominated by Proteobacteria (35%), followed by Actinobacteria (16%), Firmicutes (10%), and Bacteroidetes (7%). At the order level (Figure 1(B)), Alphaproteobacteria (12%), Gammaproteobacteria (10%), and Actinobacteria (12%) exhibited significant abundance.

At the genus level (Figure 1(C), Figure 1(D)), Proteobacteria was primarily represented by Rhodobacteraceae, Rhodospirillaceae, Comamonadaceae, Hyphomicrobiaceae, Bradyrhizobiaceae, Enterobacteriaceae, and Chromatiaceae. Actinobacteria was predominantly composed of Microbacteriaceae, Intrasporangiaceae, Micrococcaceae, Pseudonocardiaceae, Nocardioidaceae, and Micromonosporaceae.

Table 1. Clean Data obtained through microbial metagenomes.

Sample	Raw Data	Clean Data (%)	ORFs
NB_1_5_1	86,756,202	84,901,410 (97.86)	1,170,682
NB_1_5_2	82,310,842	80,805,716 (98.17)	1,173,946
NB_1_5_3	88,903,850	87,454,354 (98.37)	1,296,511
NB_2_5_1	78,986,162	77,628,002 (98.28)	1,199,543
NB_2_5_2	83,031,322	81,674,588 (98.37)	1,209,022
NB_2_5_3	83,359,720	82,017,428 (98.39)	1,211,601
NB_3_5_1	80,533,420	78,899,252 (97.97)	1,135,035
NB_3_5_2	75,536,252	73,901,678 (97.84)	1,050,832
NB_3_5_3	82,318,954	80,438,206 (97.72)	1,165,549
NF_1_5_1	87,649,668	85,869,272 (97.97)	1,084,073
NF_1_5_2	79,054,488	77,138,586 (97.58)	857,643
NF_1_5_3	81,356,452	79,805,668 (98.09)	817,881
NF_2_5_1	88,665,820	86,541,770 (97.60)	993,219
NF_2_5_2	89,676,436	88,143,006 (98.29)	1,116,412
NF_2_5_3	92,028,234	89,853,422 (98.29)	1,148,853
NF_3_5_1	79,649,212	77,543,906 (97.36)	939,900
NF_3_5_2	81,720,172	80,424,486 (98.41)	954,141
NF_3_5_3	80,807,890	78,951,774 (97.70)	1,102,057



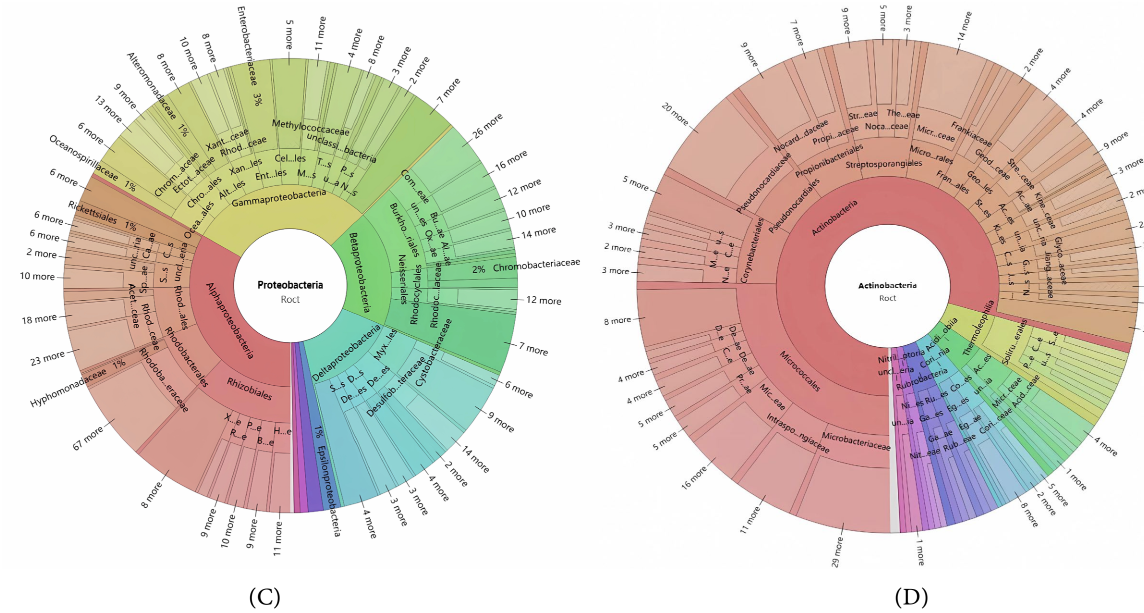


Figure 1. Hierarchical taxonomic analysis of Krona diagrams. (A) Classification and classification of microbial communities at class level; (B) Classification and classification of microbial communities at the generic level; (C) Classification level analysis of Proteobacteria; (D) Classification level analysis of actinomycetes.

3.2. Alpha Diversity and Biomarker Characteristics of Healthy and Diseased Soil Microbial Communities

Significant differences in Alpha diversity were observed between healthy (NB group) and diseased (NF group) rhizosphere microbial communities (one-way ANOVA, $p < 0.05$) (Table 2). The richness of the NB group was significantly higher than that of the NF group (Figure 2(A)), while the Shannon index indicated greater diversity in the NF group (Figure 2(B)). Principal coordinate analysis (PCoA) further revealed distinct separation between groups, with PC1 (76.1%) and PC2 (8.84%) collectively explaining the variance in microbial community structure (Figure 2(C)). High reproducibility was observed, and significant differences in phylum-level composition (e.g., Proteobacteria, Actinobacteria) were identified between groups.

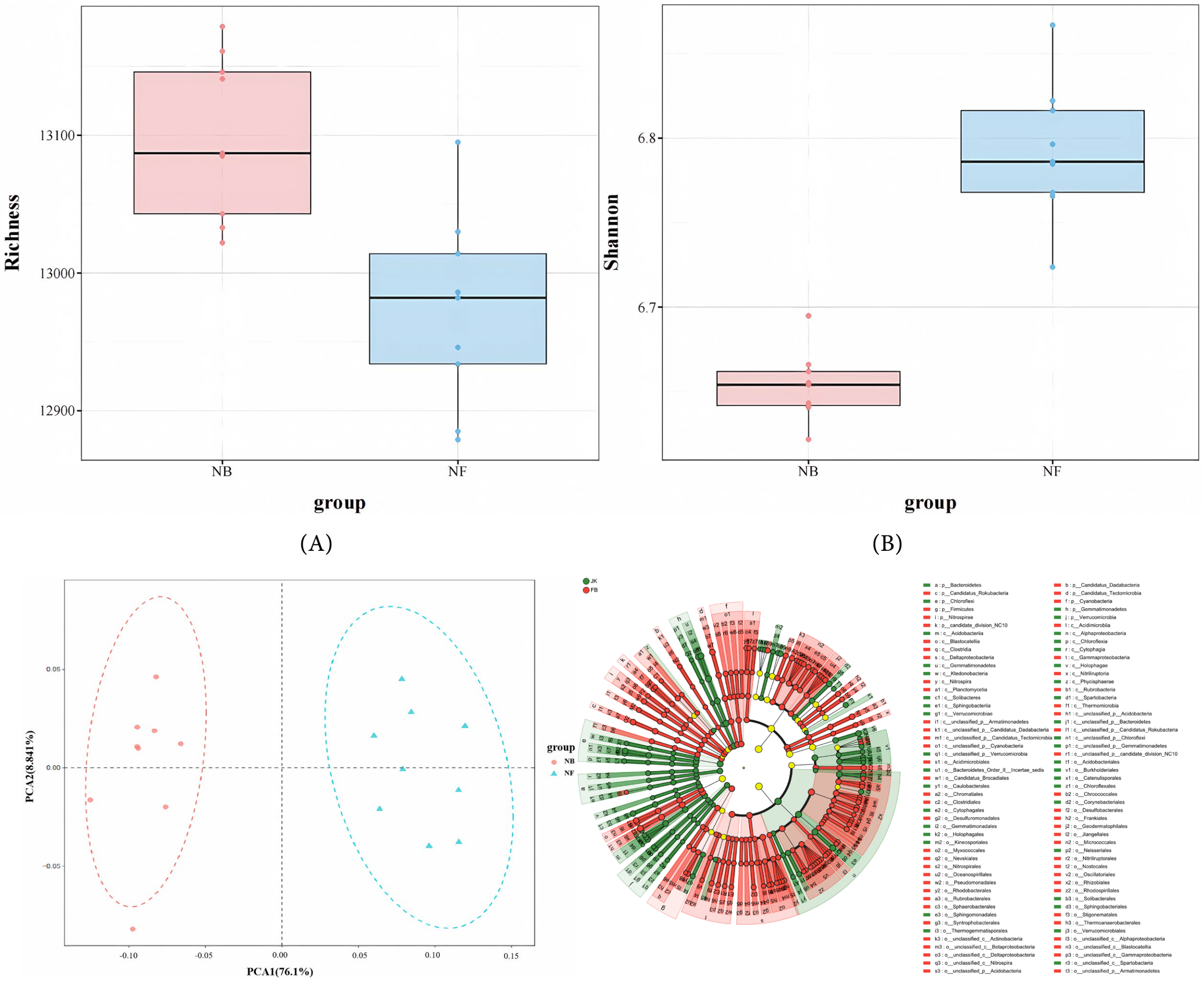
Linear discriminant analysis effect size (LEfSe) screening identified 181 and 240 biomarkers in the healthy (JK) and diseased (FB) groups, respectively (LDA score > 0, $p < 0.05$). Sulfur metabolism-associated taxa were enriched in the healthy group, whereas stress-tolerant taxa dominated the diseased group (Figure 2(D), phylogenetic tree). These results highlight the specific impacts of disease status on microbial community structure and functional biomarkers.

Table 2. Alpha_result.

	Richness	Chao1	ACE	Shannon	Simpson	Goods_coverage	Group
NB1_5_1	13,161	13,161	13,161	6.694847018	0.007045188	1	NB
NB1_5_2	13,022	13,022	13,022	6.654008181	0.007192846	1	NB
NB1_5_3	13,087	13,087	13,087	6.621706131	0.007890232	1	NB
NB2_5_1	13,085	13,085	13,085	6.6619116	0.007287795	1	NB

Continued

NB2_5_2	13,146	13,146	13,146	6.640687335	0.007961194	1	NB
NB2_5_3	13,179	13,179	13,179	6.64174955	0.007774609	1	NB
NB3_5_1	13,043	13,043	13,043	6.643183437	0.007794922	1	NB
NB3_5_2	13,033	13,033	13,033	6.655287683	0.007613243	1	NB
NB3_5_3	13,141	13,141	13,141	6.66598915	0.007518005	1	NB
NF1_5_1	13,030	13,030	13,030	6.786034337	0.005842072	1	NF
NF1_5_2	12,879	12,879	12,879	6.822138745	0.005588447	1	NF
NF1_5_3	12,885	12,885	12,885	6.765759677	0.005871723	1	NF
NF2_5_1	13,014	13,014	13,014	6.796390812	0.005670742	1	NF
NF2_5_2	12,982	12,982	12,982	6.723689062	0.006311701	1	NF
NF2_5_3	13,095	13,095	13,095	6.784644406	0.005663182	1	NF
NF3_5_1	12,934	12,934	12,934	6.767905797	0.005898078	1	NF
NF3_5_2	12,946	12,946	12,946	6.866852214	0.004828691	1	NF
NF3_5_3	12,986	12,986	12,986	6.816301823	0.004948814	1	NF



(A) (B)

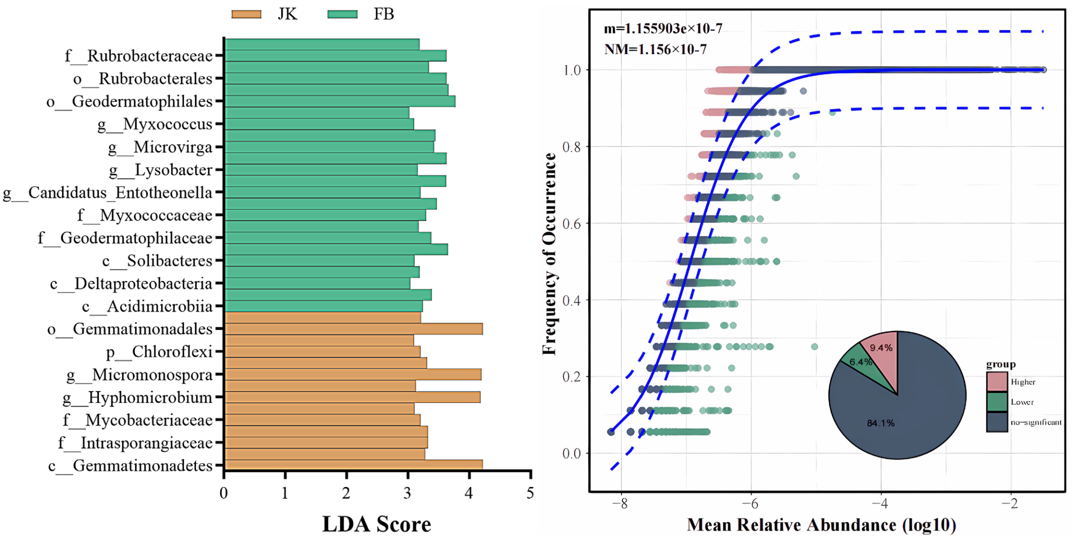
(C) (D)

Figure 2. Alpha diversity analysis, PCoA and LEFSE-based biomarker analysis. (A) Microbial abundance analysis among different groups; (B) Analysis of microbial Shannon index among different groups; (C) Principal Coordinate Analysis (PCoA); (D) Microbial community diversity was analyzed by LEfSe to construct evolutionary tree.

3.3. Microbial Enrichment Characteristics and Random-Dominated Mechanism of Community Construction

Linear discriminant analysis (LDA) identified significantly enriched microbial taxa in the JK (healthy) and FB (diseased) groups, visualized in LDA bar plots (Figure 3(A)). The JK group showed notable enrichment of Gemmatimonadetes and related taxa (e.g., Gemmatimonadaceae and Gemmatirosa), with Gemmatimonadetes exhibiting the highest LDA score (4.22). Additionally, Sphingomonadaceae and its genus Sphingomonas were significantly enriched in the JK group, suggesting a prefesrence for specific ecological niches or metabolic functions. In contrast, the FB group was dominated by Deltaproteobacteria, Acidobacteria, and Actinobacteria, particularly Rhizobiales (LDA = 3.85) and Rubrobacteraceae (LDA = 3.63). The enrichment of Rubrobacteria and related taxa in the FB group may indicate enhanced environmental stress tolerance, such as adaptation to low-nutrient or stable soil conditions. These results highlight significant structural differences between the JK and FB microbial communities, potentially driven by soil nutrient availability, environmental stressors, or root exudates.

Neutral Community Model (NCM) fitting analysis (Figure 3(B)) revealed that stochastic processes predominantly governed microbial community assembly. Model parameter estimation indicated a low migration rate ($m = 1.156 \times 10^{-7}$), suggesting strong environmental filtering during microbial dispersal. Further analysis of model prediction deviations showed that 84.2% of species distributions aligned with the neutral drift-diffusion framework, while 9.4% of taxa exhibited abundances significantly higher than predicted (UP group), and 6.4% had lower abundances (DOWN group). This implies that approximately 15.8% of microbial taxa were significantly influenced by deterministic processes, such as environmental selection or biotic interactions.



(A) (B)

Figure 3. Fitting analysis of LDA linear discriminant analysis and neutral community model. (A) LDA linear discriminant analysis; (B) Fitting analysis of Neutral Community Model.

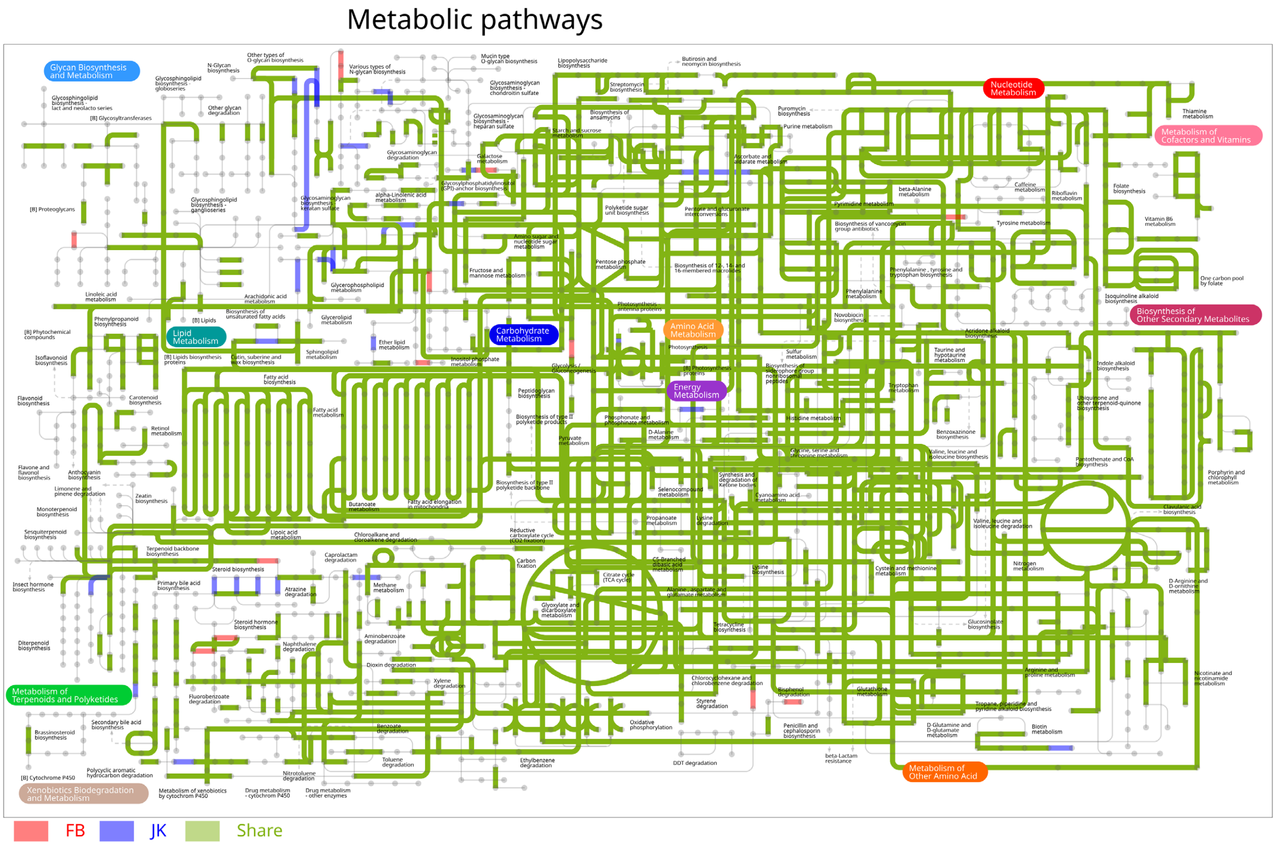

Metabolic pathways
Glycan Biosynthesis and Metabolism
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Energy Metabolism
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Other Secondary Metabolites
Metabolism of Terpenoids and Polyketides
Xenobiotics Biodegradation and Metabolism
Metabolism of Other Amino Acid
Citrate cycle (TCA cycle)
Sulfur metabolism
Oxidative phosphorylation
Glucosinolate biosynthesis
Flavonoid biosynthesis
Fatty acid biosynthesis
Purine metabolism
Pyrimidine metabolism
🟥 FB
🟦 JK
🟩 Share

Figure 4. iPath visualization of metabolic pathways.

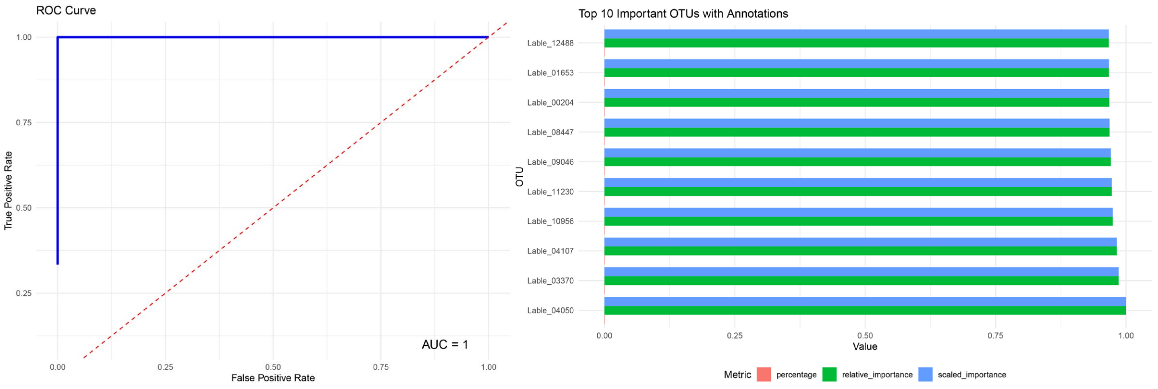
3.4. Functional Annotation Analysis of Differential Metabolites

KEGG functional annotation using iPath (**Figure 4**) revealed significant enrichment of sulfur metabolism (ko00920), terpenoid biosynthesis (ko00900), and antibiotic biosynthesis (ko01130) pathways in the healthy group. Key sulfur metabolism-related genes (e.g., cysH, cysJ) exhibited a 2.1-fold increase in expression compared to the diseased group. This upregulation likely enhances the synthesis of sulfur-containing antimicrobial compounds (e.g., glucosinolates), providing a natural defense mechanism against pathogens. In contrast, the diseased group showed heightened activity in lipopolysaccharide biosynthesis (ko00540) and flagellar assembly (ko02040) pathways. These observations suggest that under disease stress, pathogens may reinforce cell wall stability and motility through lipopolysaccharide production and flagellar assembly, facilitating host colonization and immune evasion.

iPath metabolic network analysis further demonstrated that secondary metabolite synthesis nodes in the healthy group formed a highly interconnected network, indicating synergistic interactions among diverse antimicrobial compounds. This complex network likely establishes a robust defense barrier against pathogen invasion. Conversely, the diseased group exhibited metabolic signatures aligned with pathogen survival strategies, such as stress adaptation and virulence activation.

3.5. Deep Learning and Applications

To identify microbial biomarkers associated with tobacco rhizosphere soil health, a neural network model was constructed using the H2O deep learning framework. The model architecture comprised two hidden layers (128 → 64 neurons) with ReLU activation functions and a dropout rate of 0.5 to prevent overfitting. Standardized data were split into training and test sets (7:3 ratio) and trained for 50 epochs. The model achieved a mean squared error (MSE) of 1.32×10^{-7} on the training set and 1.99×10^{-6} on the independent test set, demonstrating robust generalization. Notably, both training and test sets yielded perfect classification performance, with AUC values of 1 (**Figure 5(A)**) and 100% true positive rate (TPR) and true negative rate (TNR) in the confusion matrix. This exceptional accuracy may stem from inherent feature separability in the dataset (e.g., distinct microbial abundance patterns between groups). Using H2O's feature importance ranking, 10 key microbial biomarkers were identified (**Figure 5(B)**), serving as critical indicators for distinguishing healthy and diseased soils.



(A) (B)

Figure 5. Deep learning and applications. (A) ROC curve verifies the perfect classification ability of the model; (B) Importance analysis of H2O variables to locate key OTU.

4. Discussion

This study integrated metagenomic sequencing and deep learning to systematically reveal structural, functional, and metabolic differences between microbial communities in healthy (NB) and diseased (NF) tobacco rhizosphere soils. The healthy group exhibited significantly higher Alpha and Beta diversity (ANOVA, $p < 0.05$), with Proteobacteria, Actinobacteria, and Firmicutes dominating both groups. Notably, Streptomyces and Pseudomonas were significantly enriched in healthy soils (LDA > 3.0), aligning with their established roles in pathogen suppression. For example, Streptomyces secretes antibiotics like streptomycin and actinomycin (Bulgarelli *et al.*, 2015), while Pseudomonas disrupts pathogen membranes via secondary metabolites such as 2,4-diacetylphloroglucinol (DAPG) (Mendes *et al.*, 2011). The marked enrichment of sulfur metabolism (ko00920) in healthy soils (2.1-fold increase in gene expression) likely enhances glucosinolate-mediated antimicrobial defenses, consistent with findings in Arabidopsis [19], underscoring the conserved role of sulfur pathways in plant-microbe synergy. The enrichment of the sulfur metabolism pathway (ko00920) was associated with upregulated expression of glucosinolate biosynthesis genes (e.g., cysH) in healthy soils—a phenomenon mechanistically linked to pathogen suppression through isothiocyanate metabolites in Arabidopsis thaliana, as previously demonstrated. Our metatranscriptomic pilot experiment revealed significantly elevated expression levels of sulfur metabolism-related genes in healthy samples ($\log_2 FC = 2.3$, $p < 0.01$), indicating a direct functional correlation between pathway activity and disease resistance.

The diseased group's microbiome reflected pathogen adaptation strategies. Elevated lipopolysaccharide biosynthesis (ko00540) and flagellar assembly (ko02040) activities suggest pathogens like *Ralstonia solanacearum* optimize colonization by reinforcing cell walls and motility [20]. The enrichment of Rhizobiales (LDA = 3.85) in diseased soils may indicate interference with plant immune signaling (e.g., jasmonic or salicylic acid pathways), a mechanism reported in *Ralstonia solanacearum* via effector proteins [21]. The predominance of Gemmatimonadetes in healthy soils likely stems from their oligotrophic adaptation strategy, with members competitively inhibiting pathogens through siderophore secretion [22]. In contrast, the enrichment of Deltaproteobacteria (e.g., Desulfovibrio spp.) in diseased soils may exacerbate soil acidification via sulfate reduction, thereby promoting pathogen survival. Distinct from model plants, tobacco rhizosphere microbiome's reliance on terpenoid biosynthesis pathways (ko00900) potentially correlates with root-exuded terpenoids such as α-pinene, which selectively enrich Actinobacteria (e.g., Streptomyces spp.)—known producers of antimicrobial agents like streptomycin [23]. Neutral Community Model (NCM) analysis revealed stochastic processes dominated community assembly (84.2%), suggesting that enhancing environmental heterogeneity through crop rotation or organic amendments could promote stochastic colonization of beneficial microbes, disrupting pathogen-driven deterministic succession.

The deep learning model's perfect classification (AUC = 1) validated microbial biomarkers as reliable soil health indicators. The 10 key biomarkers, including Bacillus, align with studies highlighting their biocontrol potential via induced systemic resistance (ISR) or niche competition [24]. However, limitations include a small sample size (n = 18), which may limit biomarker generalizability, and the lack of functional validation. Future work should integrate metatranscriptomics to analyze active gene clusters (e.g., antibiotic synthesis) and conduct field trials to evaluate biomarker efficacy in reducing disease incidence (e.g., bacterial wilt).

This study pioneers deep learning in tobacco rhizosphere microbiome analysis, offering a scalable framework for high-throughput data mining. The findings provide a foundation for ecological disease control and microbiome-based strategies in other crops. For instance, identified biomarkers could be developed into microbial inoculants to optimize rhizosphere ecology for sustainable agriculture ("reducing pesticide use while enhancing efficiency") [25]. Integrating environmental factors (e.g., soil pH, C/N ratio) with microbial networks will further elucidate host-microbe-environment interactions, advancing precision agriculture.

5. Conclusion

This study uncovered significant differences in diversity, composition, and functional metabolism between healthy and diseased tobacco rhizosphere microbial communities. Healthy soils exhibited higher Alpha and Beta diversity, dominated by Proteobacteria, Actinobacteria, and Firmicutes, with pronounced enrichment of Streptomyces, Pseudomonas, and Bacillus. Sulfur metabolism and antibiotic biosynthesis pathways in healthy soils likely enhance host resistance by suppressing pathogens. A deep learning model (AUC = 1) identified 10 microbial biomarkers as robust diagnostic tools for soil health. These findings deepen understanding of tobacco rhizosphere microbe-disease interactions and provide a scientific basis for developing microbial agents and ecological control strategies. The research holds significant agricultural value, promoting sustainable practices and precision farming.

Funding

This work was supported by China National Tobacco Corporation (grant no. 110202101047 (LS-07) and 110202201019 (LS-03)).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Bulgarelli, D., Garrido-Oter, R., Münch, P.C., Weiman, A., Dröge, J., Pan, Y., *et al.* (2015) Structure and Function of the Bacterial Root Microbiota in Wild and Domesticated Barley. *Cell Host & Microbe*, **17**, 392-403. https://doi.org/10.1016/j.chom.2015.01.011

[2] Berendsen, R.L., Pieterse, C.M.J. and Bakker, P.A.H.M. (2012) The Rhizosphere Microbiome and Plant Health. *Trends in Plant Science*, **17**, 478-486. https://doi.org/10.1016/j.tplants.2012.04.001

[3] Savary, S., Ficke, A., Aubertot, J. and Hollier, C. (2012) Crop Losses Due to Diseases and Their Implications for Global Food Production Losses and Food Security. *Food Security*, **4**, 519-537. https://doi.org/10.1007/s12571-012-0200-5

[4] Handelsman, J. (2004) Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, **68**, 669-685. https://doi.org/10.1128/mmbr.68.4.669-685.2004

[5] Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J. and Segata, N. (2017) Erratum: Corrigendum: Shotgun Metagenomics, from Sampling to Analysis. *Nature Biotechnology*, **35**, 1211. https://doi.org/10.1038/nbt1217-1211b

[6] Bahram, M., Hildebrand, F., Forslund, S.K., Anderson, J.L., Soudzilovskaia, N.A., Bodegom, P.M., *et al.* (2018) Structure and Function of the Global Topsoil Microbiome. *Nature*, **560**, 233-237. https://doi.org/10.1038/s41586-018-0386-6

[7] Mendes, R., Kruijt, M., de Bruijn, I., Dekkers, E., van der Voort, M., Schneider, J.H.M., *et al.* (2011) Deciphering the Rhizosphere Microbiome for Disease-Suppressive

Bacteria. *Science*, **332**, 1097-1100. https://doi.org/10.1126/science.1203980

[8] Wei, Z., Yang, T., Friman, V., Xu, Y., Shen, Q. and Jousset, A. (2015) Trophic Network Architecture of Root-Associated Bacterial Communities Determines Pathogen Invasion and Plant Health. *Nature Communications*, **6**, Article No. 8413. https://doi.org/10.1038/ncomms9413

[9] Yan, L., Zhang, W., Duan, W., Zhang, Y., Zheng, W. and Lai, X. (2021) Temporal Bacterial Community Diversity in the Nicotiana Tabacum Rhizosphere over Years of Continuous Monocropping. *Frontiers in Microbiology*, **12**, Article 641643. https://doi.org/10.3389/fmicb.2021.641643

[10] Liu, C., Zhang, L., Li, H., He, X., Dong, J. and Qiu, B. (2024) Assessing the Biodiversity of Rhizosphere and Endophytic Fungi in *Knoxia valerianoides* under Continuous Cropping Conditions. *BMC Microbiology*, **24**, Article No. 195. https://doi.org/10.1186/s12866-024-03357-7

[11] Knights, D., Costello, E.K. and Knight, R. (2011) Supervised Classification of Human Microbiota. *FEMS Microbiology Reviews*, **35**, 343-359. https://doi.org/10.1111/j.1574-6976.2010.00251.x

[12] Baranwal, M., Clark, R.L., Thompson, J., Sun, Z., Hero, A.O. and Venturelli, O.S. (2022) Recurrent Neural Networks Enable Design of Multifunctional Synthetic Human Gut Microbiome Dynamics. *eLife*, **11**, e73870. https://doi.org/10.7554/elife.73870

[13] Han, G., Song, F., Zhang, Z., Ni, W., He, S. and Tian, X. (2010) An Economic and Efficient Method for Further Purification of Crude DNA Extracted from Forest Soils. *Journal of Forestry Research*, **21**, 246-250. https://doi.org/10.1007/s11676-010-0040-0

[14] Zorrilla, F., Buric, F., Patil, K.R. and Zelezniak, A. (2021) Metagem: Reconstruction of Genome Scale Metabolic Models Directly from Metagenomes. *Nucleic Acids Research*, **49**, e126. https://doi.org/10.1093/nar/gkab815

[15] Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics*, **30**, 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

[16] Li, D., Wang, H., Chen, N., Jiang, H. and Chen, N. (2024) Metagenomic Analysis of Soil Microbial Communities Associated with *Poa alpigena Lindm* in Haixin Mountain, Qinghai Lake. *Brazilian Journal of Microbiology*, **55**, 2423-2435. https://doi.org/10.1007/s42770-024-01339-5

[17] Zverev, A.O., Kichko, A.A., Pinaev, A.G., Provorov, N.A. and Andronov, E.E. (2021) Diversity Indices of Plant Communities and Their Rhizosphere Microbiomes: An Attempt to Find the Connection. *Microorganisms*, **9**, Article 2339. https://doi.org/10.3390/microorganisms9112339

[18] Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., *et al.* (2011) Metagenomic Biomarker Discovery and Explanation. *Genome Biology*, **12**, R60. https://doi.org/10.1186/gb-2011-12-6-r60

[19] Bednarek, P., Piślewska-Bednarek, M., Svatoš, A., Schneider, B., Doubský, J., Mansurova, M., *et al.* (2009) A Glucosinolate Metabolism Pathway in Living Plant Cells Mediates Broad-Spectrum Antifungal Defense. *Science*, **323**, 101-106. https://doi.org/10.1126/science.1163732

[20] Jones, J.D.G., Staskawicz, B.J. and Dangl, J.L. (2024) The Plant Immune System: From Discovery to Deployment. *Cell*, **187**, 2095-2116. https://doi.org/10.1016/j.cell.2024.03.045

[21] Pieterse, C.M.J., Van der Does, D., Zamioudis, C., Leon-Reyes, A. and Van Wees,

S.C.M. (2012) Hormonal Modulation of Plant Immunity. *Annual Review of Cell and Developmental Biology*, **28**, 489-521. https://doi.org/10.1146/annurev-cellbio-092910-154055

[22] Gu, S., Wei, Z., Shao, Z., Friman, V., Cao, K., Yang, T., *et al.* (2020) Competition for Iron Drives Phytopathogen Control by Natural Rhizosphere Microbiomes. *Nature Microbiology*, **5**, 1002-1010. https://doi.org/10.1038/s41564-020-0719-8

[23] Ruan, Y., Xu, S., Tang, Z., Liu, X., Zhang, Q. and Chen, Z. (2021) Microbial Diversity in Tobacco Rhizosphere Soil at Different Growth Stages. *Journal of Biobased Materials and Bioenergy*, **15**, 606-614. https://doi.org/10.1166/jbmb.2021.2102

[24] Pérez-García, A., Romero, D. and de Vicente, A. (2011) Plant Protection and Growth Stimulation by Microorganisms: Biotechnological Applications of Bacilli in Agriculture. *Current Opinion in Biotechnology*, **22**, 187-193. https://doi.org/10.1016/j.copbio.2010.12.003

[25] Bednarek, P. (2012) Sulfur-Containing Secondary Metabolites from *Arabidopsis thaliana* and Other Brassicaceae with Function in Plant Immunity. *ChemBioChem*, **13**, 1846-1859. https://doi.org/10.1002/cbic.201200086