# Microbiome-enabled genomic selection improves prediction accuracy for nitrogen-related traits in maize

Zhikai Yang[1,2], Tianjing Zhao[3], Hao Cheng[3] and Jinliang Yang[1,2,*]

[1]Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68588, USA.
[2]Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE 68583, USA.
[3]Department of Animal Science, University of California Davis, Davis, CA 95616, USA.

*Corresponding author: jinliang.yang@unl.edu

## Abstract

Root-associated microbiomes are increasingly known to play an important role in nutrient acquisition, stress tolerance, and disease resistance of plants. But it remains largely unclear to what extent these microbiomes contribute to trait variation for different genotypes and if their inclusion in the genomic selection (GS) protocol can enhance prediction accuracy. To address these questions, we developed a microbiome-enabled GS (MEGS) method that incorporated host SNPs and ASVs (amplicon sequence variants) from plant root-associated microbiomes in a maize diversity panel under high and low nitrogen (N) field conditions. Our cross-validation results showed that the MEGS model significantly outperformed the conventional GS model for nearly all time-series traits related to plant growth and N responses, with an average relative improvement of 4%. The improvement was more significant for traits measured near the microbiome data collection date and was more pronounced under low N conditions, consistent with the view that some beneficial microbes can enhance N nutrient uptake, particularly in low N fields. Using a high-dimensional mediation analysis method, our study has also identified microbial mediators that establish a link between plant genotype and phenotype. One of the detected mediator microbes, *Massilia putida*, was previously reported to promote plant growth under low N conditions. These large-effect ASVs or microbial mediators can be utilized as biofertilizers or soil additives to enhance crop performance. The improved prediction accuracy of the MEGS models, coupled with the presence of beneficial microbes, paves the way for the future implementation of microbiome-enabled plant breeding in sustainable agriculture.

**Keywords:** genomic prediction; intermediate omics; microbiomes; mediation analysis; maize

## Introduction

An increasing number of studies have suggested that plant-associated microbial communities, especially the microbial species colonizing in the plant roots, can stimulate plant growth (Saleem *et al.* 2019), enhance nutrient availability in soils (Gomes *et al.* 2018; Zhu *et al.* 2016), and decrease abiotic stress responses (Hussain *et al.* 2018; Xu *et al.* 2018). Harnessing these beneficial microbes in crop production provides a promising opportunity for crop improvement to fight against climate challenges, reduce dependency on chemical fertilizers, and boost genetic gain. Indeed, from the beginning of plant domestication, root-associated microbes were reported to be involved in crop performance (Soldan *et al.* 2021). For example, studies have shown that domesticated plants exhibited distinct microbial compositions as compared to the wild ancestors and showed a reduced ability to establish symbiotic relationships with beneficial microbes (Abdelfattah *et al.* 2022; Abdullaeva *et al.* 2021). Recent crop improvement further reduced the microbial diversity in a number of different crop species, including wheat (Hetrick *et al.* 1992), maize (Sangabriel-Conde *et al.* 2014), and soybean (Kiers *et al.* 2007). Realizing the importance of microbiomes in contributing to crop production, recently, efforts have been made to screen for beneficial microbes as potential seed additives to promote plant performance (Singer *et al.* 2021; Yee *et al.* 2021). Though promising results were found in controlled environments (Eida *et al.* 2017; Kaur *et al.* 2020; Sessitsch *et al.* 2019), it is difficult for many microbial inoculants to survive for a long period of time under field conditions (Piromyou *et al.* 2011). Considering the collective genomes of microbial species colonized in plants as the secondary genome of a plant (Berendsen *et al.* 2012), a hologenome (i.e., plant genome with its associated endocellular or extracellular microbiome) approach to incorporate the naturally occurring genotype-specific microbiomes into the genomic selection (GS) protocol provides an alternative strategy to improve tomorrow's crops.

GS, an innovative plant and animal breeding technology, enables the selection of promising individuals (and the associated heritable microbes) to advance to the next generation before or without phenotyping, therefore reducing the generation interval and increasing the genetic gain per unit of time. After the initial introduction of the landmark GS research conducted by Meuwissen *et al.* (2001), animal and plant breeders embraced the unprecedented acceleration of GS with the development of statistical methods and computational tools (de Koning 2016), including G-BLUP (genomic best linear unbiased prediction), RR-BLUP (ridge regression best linear unbiased prediction), and the Bayesian regression models (BayesA, BayesB, BayesC$\pi$, Bayes LASSO, etc.) (Krishnappa *et al.* 2021; Wang *et al.* 2015; Burgueño *et al.* 2012; Wang *et al.* 2018; Gianola *et al.* 2009; Habier *et al.* 2011). The Bayesian regression models (with different assumptions for prior distributions of marker effects) allow genome-wide markers to have different effects and variances and usually achieve high prediction accuracy (Wang *et al.* 2018). However, Bayesian regression models might be more sensitive to the number of QTLs as compared to the RR-BLUP and G-BLUP methods (Wang *et al.* 2015). In addition to these parametric methods, non-parametric methods, such as reproducing kernel Hilbert space (RKHS) (Gianola *et al.* 2006) and deep learning (DL) based methods (Gianola *et al.* 2011), are likely be more efficient for capturing non-additive genetic effects than conventional methods, while there is no clear superiority in terms of prediction power (Montesinos-López *et al.* 2021). Albeit the rapid development of the GS methods in the past 20 years, little attention has been paid to incorporating host-associated microbiomes into the prediction protocols (but see a recent simulation study in dairy cattle (Pérez-Enciso *et al.* 2021)).

For the microbiome-enabled GS (MEGS) modeling, it is straightforward to consider different microbes in a linear mixed model as the random variables, similar to the SNP markers used for conventional GS. However, different from SNPs, microbiomes are dynamic and affected profoundly by root exudates that are composed of organic acids, polysaccharides, and other metabolites (Canarini *et al.* 2019). These differentially recruited microbiomes by different genotypes may, in turn, affect soil physicochemical characteristics and nutrient bioavailability for plants, i.e., nitrogen (N) availability, and eventually regulate plant physiological processes and lead to phenotypic variation. Under such a scenario, the microbiomes can be modeled as an intermediate process to bridge the host genotype and phenotype. Previously, a GS framework integrated Omics data as the intermediate process has been developed by extending the conventional linear mixed model to the neural network (Zhao *et al.* 2022). A similar concept implemented into an associated model allows us to conduct genome-wide mediation analysis to identify large-effect intermediate mediators (Yang *et al.* 2022).

In this study, by leveraging the conventional GS methods and our recently developed mediation models, we conducted integrative analyses using our previously published datasets (Meier *et al.* 2022; Rodene *et al.* 2022) collected on the maize diversity panel — a panel representing maize genetic diversity in temperate latitudes (Flint-Garcia *et al.* 2005). The root-associated microbiomes were collected under high N (HN) and low N (LN) field conditions (Meier *et al.* 2022). From the same fields, phenotypic data were collected using an unmanned aerial vehicle (UAV) in a time-series manner (Rodene *et al.* 2022). In our MEGS analysis, a linear mixed model was used to predict maize phenotype by including both host genotype and root-associated microbiome, where SNPs and ASVs were treated as random effects with different variance components. Additionally, we modeled a presumed causal chain from host genotype to host-associated microbe to host phenotype using our previously developed high-dimensional mediation analysis method (Yang *et al.* 2022). These methods serve as an initial trial to integrate genome and microbiome data, aiming to enhance prediction accuracy and identify beneficial microbes for use as seed additives in field conditions. Overall, our study highlights the potential of beneficial microbes in enhancing crop performance and emphasizes the importance of considering plant-microbe interactions in plant breeding.

## Materials and methods

### Microbiome and phenotype data in the maize diversity panel

The root-associated microbiome data was obtained from our previously published study collected from a subset of the maize diversity panel ($n = 230$ genotypes) eight weeks after planting in both high N and low N field conditions (Meier *et al.* 2022). The microbiome data included $n = 3,626$ amplicon sequence variants (ASVs) that can be clustered into 150 microbial groups. These microbial groups spanned 19 major classes of rhizosphere microbiota. In summary, the relative abundances of $n = 3,626$ ASVs were collected for $n = 795$ observations.

Meanwhile, high throughput phenotyping data were collected from the same field in a time-series manner using UAV (Rodene *et al.* 2022). After image analysis at the plot level, a number of vegetation indices (VIs) were obtained, and some of which were highly correlated with conventional agronomic traits, such as leaf N content and 20 kernel weight. The environmental conditions during which VIs were collected were different, resulting in different qualities. Here we selected the data on 11, 21, and 35 days after microbiome sampling with similar high qualities. More details of phenotypic data can be found in Rodene *et al.* (2022).

### Genotypic data processing and linkage disequilibrium (LD) pruning

The genotypic data for maize HapMap V3.2.1 (with imputation, AGPv4) were obtained from the Panzea database (https://www.panzea.org/genotypes) (Bukowski *et al.* 2017). Using the PLINK software (Purcell *et al.* 2007), we merged the variants on different chromosomes and retained the bi-allelic SNPs only. We then performed SNP filtration by discarding variants with the missing rate $> 0.3$ across lines and a minor allele frequency (MAF) $< 0.05$, resulting in a subset of 22.5 million SNPs. Then, the LD-based SNP pruning was performed by calculating LD between each pair of SNPs in the window of 10 kb, and one of a pair of SNPs was removed if the LD was greater than 0.1. We then shifted the window 10 bp forward and repeated the procedure and resulting in a final subset of 770k SNPs.

### Linear mixed model for microbiome-enabled genomic prediction

We conducted the MEGS using the rrBLUP software (Endelman 2011), where $n = 50,000$ randomly selected SNPs and $n = 3,626$ reproducible ASVs were included simultaneously (Meier *et al.* 2022). Below is the model used for genomic prediction with both maize SNPs and root-associated ASVs:

$$y_{ijk} = \mu + f_i + b_j + \sum_{l=1}^{n} \alpha_l g_{kl} + \sum_{t=1}^{s} \gamma_t m_{ijkt} + \epsilon_{ijk}$$

$$\alpha_l \overset{iid}{\sim} N(0, {\sigma_\alpha}^2)$$
$$\gamma_t \overset{iid}{\sim} N(0, {\sigma_\gamma}^2)$$
$$\epsilon_{ijk} \overset{iid}{\sim} N(0, {\sigma_\epsilon}^2)$$

where, $y_{ijk}$ is the observation of phenotype for the $k$th genotype in the $j$th block with the $i$th N treatment level, and there are 795 observations in total; $\mu$ is the intercept; $f_i$ is the fixed effect of the $i$th N treatment ($i = 1, 2$); $b_j$ is the fixed effect of the $j$th block ($j = 1, 2$); $\alpha_l$ is the random coefficient of the $l$th SNP; $g_{kl}$ is the value of the $l$th SNP for $k$th genotype ($l = 1, ..., n$, where $n = 50,000$ is the total number of SNPs); $\gamma_t$ is the random coefficient of the $t$th ASV ($t = 1, ..., s$, where $s = 3,626$ is the total number of ASVs); $m_{ijkt}$ is the value (log relative abundance from 16S sequencing) of the $t$th ASV for $k$th genotype in $j$th block with the $i$th N level; and $\epsilon_{ijk}$ is the residual error.

In the model, we assumed the random coefficients of the $l$th SNP ($\alpha_l$), the $t$th ASV ($\gamma_t$), and the residual error ($\epsilon$) are independent variables following normal distributions with a mean of zero and estimated variances of $\sigma^2_\alpha$ (i.e., $\sigma^2_\alpha = 2.5 \times e - 8$ for VARI trait at 11 days after microbiome sampling), $\sigma^2_\gamma$ (i.e., $\sigma^2_\gamma = 6.5 \times e - 6$ for VARI trait at 11 days after microbiome sampling), and $\sigma^2_\epsilon$, respectively. To estimate the marker effect variance $\sigma^2_\alpha$, we only fitted SNPs with random effects without ASVs in an RR-BLUP model. Similarly, to get an estimate for ASV effects variance $\sigma^2_\gamma$, we only fitted ASVs with random effects with the first three principal components (PCs) of SNPs (fixed effects) to control for the genomic background effect in an RR-BLUP model.

**Mediation analysis by considering microbes as intermediate variables**

Mediation analysis introduces a variable called a mediator to infer the underlying mechanism of the relationship between an independent variable and a dependent variable (Baron and Kenny 1986). Similar to our previous study (Yang *et al.* 2022), here we conducted genome-wide mediation analysis using those 50,000 SNPs as exposures, 3,626 ASVs as microbe mediators, and the first three principal components (PCs) as confounders. We chose to use the "MedFix" algorithm, which showed higher power in our previous study (Yang *et al.* 2022). The mediation analysis consists of two models: the mediator model and the outcome model. Specifically, the mediator model is shown as follows:

$$\mathbf{M_j} = \mathbf{QA_j} + \mathbf{ZB_j} + \mathbf{e_j}$$

where, $\mathbf{M_j}$ (a $n \times 1$ vector) represents the abundance of the $j$th ASV; $\mathbf{Q}$ (a $n \times s$ matrix) is the first $s$ PCs of genotypes ($s = 3$ in our analysis); $\mathbf{A_j}$ (a $s \times 1$ vector) is the coefficients of the first $s$ PCs to the $j$th ASV; $\mathbf{Z}$ (a $n \times q$ matrix) represents the SNP set of the population composed of $n$ individuals with $q$ number of SNPs; $\mathbf{B_j}$ (a $q \times 1$ vector) is the coefficients of the $q$ SNPs to the $j$th ASV ($j = 1, ..., p$); and $\mathbf{e_j}$ is the vector of the residual errors with $\mathbf{e_j} \sim N(\mathbf{0}, \sigma_j^2 \mathbf{I_n})$.

Additionally, we fitted the SNPs and the microbe mediators (i.e., ASVs) in the outcome model, which uses phenotype as the response variable, as shown below:

$$\mathbf{y} = \mathbf{Qv} + \mathbf{Za} + \mathbf{Mc} + \mathbf{e}$$

where, $\mathbf{y}$ (a $n \times 1$ vector) represents the phenotype; $\mathbf{Q}$ (a $n \times s$ matrix) and $\mathbf{Z}$ (a $n \times q$ matrix) are the same matrices as the above mediator model; $\mathbf{v}$ (a $s \times 1$ vector) is the coefficients of the first $s = 3$ PCs; $\mathbf{a}$ (a $q \times 1$ vector) denotes the coefficients of the SNPs to the phenotype; $\mathbf{M}$ (a $n \times p$ matrix) is the abundance of the ASVs (i.e., a matrix combines all the vectors of $\mathbf{M_j}$); $\mathbf{c}$ (a $p \times 1$ vector) is the coefficients of ASVs to the phenotype; $\mathbf{e}$ is the vector the residual errors with $\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I_n})$.

The SNPs are determined as indirect SNPs for the $j$th ASV if $p_{B_j}$ (the p-values of $\chi^2$ likelihood ratio test for the corresponding $\mathbf{B_j}$) are smaller than 0.05; and the $j$th ASV is determined as a mediator if the $max(p_{B_j}, p_{cj})$

($p_{cj}$ is the p-value of the z-test for the $j$th element of **c**) is smaller than 0.05, with a step-down procedure to control for the false discovery proportion (Zhang 2021).

## Results

**Microbiome incorporation into GS model improves prediction accuracy for host phenotype**

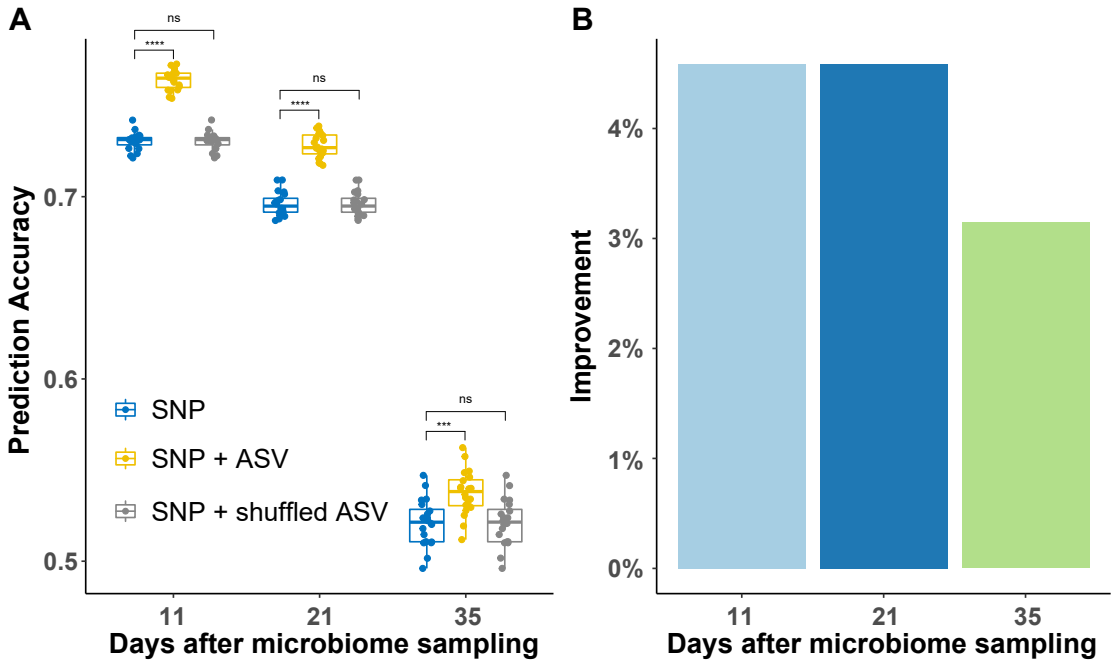


**Figure 1 Comparison of prediction accuracy for an image-extracted trait using different variables in GS models.** (**A**) Genomic prediction results using SNPs only (blue), both SNPs and microbiomes (yellow), and both SNPs and shuffled microbiomes (grey). The statistical significance of the difference in accuracy between the models is indicated by asterisks: ns (not significant), * ($0.01 < \text{p-value} \leq 0.05$), ** ($0.001 < \text{p-value} \leq 0.01$), ***( $0.0001 < \text{p-value} \leq 0.001$), and **** (p-value $\leq 0.0001$). (**B**) Percentage improvement ($\frac{r_{SNP+ASV} - r_{SNP}}{r_{SNP}}$, where $r_{SNP+ASV}$ is the prediction accuracy for the SNP and ASV model and $r_{SNP}$ is the accuracy for the SNP-only model) in prediction accuracy on different dates.

Plant roots are colonized by tens of thousands of microbial species, many of which are heritable (Peiffer *et al.* 2013; Meier *et al.* 2021). To test whether these root-associated microbiomes can be leveraged to predict the plant phenotypic performance, we obtained the composition of microbiomes quantified on a diversity panel of 230 maize inbred lines under high N (HN) and low N (LN) field conditions (Meier *et al.* 2022). Additionally, we obtained the image-extracted vegetation index traits collected from the same field (Rodene *et al.* 2022). Since most of the vegetation index traits exhibit high correlation (see **Figure S1**), we focused on one of the most representative vegetation indices — Visible Atmospherically Resistant Index (VARI) — collected on three dates, i.e., 11, 21, and 35 days after microbiome sampling (DAMS) in the field.

We fitted the ASVs (host root-associated microbiomes) and SNPs (host genomes) as the explanatory variables with random effects into a linear mixed model (**Materials and Methods**). After 20 randomized five-fold cross-validations, results suggested that incorporating ASVs into the model significantly increased the GS prediction accuracy compared to the conventional SNP-only model for all three dates (**Figure 1A**). As the number of days after microbiome sampling increased, the overall prediction accuracy declined, which can be attributed to the low heritability of VAGI traits as plants approach senescence. To ensure that the greater prediction accuracy was not simply due to the larger number of explanatory variables, we randomly shuffled the ASVs to create the same set of dummy variables. We found that the extra number of shuffled variables had no significant effect on prediction accuracy (**Figure 1A**). Compared with the conventional GS model using only SNPs, the percentages

of improvement by adding ASVs were 4.6%, 4.6%, and 3.1% on the 11, 21, and 35 days, respectively (**Figure 1B**). Even compared with results using randomly shuffled ASVs, an average of 4.1% improvement was achieved. The improvement in prediction accuracy was notably higher for traits observed at 11 and 21 days after the sampling of microbiomes in the field, but decreased significantly for traits observed after 35 days. This decline is likely attributed to the diminishing effects of microbiomes on the plant phenotype as plant development progresses. Similar results were also observed for other image-extracted traits (see **Figure S2**).

**Prediction improvement is predominantly observed under low N conditions**

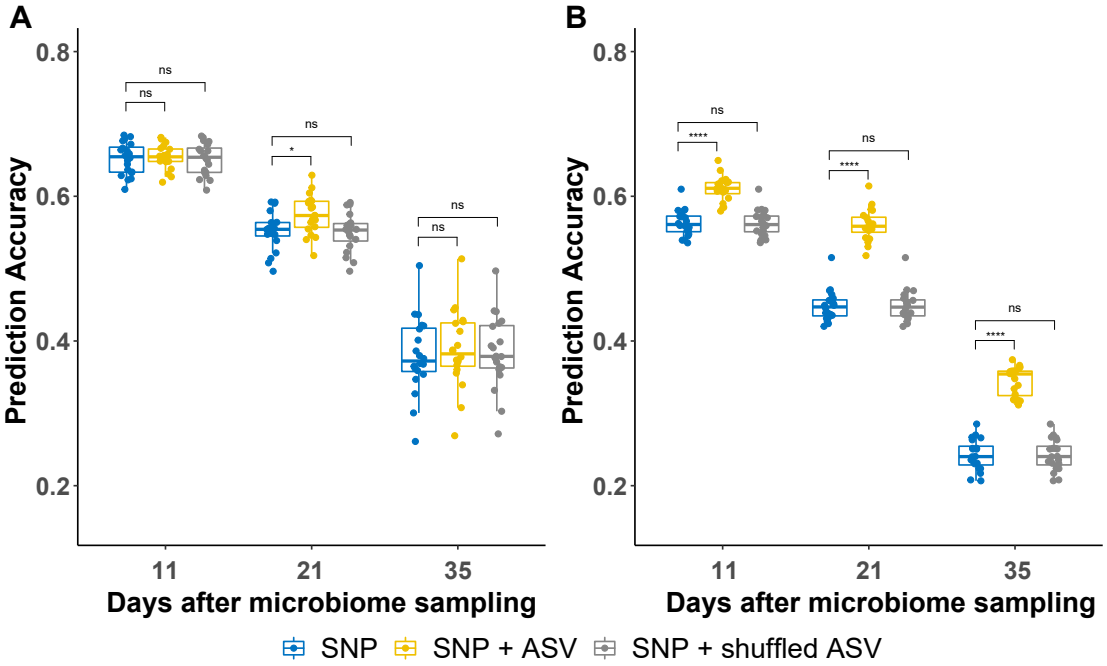


**Figure 2 Prediction accuracy of incorporating microbiome under different N conditions.** The prediction accuracy of using SNPs only (blue), using both SNPs and microbiome (yellow), and using both SNPs and randomly shuffled microbiome (grey) in high N (**A**) and low N (**B**) conditions. The statistical significance of the difference in accuracy between the models is indicated by asterisks: ns (not significant), * ($0.01 < \text{p-value} \leq 0.05$), ** ($0.001 < \text{p-value} \leq 0.01$), ***( $0.0001 < \text{p-value} \leq 0.001$), and **** ($\text{p-value} \leq 0.0001$).

To investigate the impact of N treatment on MEGS results, we fitted the model separately under high N and low N conditions. In the high N conditions, the MEGS model did not outperform the conventional SNP-only model, except for a slight improvement observed at 21 DAMS (**Figure 2A**). Interestingly, significant improvements were observed for all three days under the low N conditions. Specifically, the prediction accuracy increased from 56.3% to 61.1% at 11 DAMS, from 44.8% to 56.0% at 21 DAMS, and from 24.2% to 34.5% at 35 DAMS (**Figure 2B**). These findings support the idea that microbiomes can enhance N nutrient bioavailability in the soil (Jacoby *et al.* 2017), suggesting their potential as functional markers for improving the prediction of plant phenotypes.

**Microbiomes exhibit more pronounced effects in the early days after microbiome sampling**

In order to better understand the impact of specific microbes on prediction, we evaluated the effect sizes of the ASVs and SNPs included in the prediction model (**Figure 3**). As expected, the overall SNP effects remained largely consistent across different dates (**Figure 3B**). However, the effects of ASVs declined as the number of days between initial microbiome sampling and plant phenotype collection increased (**Figure 3A**). We identified four ASVs that consistently had relatively large effects across all dates (i.e., they were among the top 1% of absolute coefficients listed in **Supplementary Table S1**). These ASVs were annotated as *Candidatus*

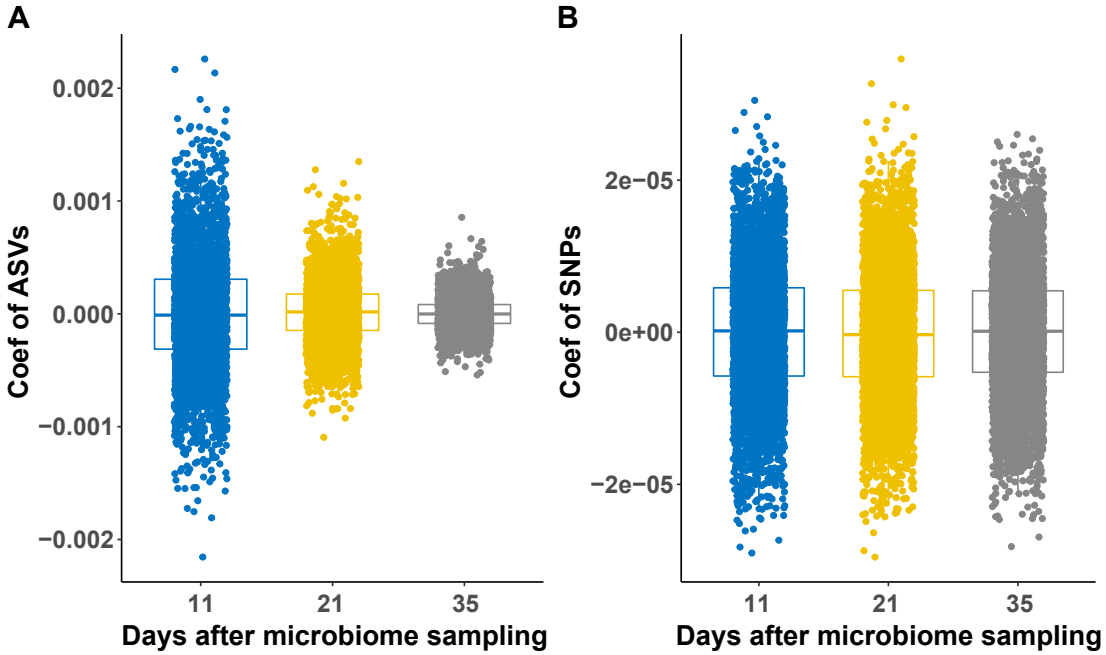


**Figure 3 The effect sizes of ASVs and SNPs on the VARI phenotype.** The distribution of effect size of ASVs (**A**) and SNP (**B**) at different DAMS.

*Udaeobacter copiosus*, *Massilia niabensis*, *Pseudomonas parafulva*, and *Sphingomonas limnosediminicola*. Interestingly, *Pseudomonas parafulva* (Oteino *et al.* 2015; Preston 2004) and *Massilia niabensis* (Yu *et al.* 2021) have previously been reported to promote plant growth. Some of the other top 1% ASVs, which were only detected on certain dates, may also be important candidates for further investigation. For example, *Bacillus fumarioli* was only detected as one of the top 1% ASVs at 11 DAMS. Previous studies have suggested that *Bacillus fumarioli* is likely to be under host selection (Meier *et al.* 2022) and may promote plant growth (Kumar *et al.* 2012).

To identify shared features of the ASVs with the largest effects, we compared the top 1% ASVs with all the other ASVs with regard to their heritability and selection scores, as calculated by Meier *et al.* (2022). The results indicated that although the top 1% ASVs were not more heritable under either high or low N field conditions, they had significantly higher selection scores under low N conditions on 11 and 21 DAMS (**Figure S3**). These findings are consistent with the greater improvement in prediction accuracy observed in the low N field due to the microbiome. Furthermore, they suggest that plant hosts may have a genetic mechanism for recruiting certain microbes to mitigate low N stress.

### Microbiome mediation analysis reveals promising microbial mediators

In order to establish a causal chain from plant genotype to microbiome to plant phenotype, we sought to model ASVs as the intermediate mediators that are selectively recruited by different plant genotypes and have a significant effect on plant phenotype. Therefore, we supplemented the MEGS with our previously developed genome-wide mediation analysis (Yang *et al.* 2022). Through this approach (see **Materials and Methods**), we identified five unique ASVs that act as mediators for 8 VIs (**Supplementary Table S2**). Among these, *Massilia putida* and *Burkholderia pseudomallei* were the top 1% largest effect ASVs in the MEGS model.

To examine whether there were significant differences in the accumulation of *Massilia putida* among different genotypes, we divided the population based on SNP genotype at the largest effect indirect SNP (or iSNP, an SNP that significantly associated with the relative abundance of an ASV). The iSNP 1-240925424, located on the gene body of GDP-mannose-3′5′-epimerase1 (Zm00001d032950), showed the minor allele of T and the major allele of A. Our microbiome mediation analysis revealed significant differences between the two alleles under both high N and low N conditions (**Figure 4A**). To investigate the effects of *Massilia putida* on plant

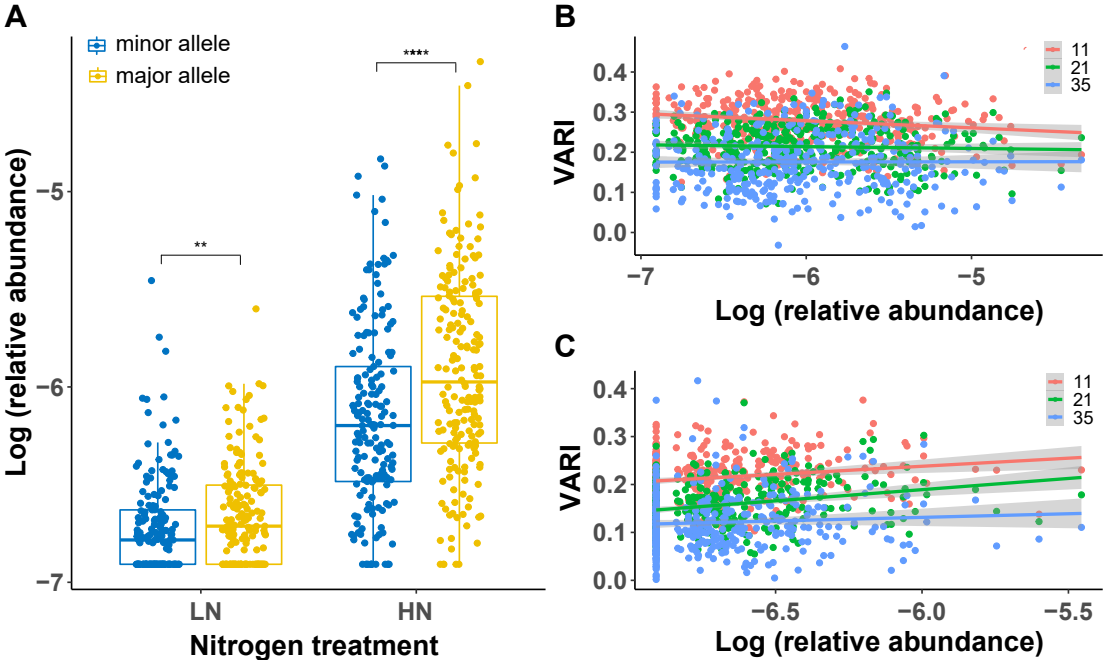


**Figure 4 The relative abundance of *Massilia putida* and its relationship with the VARI phenotype.** (**A**) The differences of the relative abundance of *Massilia putida* between major and minor alleles at the iSNP under low N (LN) and high N (HN) conditions. The statistical significance of the Wilcoxon rank sum and signed rank test is indicated by asterisks: ns (not significant), * ($0.01 <$ p-value $\leq 0.05$), ** ($0.001 <$ p-value $\leq 0.01$), ***( $0.0001 <$ p-value $\leq 0.001$), and **** (p-value $\leq 0.0001$). The correlations between the relative abundance of *Massilia putida* and the VARI phenotype collected on 11, 21, and 35 DAMS under high N (**B**) and low N (**C**) conditions.

phenotype, we performed Spearman's rank correlation test to determine if there was a significant correlation between the abundance of *Massilia putida* and the VARI phenotype. Under high N conditions, we found barely negative correlations on 11 ($r = -0.16$, $p$-value $= 3.2 \times e - 3$), 21 ($r = -0.026$, $p$-value $= 0.62$), and 35 ($r = 0.023$, $p$-value $= 0.66$) days (**Figure 4B**). Under low N conditions, results showed significantly larger positive correlations on days 11 ($r = 0.19$, $p$-value $= 2.9 \times e - 4$) and 21 ($r = 0.25$, $p$-value $= 2.1 \times e - 6$), yet the correlation on 35 DAMS was insignificant ($r = 0.08$, $p$-value $= 0.13$) (**Figure 4C**). These results are consistent with the previous finding that *Massilia putida* promotes plant growth under low N conditions (Yu *et al.* 2021).

## Discussion

In this study, we developed a microbiome-enabled genomic selection (MEGS) method and provided empirical evidence that incorporating microbiome data led to a significant increase (about 4%) in prediction accuracy for most traits extracted from UAV images across different dates. We acknowledged that there might be correlations between SNPs (host genome) and ASVs (microbiome), leading to multicollinearity, but this won't affect the screen for beneficial ASVs as seed additives for improving crop production. Besides, the spatial and temporal stability of the microbiome affects the efficacy of prediction models (Pérez-Enciso *et al.* 2021). Several studies suggest that the microbiome residing in the roots undergo alterations throughout the lifespan of an individual plant (Mougel *et al.* 2006; Houlden *et al.* 2008; Yu *et al.* 2012; Chaparro *et al.* 2014). However, research has also indicated that the microbiome could potentially reach a relatively stable condition after the initial two weeks of growth (Ibekwe and Grieve 2004; Edwards *et al.* 2015; Aleklett *et al.* 2022). Consequently, to enhance the accuracy of predictions, incorporating microbial data collected during a pre-determined stable stage would be recommended. This preference for stable stage significantly restricts the application of microbiome in

prediction within breeding programs when compared to genomic data, which remains unchanged (Pérez-Enciso *et al.* 2021). Nonetheless, significant potential applications still exist given the increased prediction accuracy (about 4% on average) after incorporating the microbiome data into the prediction models. As the cost of obtaining ASV data through 16S rRNA sequencing decreases with advances in sequencing technology, MEGS presents an unprecedented opportunity to predict complex traits, such as N or water usage efficiency, by including ASVs as additional explanatory variables.

We analyzed the prediction accuracy under high and low N conditions separately and discovered that microbiome data was more beneficial under the latter. This is consistent with the idea that plants may recruit specific microbes to facilitate nutrient absorption when soil N is inadequate. Our mediation analysis, which considered the microbes as intermediaries between plant genotype and plant phenotype, pinpointed five microbe mediators. Among these, two mediator microbes, namely *Massilia putida* and *Burkholderia pseudomallei*, played crucial roles in predicting the plant phenotype in the MEGS analysis. Notably, *Massila putida* was enriched in the rhizosphere and is thought to enhance plant growth and N uptake by inducing lateral root formation under low N conditions (Yu *et al.* 2021). We also detected several other large-effect ASVs, such as *Pseudomonas parafulva* and *Bacillus fumarioli*, some of which have been linked to plant development (Oteino *et al.* 2015; Preston 2004; Kumar *et al.* 2012). However, further phenotypic and functional validation of these identified microbe mediators and large-effect microbes is necessary to reveal their effects on the host plants.

Due to computational constraints, we randomly sampled a subset of SNPs in low LD across the whole genome to fit the MEGS models. To determine the influence of SNP size on prediction accuracy, we performed a sensitivity test and found no significant difference in prediction accuracy when random sampling 10k, 25k, and 50k SNPs out of the 770k low LD SNPs. The prediction accuracy was reasonably high, ranging from 52.1%-76.4% for both SNP-only and SNP-ASV models across different dates (**Figure S4**). To maintain high fidelity under the current computation resource, we chose 50k SNPs for all traits and dates in the study. However, the use of more computationally efficient methods is necessary to handle larger SNP datasets in the future. Although the present study was able to achieve reasonable prediction accuracy with a limited set of SNPs and the whole set of ASVs, the use of more comprehensive datasets would undoubtedly improve the accuracy of the MEGS models.

In summary, our findings suggest that incorporating ASVs into the GS model enhances prediction accuracy compared to the traditional SNP-only model, particularly for traits collected under low N conditions and at earlier time points after microbiome sampling in the field. However, the overall prediction accuracy for different dates decreases as plants approach senescence, indicating that the microbiome, as a functional marker, is sensitive to both time and environment. To successfully implement MEGS in a breeding program, it is crucial to carefully design experiments for in-field microbiome data collection to ensure the quality of prediction results.

## Data and Code Availability

The data and code used for the analyses can be accessed through GitHub (https://github.com/ZhikaiYang/GP_microbiome).

## Acknowledgements

This work is supported by the Agriculture and Food Research Initiative Grant number 2019-67013-29167 and 2022-67013-3656 from the USDA National Institute of Food and Agriculture, and by the U.S. Department of Energy under Award Number DE-SC0023138. This work is conducted using the Holland Computing Center of the University of Nebraska-Lincoln Start-up, which receives support from the Nebraska Research Initiative.

## Conflicts of interest

The authors declare no competing interests.

## Literature cited


Abdelfattah A, Tack AJ, Wasserman B, Liu J, Berg G, Norelli J, Droby S, Wisniewski M. 2022. Evidence for host–microbiome co-evolution in apple. New Phytologist. 234:2088–2100.

Abdullaeva Y, Ratering S, Manirajan BA, Rosado-Porto D, Schnell S, Cardinale M. 2021. Domestication impacts the wheat-associated microbiota and the rhizosphere colonization by seed-and soil-originated microbiomes, across different fields. Frontiers in plant science. 12.

Aleklett K, Rosa D, Pickles BJ, Hart MM. 2022. Community assembly and stability in the root microbiota during early plant development. Frontiers in Microbiology. 13:826521.

Baron RM, Kenny DA. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of personality and social psychology. 51:1173.

Berendsen RL, Pieterse CM, Bakker PA. 2012. The rhizosphere microbiome and plant health. Trends in plant science. 17:478–486.

Bukowski R, Guo X, Lu Y, Zou C, He B, Rong Z, Wang B, Xu D, Yang B, Xie C *et al*. 2017. Construction of the third-generation zea mays haplotype map. GigaScience. 7:gix134.

Burgueño J, de los Campos G, Weigel K, Crossa J. 2012. Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. Crop Science. 52:707–719.

Canarini A, Kaiser C, Merchant A, Richter A, Wanek W. 2019. Root exudation of primary metabolites: mechanisms and their roles in plant responses to environmental stimuli. Frontiers in Plant Science. 10:157.

Chaparro JM, Badri DV, Vivanco JM. 2014. Rhizosphere microbiome assemblage is affected by plant development. The ISME journal. 8:790–803.

de Koning DJ. 2016. Meuwissen et al. on genomic selection. Genetics. 203:5–7.

Edwards J, Johnson C, Santos-Medellín C, Lurie E, Podishetty NK, Bhatnagar S, Eisen JA, Sundaresan V. 2015. Structure, variation, and assembly of the root-associated microbiomes of rice. Proceedings of the National Academy of Sciences. 112:E911–E920.

Eida AA, Hirt H, Saad MM. 2017. Challenges faced in field application of phosphate-solubilizing bacteria, In: , Springer. pp. 125–143.

Endelman JB. 2011. Ridge regression and other kernels for genomic selection with r package rrblup. The Plant Genome. 4.

Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES. 2005. Maize association population: a high-resolution platform for quantitative trait locus dissection. The Plant Journal. 44:1054–1064.

Gianola D, De Los Campos G, Hill WG, Manfredi E, Fernando R. 2009. Additive genetic variability and the bayesian alphabet. Genetics. 183:347–363.

Gianola D, Fernando RL, Stella A. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics. 173:1761–1776.

Gianola D, Okut H, Weigel KA, Rosa GJ. 2011. Predicting complex quantitative traits with bayesian neural networks: a case study with jersey cows and wheat. BMC genetics. 12:1–14.

Gomes E, Lana U, Quensen J, de Sousa S, Oliveira C, Guo J, Guimarães L, Tiedje J. 2018. Root-associated microbiome of maize genotypes with contrasting phosphorus use efficiency. phytobiomes j 2: 129–137.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ. 2011. Extension of the bayesian alphabet for genomic selection. BMC bioinformatics. 12:1–12.

Hetrick B, Wilson G, Cox T. 1992. Mycorrhizal dependence of modern wheat varieties, landraces, and ancestors. Canadian Journal of Botany. 70:2032–2040.

Houlden A, Timms-Wilson TM, Day MJ, Bailey MJ. 2008. Influence of plant developmental stage on microbial community structure and activity in the rhizosphere of three field crops. FEMS microbiology ecology. 65:193–201.

Hussain SS, Mehnaz S, Siddique KH. 2018. Harnessing the plant microbiome for improved abiotic stress tolerance, In: , Springer. pp. 21–43.

Ibekwe A, Grieve C. 2004. Changes in developing plant microbial community structure as affected by contaminated water. FEMS microbiology ecology. 48:239–248.

Jacoby R, Peukert M, Succurro A, Koprivova A, Kopriva S. 2017. The role of soil microorganisms in plant mineral nutrition—current knowledge and future directions. Frontiers in plant science. 8:1617.

Kaur T, Rana KL, Kour D, Sheikh I, Yadav N, Kumar V, Yadav AN, Dhaliwal HS, Saxena AK. 2020. Microbe-mediated biofortification for micronutrients: present status and future challenges, In: , Elsevier. pp. 1–17.

Kiers ET, Hutton MG, Denison RF. 2007. Human selection and the relaxation of legume defences against ineffective rhizobia. Proceedings of the Royal Society B: Biological Sciences. 274:3119–3126.

Krishnappa G, Savadi S, Tyagi BS, Singh SK, Mamrutha HM, Kumar S, Mishra CN, Khan H, Gangadhara K, Uday G *et al*. 2021. Integrated genomic selection for rapid improvement of crops. Genomics. 113:1070–1086.

Kumar P, Dubey R, Maheshwari D. 2012. Bacillus strains isolated from rhizosphere showed plant growth promoting and antagonistic activity against phytopathogens. Microbiological research. 167:493–499.

Meier MA, Lopez-Guerrero MG, Guo M, Schmer MR, Herr JR, Schnable JC, Alfano JR, Yang J. 2021. Rhizosphere microbiomes in a historical maize-soybean rotation system respond to host species and nitrogen fertilization at the genus and subgenus levels. Applied and environmental microbiology. 87:e03132–20.

Meier MA, Xu G, Lopez-Guerrero MG, Li G, Smith C, Sigmon B, Herr JR, Alfano JR, Ge Y, Schnable JC *et al*. 2022. Association analyses of host genetics, root-colonizing microbes, and plant phenotypes under different nitrogen conditions in maize. eLife. 11:e75790.

Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819–1829.

Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Barrón-López JA, Martini JW, Fajardo-Flores SB, Gaytan-Lugo LS, Santana-Mancilla PC, Crossa J. 2021. A review of deep learning applications for genomic selection. BMC genomics. 22:1–23.

Mougel C, Offre P, Ranjard L, Corberand T, Gamalero E, Robin C, Lemanceau P. 2006. Dynamic of the genetic structure of bacterial and fungal communities at different developmental stages of medicago truncatula gaertn. cv. jemalong line j5. New Phytologist. 170:165–175.

Oteino N, Lally RD, Kiwanuka S, Lloyd A, Ryan D, Germaine KJ, Dowling DN. 2015. Plant growth promotion induced by phosphate solubilizing endophytic pseudomonas isolates. Frontiers in microbiology. 6:745.

Peiffer JA, Spor A, Koren O, Jin Z, Tringe SG, Dangl JL, Buckler ES, Ley RE. 2013. Diversity and heritability of the maize rhizosphere microbiome under field conditions. Proc Natl Acad Sci U S A. 110:6548–6553.

Pérez-Enciso M, Zingaretti LM, Ramayo-Caldas Y, de Los Campos G. 2021. Opportunities and limits of combining microbiome and genome data for complex trait prediction. Genetics Selection Evolution. 53:1–20.

Piromyou P, Buranabanyat B, Tantasawat P, Tittabutr P, Boonkerd N, Teaumroong N. 2011. Effect of plant growth promoting rhizobacteria (pgpr) inoculation on microbial community structure in rhizosphere of forage corn cultivated in thailand. European Journal of Soil Biology. 47:44–54.

Preston GM. 2004. Plant perceptions of plant growth-promoting pseudomonas. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences. 359:907–918.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ *et al*. 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 81:559–575.

Rodene E, Xu G, Palali Delen S, Zhao X, Smith C, Ge Y, Schnable J, Yang J. 2022. A uav-based high-throughput phenotyping approach to assess time-series nitrogen responses and identify trait-associated genetic components in maize. The Plant Phenome Journal. 5:e20030.

Saleem M, Hu J, Jousset A. 2019. More than the sum of its parts: microbiome biodiversity as a driver of plant growth and soil health. Annual Review of Ecology, Evolution, and Systematics. 50:145–168.

Sangabriel-Conde W, Negrete-Yankelevich S, Maldonado-Mendoza IE, Trejo-Aguilar D. 2014. Native maize landraces from los tuxtlas, mexico show varying mycorrhizal dependency for p uptake. Biology and fertility of soils. 50:405–414.

Sessitsch A, Pfaffenbichler N, Mitter B. 2019. Microbiome applications from lab to field: facing complexity. Trends in plant science. 24:194–198.

Singer E, Vogel JP, Northen T, Mungall CJ, Juenger TE. 2021. Novel and emerging capabilities that can provide a holistic understanding of the plant root microbiome. Phytobiomes Journal. 5:122–132.

Soldan R, Fusi M, Cardinale M, Daffonchio D, Preston GM. 2021. The effect of plant domestication on host control of the microbiota. Communications Biology. 4:1–9.

Wang X, Xu Y, Hu Z, Xu C. 2018. Genomic selection methods for crop improvement: Current status and prospects. The Crop Journal. 6:330–340.

Wang X, Yang Z, Xu C. 2015. A comparison of genomic selection methods for breeding value prediction. Science Bulletin. 60:925–935.

Xu L, Naylor D, Dong Z, Simmons T, Pierroz G, Hixson KK, Kim YM, Zink EM, Engbrecht KM, Wang Y *et al*. 2018. Drought delays development of the sorghum root microbiome and enriches for monoderm bacteria. Proceedings of the National Academy of Sciences. 115:E4284–E4293.

Yang Z, Xu G, Zhang Q, Obata T, Yang J. 2022. Genome-wide mediation analysis: an empirical study to connect phenotype with genotype via intermediate transcriptomic data in maize. Genetics. 221:iyac057.

Yee MO, Kim P, Li Y, Singh AK, Northen TR, Chakraborty R. 2021. Specialized plant growth chamber designs to study complex rhizosphere interactions. Frontiers in Microbiology. p. 507.

Yu L, Nicolaisen M, Larsen J, Ravnskov S. 2012. Succession of root-associated fungi in pisum sativum during a plant growth cycle as examined by 454 pyrosequencing. Plant and soil. 358:225–233.

Yu P, He X, Baer M, Beirinckx S, Tian T, Moya YAT, Zhang X, Deichmann M, Frey FP, Bresgen V *et al*. 2021. Plant flavones enrich rhizosphere oxalobacteraceae to improve maize performance under nitrogen deprivation. Nature Plants. 7:481–499.

Zhang Q. 2021. High-dimensional mediation analysis with applications to causal gene identification. Statistics in Biosciences. .

Zhao T, Zeng J, Cheng H. 2022. Extend mixed models to multilayer neural networks for genomic prediction including intermediate omics data. Genetics. 221:iyac034.

Zhu S, Vivanco JM, Manter DK. 2016. Nitrogen fertilizer rate affects root exudation, the rhizosphere microbiome and nitrogen-use-efficiency of maize. Applied Soil Ecology. 107:324–333.

## Supplemental Tables

**Table S1** Top 1% ASVs detected by the MEGS model. (https://github.com/ZhikaiYang/GP_microbiome/blob/master/data/supplementary/Supplemental_Table_S1_top_one_percent_asvs.txt)

**Table S2** Mediators ASVs identified by mediation analysis. (https://github.com/ZhikaiYang/GP_microbiome/blob/master/data/supplementary/Supplemental_Table_S2_mediator_asvs.txt)

## Supplemental Figures

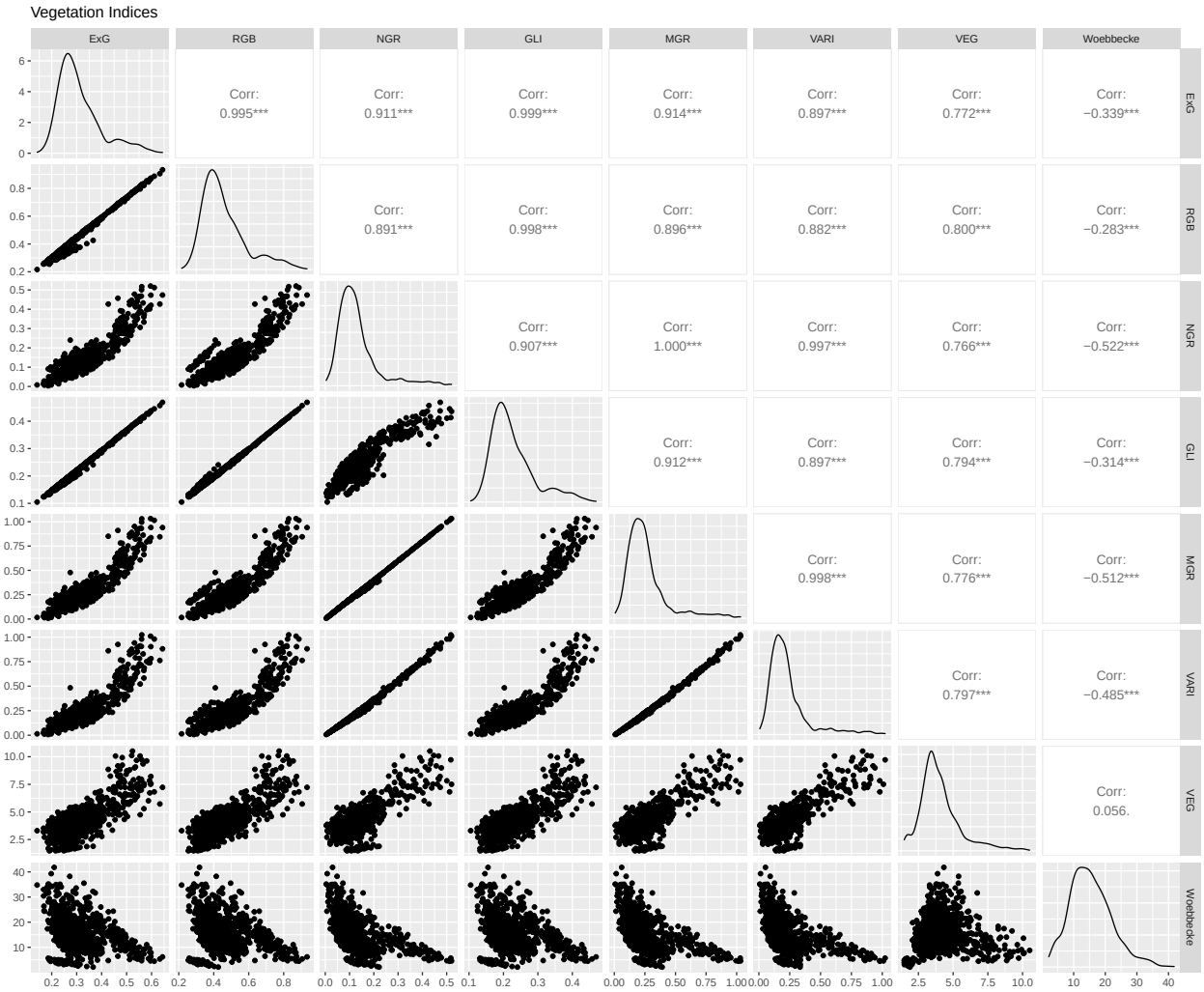


**Figure S1 The distributions and correlations of the eight vegetation indices (VI).** The upper right panel shows the pair-wise correlations, the lower left panel shows the corresponding scatter plots of each VI, and the diagonal shows the density plots of each VI. The statistical significance of the Pearson correlation is indicated by asterisks: * ($0.01 < \text{p-value} \leq 0.05$), ** ($0.001 < \text{p-value} \leq 0.01$), ***( $0.0001 < \text{p-value} \leq 0.001$), and **** ($\text{p-value} \leq 0.0001$).

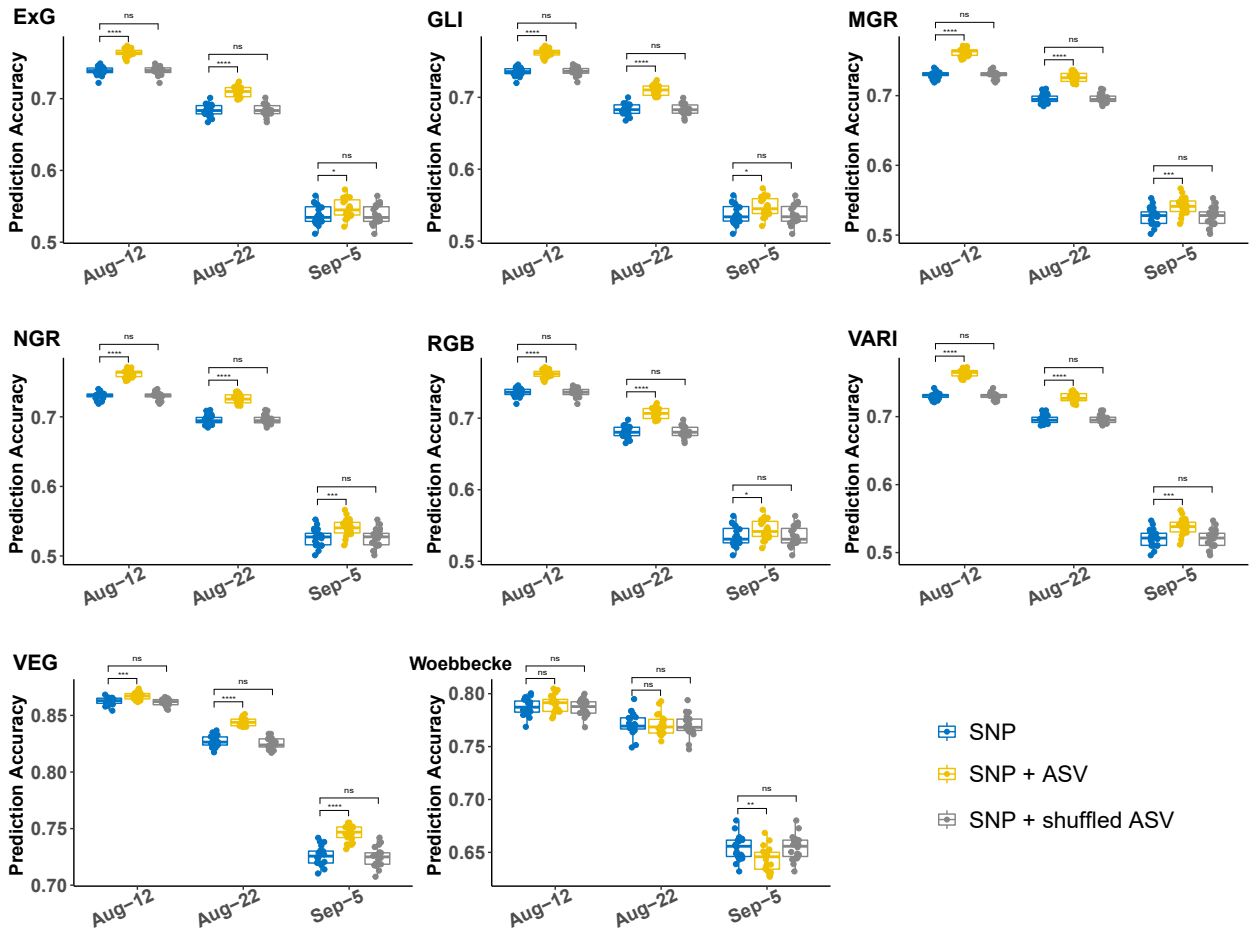


**Figure S2 Prediction accuracy of VIs by incorporating microbiome into the prediction model.** Genomic prediction results using SNPs only (blue), both SNPs and microbiomes (yellow), and both SNPs and shuffled microbiomes (grey). The statistical significance of the difference in accuracy between the models is indicated by asterisks: ns (not significant), * ($0.01 <$ p-value $\leq 0.05$), ** ($0.001 <$ p-value $\leq 0.01$), ***( $0.0001 <$ p-value $\leq 0.001$), and **** (p-value $\leq 0.0001$).

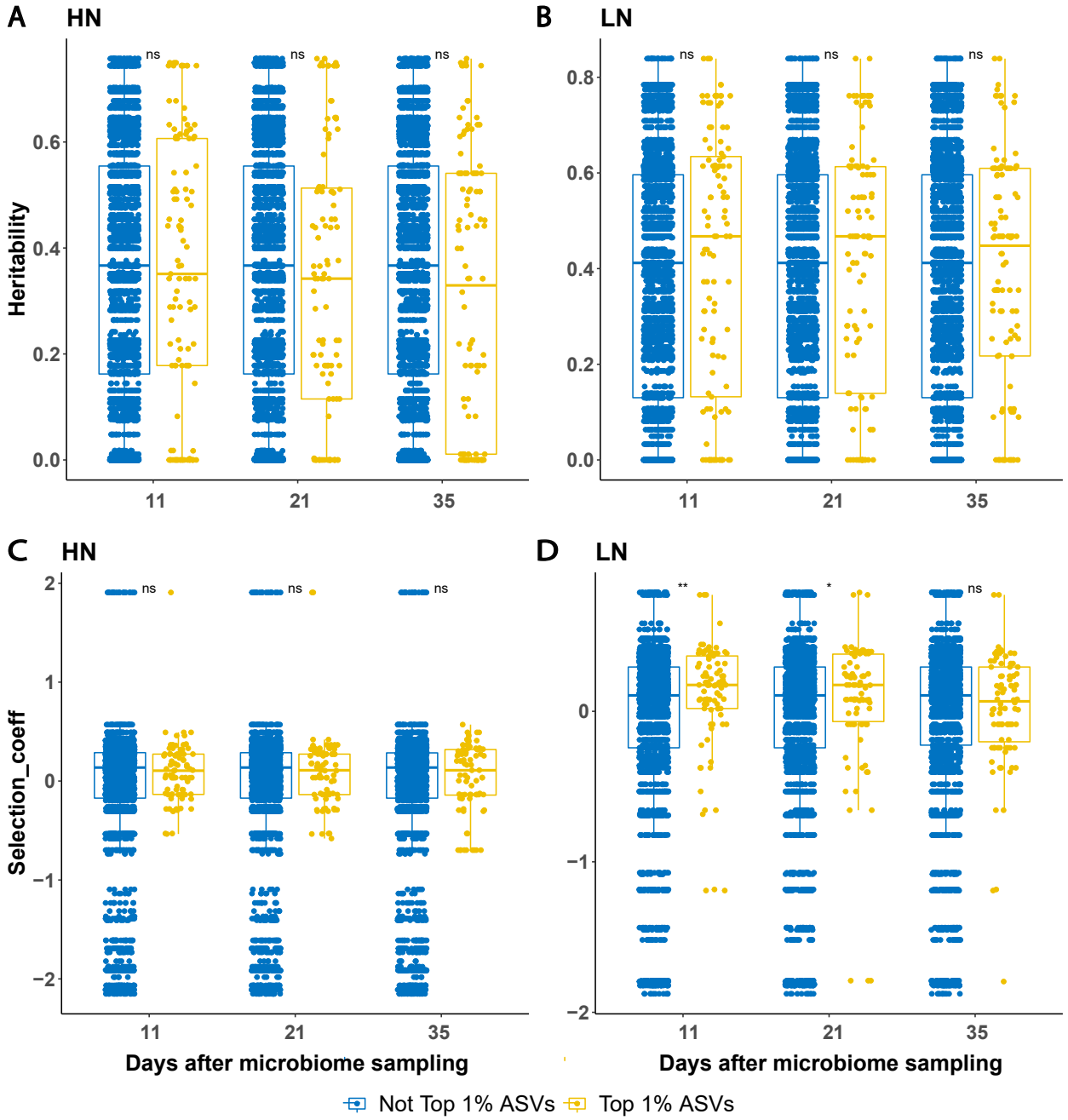


**Figure S3 Comparison between top 1% ASVs and the remaining ASVs for heritability and selection coefficient.** The Student's t-test results between top 1% ASVs (yellow) and the remaining 99% ASVs (blue) for heritability under HN (**A**) and LN (**B**), and selection coefficient under HN (**C**) and LN (**D**), respectively. The statistical significance of the difference of the Student's t-test result is indicated by asterisks: ns (not significant), * ($0.01 < \text{p-value} \leq 0.05$), ** ($0.001 < \text{p-value} \leq 0.01$), ***( $0.0001 < \text{p-value} \leq 0.001$), and **** ($\text{p-value} \leq 0.0001$).

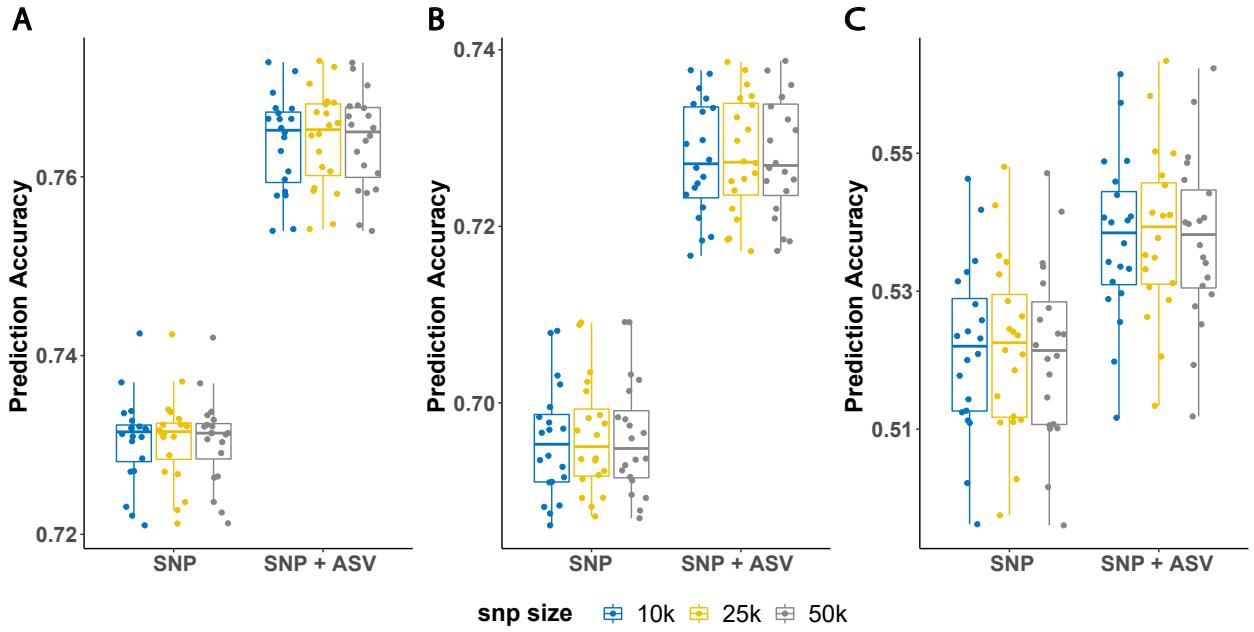


**Figure S4 Sensitivity test of the SNP size on prediction accuracy for VARI trait.** The Student's t-test results using randomly selected 10k SNPs (blue), 25k SNP (yellow), and 50k SNP (grey) in 11 (**A**), 21 (**B**) and 35 (**C**) days after microbiome sampling. The test results are not significant with p-values greater than 0.1.