Soil diversity metabarcoding from cacao crop wild relatives in a tropical biodiversity hot spot in Colombia

Luis Alberto Chica Cárdenas[(1,2)] (iD), **Monica Arias Peñaloza**[(1)] (iD), **Martha Lucía Cepeda Hernández**[(3,4)] (iD) **and Martha Josefina Vives Flórez**[(1)]* (iD)

[(1)] Universidad de los Andes, Centro de Investigaciones Microbiológicas CIMIC, Department of Biological Sciences, Bogotá D.C., Colombia.

[(2)] Universidad de los Andes, Max Planck Tandem Group in Computational Biology, Department of Biological Sciences, Bogotá, Colombia. Research Group on Computational Biology and Microbial Ecology, Department of Biological Sciences, Universidad de Los Andes, Bogotá, Colombia.

[(3)] Universidad de los Andes, School of Sciences, Bogotá D.C., Colombia.

[(4)] Corporación Corpogen, Bogotá D.C., Colombia.

ABSTRACT: *Theobroma cacao*, the source for chocolate fabrication, is a high-value crop that faces challenges such as the impact of climate change, pathogens, and cadmium accumulation. Soil associated with *T. cacao* has been extensively studied, looking for bio-controllers and microorganisms capable of Cd accumulation. However, there is no information about the microbial structure and interactions occurring in soil associated with cacao wild relatives, which represent a repository for biological diversity and its potential for biotechnological applications. We performed an extracellular DNA metabarcoding on soil samples associated with *Theobroma* spp. and *Herrania* sp. plants in two localities of the Biogeographic Chocó. We found microbial high diversity indexes and no correlation with plants or sampling locations. Potential cacao pathogens and bio-controllers and unexpected differences in the physicochemical soil properties from close locations were detected. Cadmium, an important economic factor for the commercialization of cacao products, showed significant differences between locations associated with a particular *Theobroma* species. We discuss some important relationships with soil physicochemical properties, the urgent need to complete the missing information on the diversity of bacterial, fungal and insect groups, and the potential of comprehensive analyses for decision-making regarding land-use and vocation. Also, we did not find the only *T. hylaeum* tree after three years, evincing the urgent need for conservation strategies in Colombia's Chocó region.

Keywords: cacao wild relatives, soil metabarcoding, eDNA, *Theobroma*, *Herrania*.

INTRODUCTION

Biological and genetic resources surrounding *Theobroma cacao* species and their crop wild relatives (CWR) are of great value for crop development in Colombia and the world. This species group may present genetic variations that confer beneficial agronomic characteristics or new flavor and aroma profiles. Variations in the traits hold an economic value potential for the chocolate industry, and, therefore, knowledge building on *Theobroma*'s native genetic material and associated biological diversity is important. Commercial cacao cultivation demands urgent advances in the production chain development. Given its use in illicit crop substitution programs in Colombia (cacaobp.org) and the consequent involvement of vulnerable communities, a comprehensive consideration of the factors affecting income generation possibilities is essential. Unfortunately, the information currently available on the *Theobroma* genus and its wild relatives is limited (Callejas, 2011).

Biogeographic Chocó is recognized as a hotspot of tropical biodiversity and a center of production of new species of *Theobroma* and *Herrania* (Richardson et al., 2015). For decades, the scientific community has experienced difficulties in accessing the region for study, which is subject to great pressures that lead to biodiversity loss. Deforestation rate from sea level to 1200 m a.s.l., the natural habitat of *Theobroma* species and their taxonomic relatives, is the highest in Colombia due to proximity to rural populations, growth of agriculture, and effects of the armed conflict (Dávalos et al. 2011).

The knowledge strategy of the national biodiversity policy (Colombian Ministry of Environment and Sustainable Development, 2012) establishes characterization of biodiversity components at ecosystem, population, species, and genetic levels is a fundamental part of our national policy. Under this framework, the CacaoBio Expedition brought together scientists from national and regional government agencies, universities, and local communities to determine the diversity of *Theobroma cacao*, its sister species in the wild in little-explored endemic areas, and its associated biota of plants, insects, and microbiota. As far as the latter is concerned, the associate microbiota may have beneficial functions for cacao's sustainable agriculture, and therefore, understanding the microbial communities linked to the CWR of cacao represents an initial stride toward the development of targeted microbiome interventions aimed at enhancing agricultural productivity and sustainability within cacao agroecosystems (Schmidt et al., 2022). Species associated with cacao can contribute to chocolate productivity and quality, making it vital to understand the ecosystem in which this species grows naturally.

Traditional study techniques such as captures and in-lab cultivation have limitations that are overcome by complementing the data with results obtained using molecular methods, which allow a closer approximation to the real composition of the associated microbiota. In agriculturally important plants, knowledge of soil biota has shown great potential for development of more sustainable agricultural technologies and practices, such as reducing diseases incidence (Andrews, 1992; Bloemberg and Lugtenberg, 2001), increasing agricultural production (Bakker et al., 2012), decreasing the use of chemical inputs (Adesemoye et al., 2009), and reducing greenhouse gas emissions (Singh et al., 2010). Some studies show the microbiome can have a direct effect on plant phenology, for example, on flowering time (Wagner et al., 2014; Panke-Buisse et al., 2015).

Among the most recent techniques for studying diversity in complex samples is the study of environmental DNA or extracellular DNA (eDNA, exDNA). In a comparative study, intracellular DNA (iDNA) was evaluated with extracellular DNA (exDNA), showing that some sequences found in the exDNA fraction are not found in the iDNA fraction (Nagler et al., 2018). Such DNA, which can persist in soil for extended periods, reflects historical biodiversity of the environment and can provide important information about past climatic and ecological conditions. Another study found that with exDNA analysis,

up to 55 % more information on observed prokaryote and fungal richness was retrieved compared to iDNA analysis (Nagler et al., 2018).

This paper presents the results obtained from the analysis of extracellular DNA metabarcoding for the two major biological groups present in the microbiota (bacteria, fungi) from soil samples associated with *Theobroma* spp. and *Herrania* sp. plants in two localities of the Biogeographic Chocó. While the soil microbial diversity of commercially grown cacao has been described, there is not much information available on microbial diversity in soils associated with plants of wild species that are phylogenetically close to cacao. This study represents a repository of genetic and functional information that may eventually help solve current problems associated with commercial cultivation. This study aimed to discuss some important relationships with soil physicochemical properties, the presence of potential crop pathogens, the urgent need to complete the missing information on diversity of the identified groups, and the potential of comprehensive analyses for decision-makers regarding land-use and vocation.

MATERIALS AND METHODS

Soil sampling

Twenty-five soil samples from cacao crop wild relatives (CCWR) of trees previously geo-referenced during the 2019 cacao-BIO expedition to the village La Victoria in the Department of Chocó, Colombia (González-Orozco et al., 2020) were collected between March and April of 2021 as follows: Samples from six *Theobroma glaucum* trees, six *Theobroma cacao* trees, eight *Theobroma simiarum* trees, and four *Herrania* cf. *purpurea* trees. A sample from one *Theobroma* cf. *hylaeum* tree was also obtained (Table 1). Trees were in two distinct zones of the La Victoria area namely, Baudó (west of the village towards the lower Baudó Range) and Atrato (east of the village towards the Atrato River basin).

For each tree, a surrounding circular plot of 1 m radius was established and samples (~250 g) of non-rhizosphere soil [horizon A (0.00-0.30 m)] from eight equidistant points were collected after careful removal of litter and organic layer. Eight subsamples were bulked into one homogenized composite sample per tree, stored into a sterile airtight plastic bag for shipment, and refrigerated for two weeks at 4 °C. Samples were frozen immediately upon return to the laboratory and kept at -20 °C until DNA was extracted. Subsamples of 500 g from each bulk sample were sent for physicochemical analyses. Soil sampling was authorized by the Colombian Authority for Environmental Licenses (ANLA) through Resolution 1177 (Collection of Specimens of Wild Species of Biological Diversity for Non-Commercial Scientific Research Purposes) granted to Universidad de los Andes in Bogotá, Colombia.

Table 1. Number of samples taken per tree species and per location

Tree species	Samples per species	Location	Samples per sampling site
Theobroma cacao	6	Baudó	5
		Atrato	1
Herrania purpurea	4	Baudó	3
		Atrato	1
T. glaucum	6	Baudó	2
		Atrato	4
T. simiarum	8	Baudó	7
		Atrato	1
T. hylaeum	1	Baudó	1

Soil physicochemical properties

Physicochemical analysis of the soil samples was performed by AGRILAB® environmental and agricultural services in Bogotá, Colombia, following standard methods (i.e., USDA Salinity Laboratory, NTC 5403 Walkley-Black, Bouyoucos, EPA 200.9, NTC 5349, NTC 5526 and NTC 5350) and using instrumental analysis to determine the following parameters: pH, electric conductivity, effective cation exchange capacity, medium humidity saturation, oxidizable organic carbon, organic matter, total nitrogen (N), texture, apparent density, total cadmium (Cd), potassium (K), calcium (Ca), exchangeable magnesium (Mg) and sodium (Na), iron (Fe), copper (Cu), zinc (Zn), boron (B), sulfur (S), phosphorus (P), and Ca/Mg, C/K, Mg/K ratios, and (Ca+Mg)/K ratio. All samples were processed in triplicate, and the values for each physicochemical factor represent the mean of these replicates.

Extracellular DNA (eDNA) extraction

Total eDNA was extracted from 15 g of each homogenized soil sample using the NucleoSpin Soil kit (Macherey-Nagel, Düren, Germany) following the protocol described in Taberlet et al. (2012) modified by Valencia et al. (2021). Concentration and purity of the eDNA were assessed using a Nanodrop spectrophotometer (Thermo Scientific, Willington, DE, USA) based on 260/280 and 260/230 nm absorbance radios and stored at -20 °C for further analyses. For samples with low eDNA concentration, a final Speed Vac for 15 min at 50 °C step was performed.

eDNA amplification and sequencing

Amplification reaction for sequencing of the V3-V4 regions of prokaryotes 16S rRNA genes (341F/806R) with an index for sample identification was performed in triplicate for each sample, using the primers described by Kozich et al. (2013). The 25 µl Polymerase Chain Reaction (PCR) mixture comprised five µl of HOT FIREPol Blend Master Mix (Solis Biodyne, Tartu, Estonia), 1 µL of each forward and reverse primer (10 µmol L^{-1}) and 2 µL of DNA. Thermal cycling included an initial denaturation at 95 °C for 10 min; 35 cycles of denaturation for 15 s at 95 °C, annealing for 30 s at 55 °C, elongation for 45 s at 72 °C; final elongation at 72 °C for 5 min.

Primers for the ITS1 (5.8S/ITS5) region described by White et al. (1990) and Epp et al. (2012) were used to create amplicon libraries using a two-round PCR. During the first PCR, partial Illumina (San Diego, CA, USA) TruSeq adapter sequences were added to the 5' tail of the primer and the products were used as a template for the second PCR to add the remaining adapter sequence and index for sample identification. For the first amplification, the 25 µL PCR mixture comprised five µL of HOT FIREPol Blend Master Mix (Solis Biodyne, Tartu, Estonia), one µL of each forward and reverse primer (10 µmol L^{-1}) and 2 µL of DNA. Thermal cycling included an initial denaturation at 95 °C for 10 min; 35 cycles of denaturation for 15 s at 95 °C, annealing for 30 s at 55 °C, elongation for 30 s at 72 °C; final elongation at 72 °C for 5 min. For the second amplification, the 25 µL PCR mixture comprised 5 µL of HOT FIREPol Blend Master Mix (Solis Biodyne, Tartu, Estonia), 1 µL of each forward and reverse primer (10 µmol L^{-1}) with the unique index sequence per sample and 4 µL of the products from the first PCR. Thermal cycling included an initial denaturation at 95 °C for 10 min; eight cycles of denaturation for 15 s at 95 °C, annealing for 30 s at 55 °C, elongation for 30 s at 72 °C; final elongation at 72 °C for 5 min. The same protocol was used for 18S metabarcodes. However, although amplification was achieved, an incomplete taxonomic assignment did not allow any significant analyses. Hence, the results are not included.

Random PCR products (5 µL) from all the amplifications were verified using 1.5 % agarose gel. All the PCR products were purified with AMPure XP beads (Beckman Coulter, Atlanta, GA, USA) eluted five-fold in ultrapure water and quantified using the QuBit dsDNA HS Assay kit and the QuBit 2.0 fluorometer following the manufacturer's instructions (Life

Technologies, Grand Island, NY, USA). DNA quality was examined using Bioptic Qsep 100 DNA fragment analyzer (BiOptic Inc., New Taipei City, Taiwan). The pooled, quantified libraries were adjusted to 4 nmol L^{-1} and denatured following the Illumina MiSeq library denaturation and dilution guide. To improve clustering during initial sequencing, the denatured libraries (8 pmol L^{-1}) were mixed with 20 % PhiX genomic control. Library preparation and sequencing on the Illumina MiSeq platform (2 × 250 bp) using the reagent kit v2 was performed at Corporación CorpoGen (Bogotá, Colombia). Negative controls were conducted for all PCR reactions.

Processing Illumina sequencing data

Quality control of the raw reds was conducted using fastqc v0.11.9. Adapters and low-quality reads were removed (SLIDINGWINDOW:4:20, MINLEN:100, HEADCROP:10) using Trimmomatic v0.39 (Bolger et al., 2014) in single-end mode. Sequences from both amplicons (16S, ITS) were independently imported to qiime2 (Bolyen et al., 2019) and denoised using DADA2 (Callahan et al., 2016). A truncation length of 220bp was used for the 16S reads, whereas for ITS reads, the truncation length was set to 180bp. The resulting amplicon sequence variants (ASV) with a frequency lower than 10 were removed from the analysis.

Taxonomic assignment of the ASVs was made using a naïve bayes classifier, as implemented in Qiime2, and trained with Silva v138 (Quast et al., 2012) and unite v8.3 (Nilsson et al., 2019) databases for 16S and ITS, respectively. Taxonomic classification was assessed at Phylum level and at the maximum taxonomic level reached by each ASV. For this last approach, taxonomic assignments were manually curated by removing ambiguous classification (c__uncultured, o__uncultured, f__uncultured, g__uncultured, s__metagenome, s__bacterium_enrichment, s__unidentified). Subsequently, the deepest informative taxonomic level of each ASV was used.

The ITS ASVs were also classified using a second taxonomic classifier, trained with a custom database encompassing ITS sequences from the genera *Albonectria*, *Armilaria*, *Ceratabasidium*, *Ceratosystis*, *Colletotrichum*, *Diploidia*, *Lasiodiploidia*, *Phellinus*, *Pythoftora*, *Rigidoporus* and *Roselina*, considered fungal pathogens. To build the custom database, sequences were downloaded from NCBI nucleotide database. This was followed by a cleaning process in which sequences with more than 5 degenerated base pairs and homopolymers of more than 12 bp were removed. A sequence length filter was also applied using a minimum length of 150 bp and a maximum length of 2000 bp. Finally, sequences were de-replicated before training the taxonomic classifier. Curation process of the database described above was performed using the Rescript plugin for Qiime2 (Robeson et al., 2021).

Alpha and Beta diversity estimates were computed using Vegan v2.5.7 (Oksanen et al., 2013). For alfa diversity, the Shannon index was used as metric, whereas for beta diversity, a non-metric multidimensional scaling (NMDS) was computed using a Bray-Curtis dissimilarity matrix. Physicochemical factors were fitted on the ordination using the envfit function. Biplots were plotted using the five physicochemical properties with higher correlation values with any of the axes of the NMDS and p-values lower than 0.05. Distance net between species was calculated after extracting the pairwise distances between the samples of each pair of tree species from the Bray-Curtis dissimilarity matrices (16S and ITS) and computing the median of the distances.

Statistical analyses

Differences in physicochemical factors between the two sampling locations were tested using a Wilcoxon test. Alpha diversity significances were evaluated by location using a Wilcoxon test and by tree species using an Anova test (p-value<0.05).

To examine the ASVs with significant differences in their abundance profiles between the 2 sampling zones used in this study (Baudó and Atrato), a differential abundance test using Deseq2 v1.28.1 (Love et al., 2014) was conducted. Results were filtered using a log2 fold change threshold of 2 and an adjusted p-value of 0.05. Afterwards, the abundance profiles of the important taxa given by Deseq2 were correlated with information on the physicochemical factors using the cor_mat function from rstatix package.

Sequence data accession numbers

Raw sequences for 16S and ITS amplicons were submitted to the European National Archive (ENA) under the project accession number PRJEB54958.

RESULTS

Samples

Tweny five soil samples belonging to soils associated with five species from the Tribu *Theobromeae* were collected in two locations in Chocó, Colombia (Table 1). These soils are classified as Humic Dystrudepts and Typic Eutrudepts, characterized by moderately rugged and steep terrain, slopes ranging from 50-75 %, loamy-sandy texture, moderately deep soil, and well-drained conditions (Colombian Ministry of Environment and Sustainable Development, 2016). Geographical references for each tree sampled are provided in the supplementary material. Since only one sample of *T. hylaeum* - a wild species of the Tribu *Theobromeae* - was collected, this was discarded from the statistical analysis.

Physicochemical analysis

By computing the results of the physicochemical analysis of the samples grouped by location, of the 33 factors considered, ten were found to be significantly different (Wilcoxon<0.05). Of the factors with significant differences, in 9 cases, the values were higher in the Baudó area, and only the aluminum saturation showed higher values for the Atrato (Figure 1).

Metabarcoding - based taxonomic assignment

Bacterial diversity and abundance based on 16S analysis

16S analysis at phylum level showed a very similar pattern in soil samples from different trees and geographical zones, in which *Acidobacteriota* was found to be the most prevalent phylum (Figure 2). A consistent and important abundance of *Proteobacteria* and *Verrucomicrobia*, regardless of the geographical zone or tree of the cacao group from which the sample was taken, was also found. On the other hand, the abundance of *Bacteroidota* was not constant across the samples. No specific pattern was found, suggesting members of this phylum have a per-sample instability not associated with the assessed variables. It is also important to mention that even in low abundance, members of the phylum *Thermoplasmatota,* a group characterized for growing in very low pH conditions, were found in most of the samples.

In assessing abundance using ASVs at the highest taxonomic rank, several ASVs belonging to *Acidobacteria* group were found in most samples. However, no ASV was found to have a specific pattern across a determined group of samples. In ~50 % of the samples, the 20 most representative ASVs accounted for at least 10 % of the total abundance of the samples. High diversity and richness of soil suggest that this small portion of ASVs might be very important to soil dynamics.
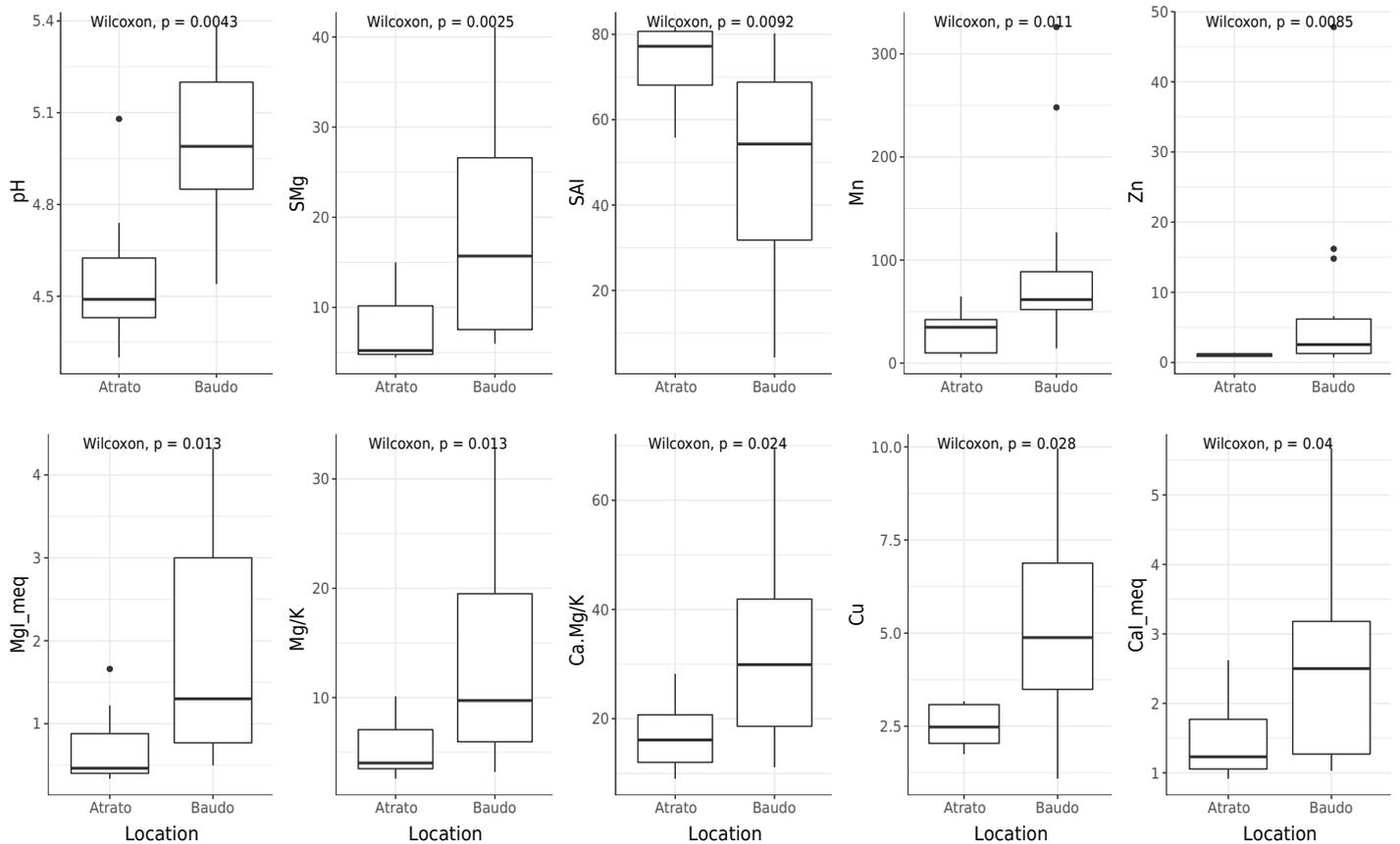
Figure 1. Top 10 factors that showed significant differences (Wilcoxon<0.05) between sampling locations. While pH, magnesium saturation (SMg), manganese (Mn), zinc (Zn), interchangeable magnesium (MgI_meq), magnesium/potassium (Mg/K), calcium+magnesium/potassium (Ca.Mg/K), copper (Cu), and interchangeable calcium (CaI_meq) relations were significantly higher in Baudó, aluminum saturation (SAl) was the only factor that was lower in the Baudó area).

Fungal diversity and abundance based on ITS analysis

The ITS assignment showed that phyla *Ascomycota*, *Mortierellomycota* and *Basidiomycota* comprised more than 98 % of the abundance in all samples. However, the abundance patterns of this phyla were found to be sample-specific (Figure 3).

The ASV analysis for ITS shows *Candida hyderabandecis* as a dominant ASV on most of the samples of *T. glaucum* and ~33 % of samples of *T. cacao*. *T. simiarum* samples were found to have a more diverse composition with a slight dominance of *Trichoderma* ASVs. Predominance of single ITS ASVs in the samples is more notorious than with 16S analysis. Multiple samples have more than 50 % of the total abundance represented by the 20 more abundant ASVs; however, the taxonomic resolution of these ASVs does not allow proper classification.

ITS detection of fungal pathogens

To discover whether the common fungal pathogens for *Theobroma* species were represented, full ITS sequences were downloaded from NCBI and used to train a naïve bayes classifier that was subsequently used to scan all ASVs sequences. By counting the number of ASVs that were classified using the pathogen database and comparing their taxonomic classification against the classification given by Unite database, we found that most of these ASVs were poorly classified taxonomically when using Unite database, whereas in the pathogen database a significant amount of these ASVs were classified to genus or species level. Many of the ASVs classified by the pathogen databases were assigned as *Fusarium* species, and their contribution to the total abundance of each sample suggests that most ASVs classified only at Phylum or order with the Unite
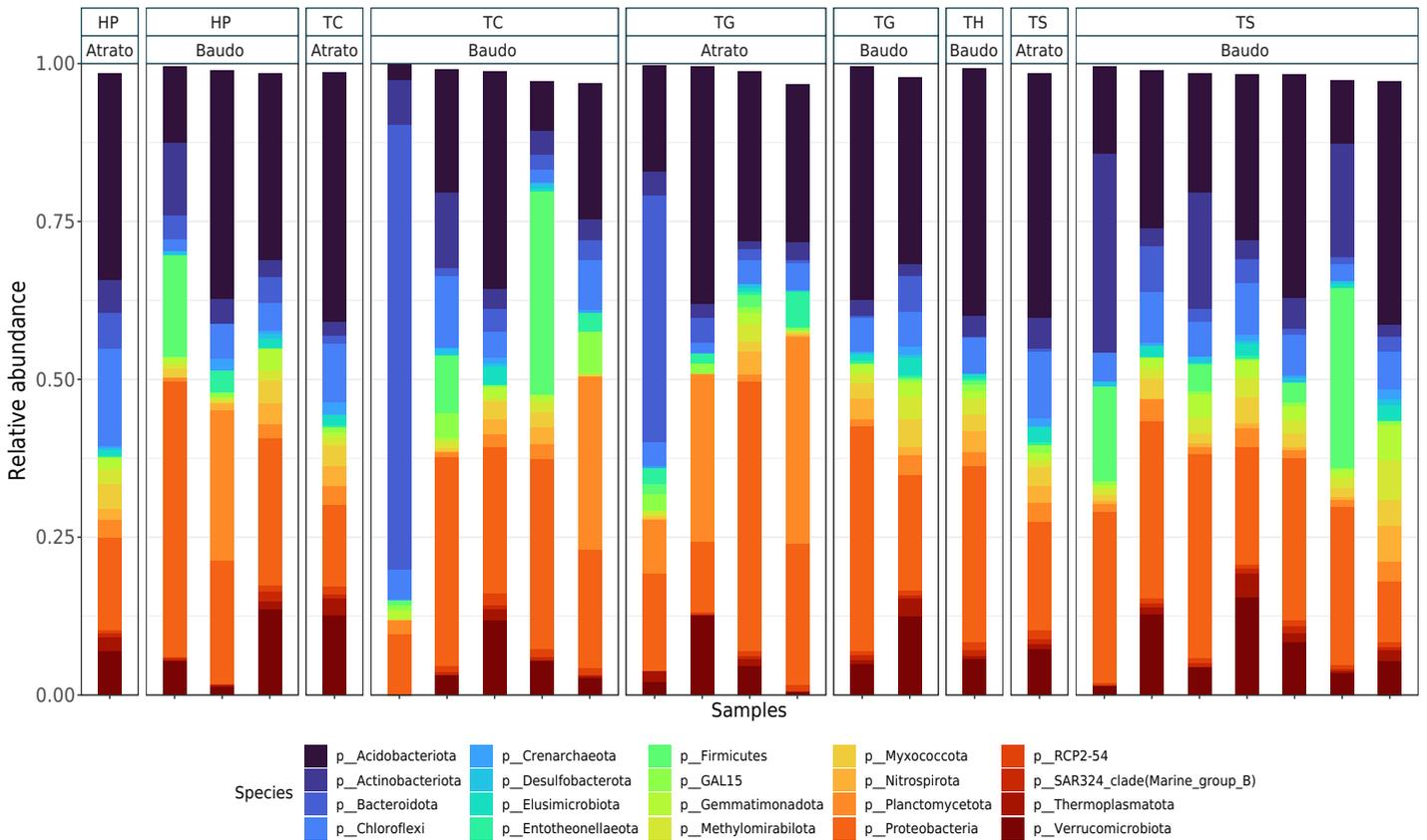
Figure 2. 16S taxonomic plots per sample using Silva database at phylum level. Plots were faceted by the tree species and the sampling location. The 20 taxa with the highest average abundance across samples were plotted. TC: *T. cacao*; HP: *H. purpurea*; TG: *T. glaucum*; TS: *T. simearum*; and TH: *T. hylaeum*.

database are *Fusarium* species. Genera *Colletotrichum* and *Diploidia* were also found to have a high number of ASVs assigned to them. In spite of the high representation of common *T. cacao* pathogens such as *Phytophthora* on the database, no ASVs were assigned to these pathogens.

Given the lack of resolution yielded by the primers targeting the V9-18S region to identify insect composition, in which 80 % of the reads were assigned to different orders, and the heterogeneity of the number of sequences per sample (max: 32707, min: 2435), this dataset was not used for further diversity analysis.

Alpha and beta diversity analysis

Alpha diversity analysis for the 16S dataset showed no significant differences when comparing samples from different trees or between sampling locations (Figures 4a and 4b). When exploring fungal diversity, we found the same pattern showed with the 16S (Figures 4c and 4d). These results suggest that the environmental or physicochemical changes in geographical zones do not affect bacterial or fungal diversity.

Beta diversity analysis showed no pattern of clustering within the samples. However, for 16S, two big groups of samples were found for which Cd, Al saturation, and interchangeable acidity were found to be drivers of the top group, whereas Ca saturation and pH were found to be associated with the bottom group (Figure 5a). For ITS, NMDS1 separates Atrato and Baudó samples. Magnesium and Al saturation may drive this separation (Figure 5b).

When computing the median of distances between samples of the same species and between species using the Bray-Curtis dissimilarity matrices, greater distances were found
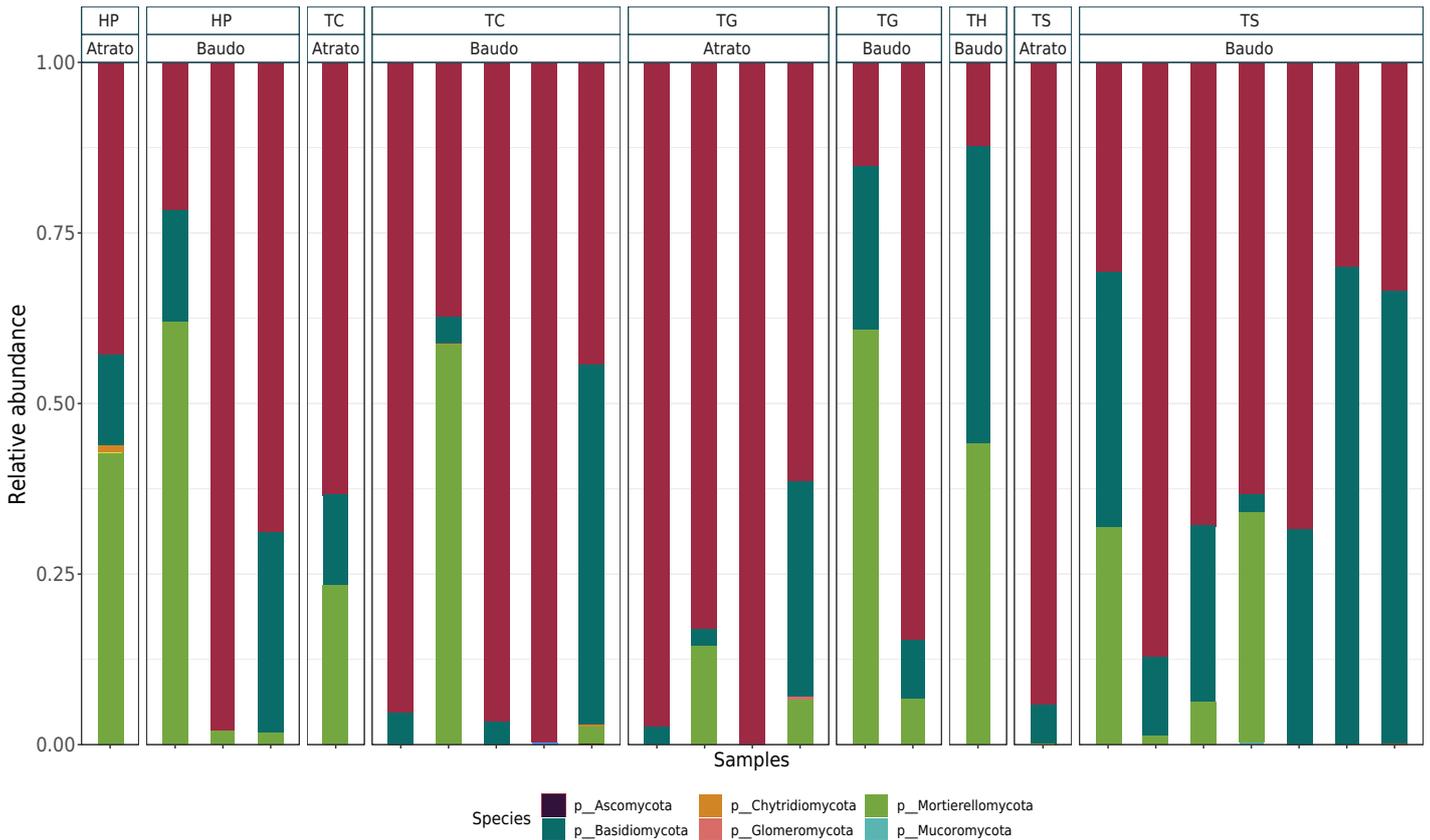
Figure 3. The ITS taxonomic plots using Unite ITS database at phylum level. Plots were divided by the sampling variables that were considered. The 20 taxa with the highest average abundance across samples were plotted. TC: *T. cacao*; HP: *H. purpurea*; TG: *T. glaucum*; TS: *T. simearum*; and TH: *T. hylaeum*.

between samples of *T. cacao* than between *T. cacao* and other species. This pattern was seen for 16S (Figure 6a) and for ITS data (Figure 6b). When comparing in-between species, 16S data showed *H. purpurea* and *T. simiarum* to be the closest species, whereas for ITS, *T. cacao* and *T. glaucum* were the species with the shortest distance between them.

Correlation to Cd contents

Since Cd content was highly correlated with ordination in NMDS, a regression analysis was performed using Cd content and the first NMDS axis to determine whether samples could be separated according to this parameter. When Cd content was correlated to the first NMDS axis, a significant correlation was observed, suggesting it has an influence on sample dispersion within the ordination, but no cluster per sample was found. When the model was run for each species, the fit was higher, but the correlation was significant for only one species: *T. glaucum.* This result suggests that further research is necessary to unravel the dynamics between this species, the associated microorganisms, and Cd contents.

Differential abundance analysis

A differential abundance test was performed using Deseq to identify the taxa that were differentially abundant in Atrato samples when compared to Baudó. When collapsing at phylum level, six phyla (*Sphirochateota, Desulfobacteriota, GAL15, MBNT15, Latescibacteriota* and Firmicutes) exclusively showed a positive fold change, indicating a significantly higher abundance in Baudó soils. On the other hand, four phyla (*Cyanobacteria, Entoheneleaota, Entotheonellaeota*, *Bdellovibrionota, Patescibacteria*) were found to have exclusively negative fold change values, indicating a significant lower abundance of these
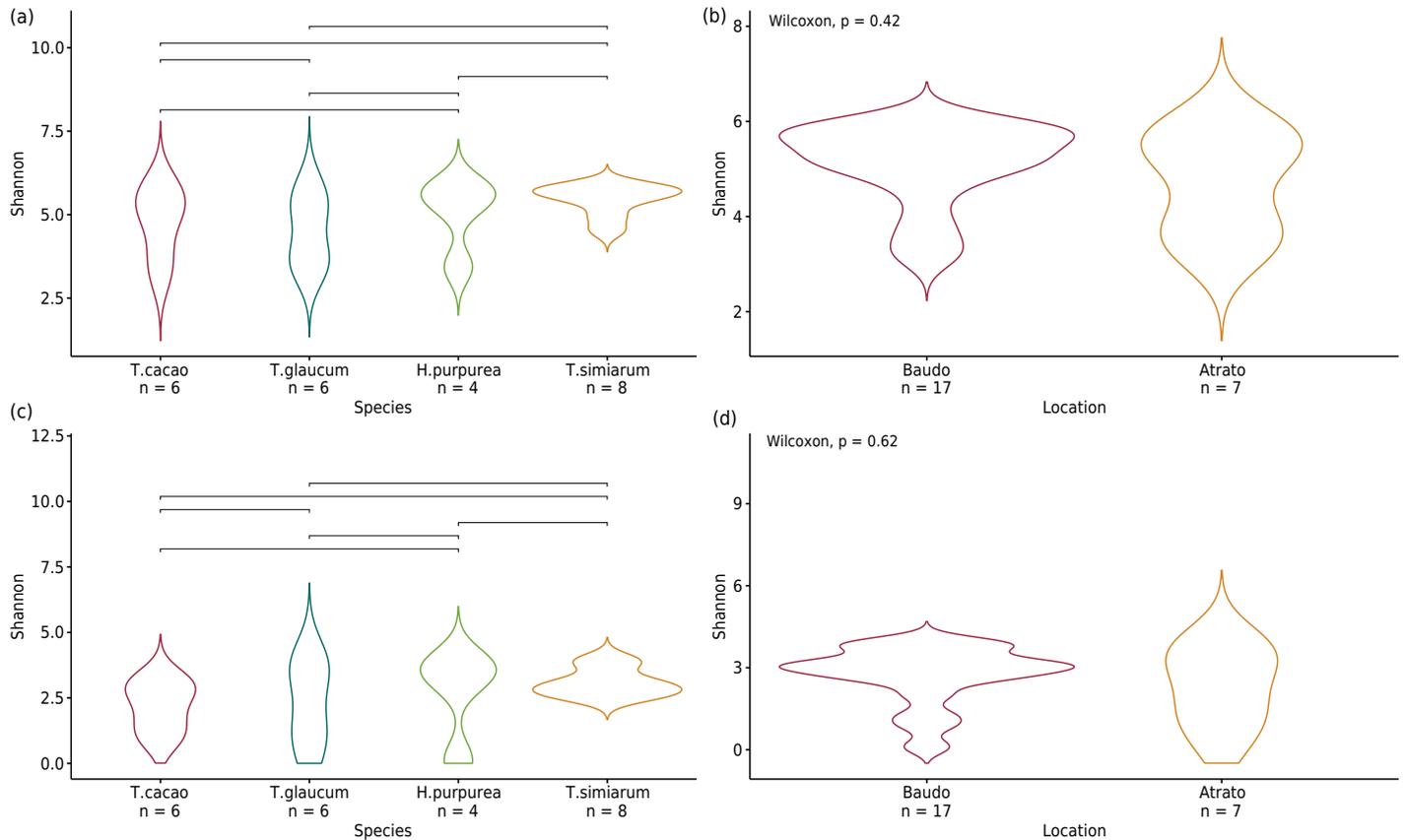
Figure 4. Alfa Diversity differences for 16S and ITS using the species of cacao trees (a and c, respectively) and the geographical zone (b and d, respectively). Wilcoxon test with adjusted p-values was used to compare each pair of tree species statistically.

in Baudó soils. Even though, phyla *Bacteroidota, Actinobacteriota, Methylomirabilota, Proteobacteria,* and *Gemmatimonadota* have differentially abundant members in both areas, their log2 fold change medians suggest these might have a greater influence in Atrato. For the ITS fungal dataset, only ASVs belonging to three phyla were found to be differentially abundant, and most of these were found to be higher in the Atrato area.

To explore whether these phyla can be associated with physicochemical properties that were shown to be different between sample zones, a correlation analysis was performed using the bacterial abundance at the phylum level (Figure 7). Interestingly, *Firmicutes* showed strong positives and significant correlations with pH, Zn, and Mg and a completely opposite pattern with Al saturation. Actinobacteria group also showed high correlation values with pH and zinc content.

DISCUSSION

Colombian Pacific has been reported as a cacao CWR biodiversity hotspot (CCWR), with up to 22 of the 26 reported wild cacao taxa identified in previous expeditions (González-Orozco et al., 2020). This plant diversity may be related to diversity in soil microbiota. However, despite obtaining a general inventory of the region's microbiota, this study did not reveal soil-associated patterns of abundance or diversity of the different CCWR species, nor were there any evident differences between zones. Nevertheless, soil physicochemical properties in the studied areas revealed differences between the Atrato and Baudó zones. In the Atrato zone, Al saturation values were significantly higher than those found in the Baudó zone, while the other variables analyzed were significantly lower in Baudó. Differences were found in such a small area (1.5 km^2), which could be
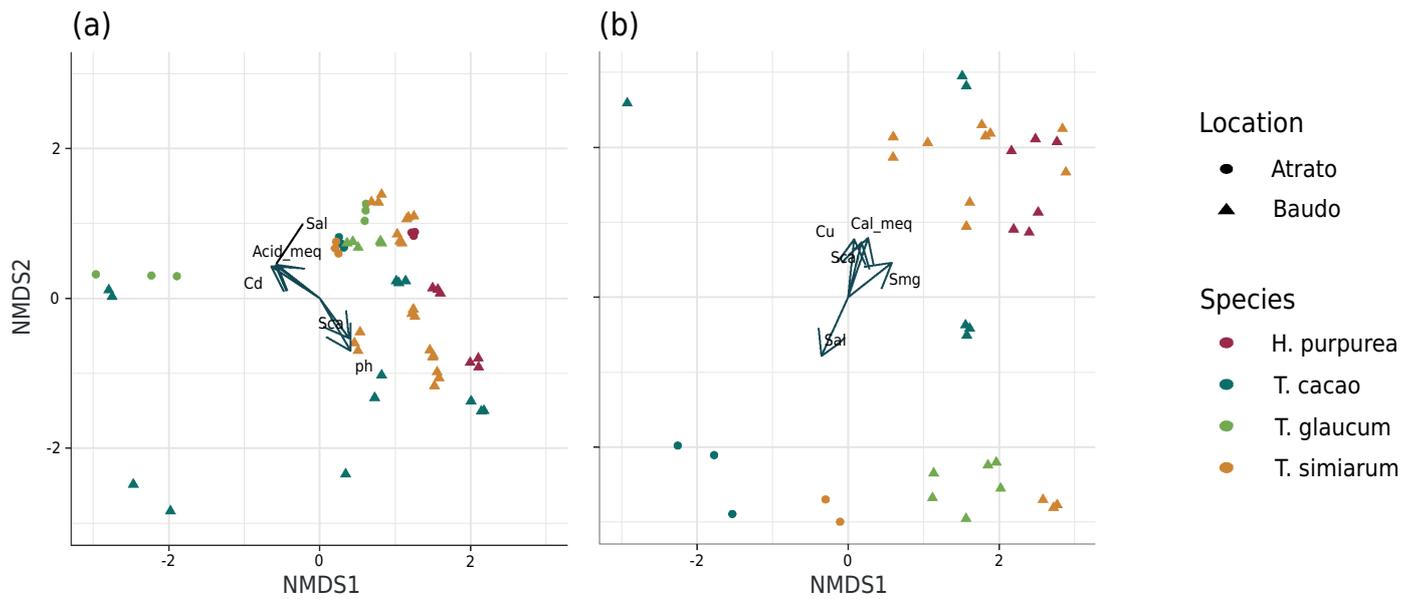
Figure 5. Beta Diversity ordination plots for 16S (a) and ITS (b) datasets. The 5 physicochemical properties with higher correlation coefficients were plotted as vectors. Multidimensional scaling plots were drawn up using a Bray-Curtis dissimilarity matrix, and the physicochemical properties were fitted to the resulting ordination.

attributed to the variety of anthropogenic activities and land-uses affecting biotic and abiotic soil variables, although no official source of information that discriminates these activities was found. This information is fundamental for an in-depth understanding of the edaphic biodiversity associated with CCWRs. Our study presents the generalities found through soil metabarcoding, allowing an initial approximation to the composition and structure of the microbial communities associated with CCWRs. The greater distance found in the microbiota within *T. cacao* samples is equally interesting. This could be an indicator of the diverse cultural practices related to the crop, but in-depth community research is needed to reveal the causes.

The most prevalent phyla *Acidobacteriota* and *Verrucomicrobia* reported in this study were found to be abundant in the soil and rhizosphere. However, these are poorly studied phyla as they are difficult to isolate and cultivate in-lab (Hackl et al., 2004; Janssen, 2006; Bergmann et al., 2011; Tanaka et al., 2017; Kalam et al., 2020). They are physiologically diverse and play an important role in the biogeochemical carbon, nitrogen, and sulfur cycles, as well as the production of exopolysaccharides and other compounds beneficial for plant growth (Navarrete et al., 2015; Kalam et al., 2020). A lesser abundance of the phylum *archaea Thermoplasmatota* was present in almost all samples. Genome studies of some representatives of these methanogens reveal metabolic activities important for the turn-over of protein residues and the conversion of methane or ammonia into oxidized forms available for other organisms, highlighting their role in biogeochemical carbon and nitrogen cycles.

Further studies are needed to determine their role in the different ecosystems they occur (Diamond et al., 2022). Specifically, for cacao, knowledge of the soil-associated microbiota of its CRW can support the search for new species of microorganisms useful for the crop. For example, the widespread presence of *Streptomicetae* in this study was to be expected given their abundance in the soil, where they play an important role as biocontrol and plant growth-promoting microbes given their ability to produce a wide variety of bioactive compounds, antibiotics, and hydrolytic exoenzymes. This is the case for *Streptomyces cameroonensis*, a new species promoting *Theobroma cacao* growth (Boudjeko et al., 2017).
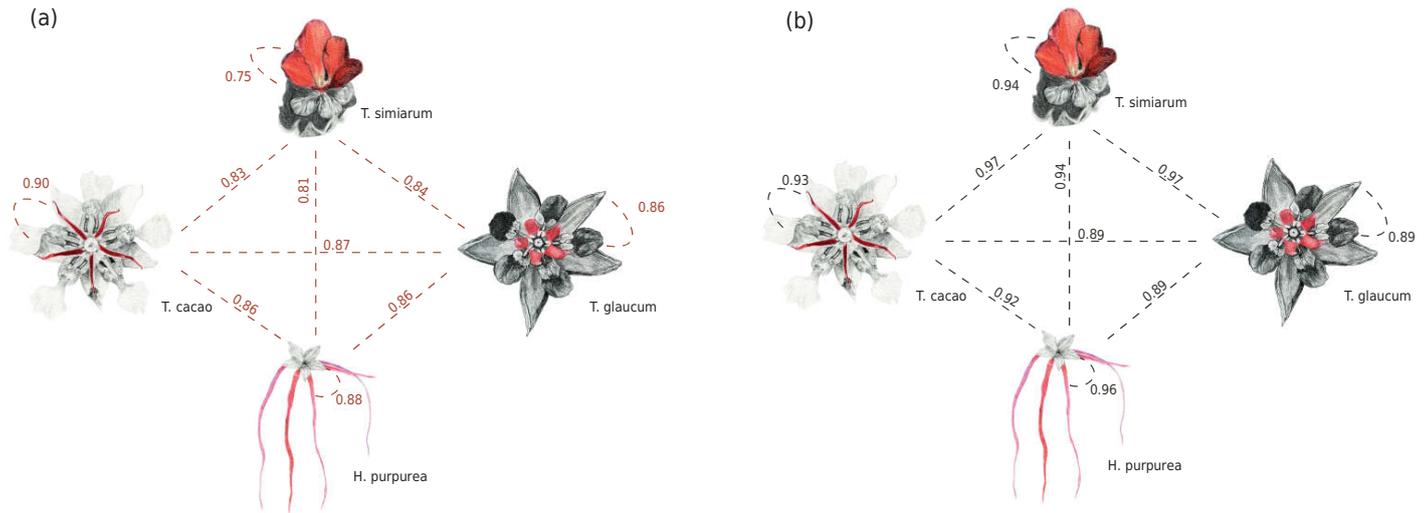
Figure 6. Distance net between plant species using (a) 16S and (b) ITS distance matrices. Distance between samples belonging to the same and different plant species was computed. Drawings by Sofia Echeverri.

In terms of fungi, it is important to mention that the taxonomic resolution did not extend to extensively analyzing such diverse phyla as *Ascomycota, Mortierellomycota* and *Basidiomycota*, highlighting the need to explore new and better metabarcoding markers for fungal identification and to strengthen reference databases for fungi (Orgiazzi et al., 2015; Estensmo et al., 2021). This is supported by a comparative analysis of ASVs with a database of specific ITS sequences of the most important cacao pathogens, where it was possible to classify some ASVs taxonomically, even at the species level. For example, approximately 10 ASVs were classified within the genus *Fusarium* when using the UNITE database, while when comparing the sequences with the database created for pathogens, about 70 ASVs matched representatives of this genus. Species of the genus *Fusarium* have been reported to cause Cushion Gall disease on young and mature cacao trees in Sri Lanka, West Africa, Colombia, Costa Rica, and Nicaragua (Hansen, 1966; Ploetz, 2006).

Species belonging to the genera *Colletotrichum* and *Diplodia*, which cause Anthracnose and Diplodia in *Theobroma cacao*, respectively, were also detected in this study. These diseases have been little investigated since their incidence is low in cacao crops (Delgado-Ospina et al., 2021). *Fusarium, Colletotrichum* and *Diplodia* are part of the *Theobroma cacao* microbiota and have been reported as endophytes (Delgado-Ospina et al., 2021). Their presence in CCWR can be related to both pathogenicity and biocontrol by competition. Thus, it is of great importance to isolate strains of these and other fungal species that may be related to CCWR, to study their ecological relationships and their potential as biocontrollers of other cacao diseases, such as black pod, caused by *Phytophthora palmivora*, and moniliasis, caused by *Moniliophthora roreri*, both of which have significant economic consequences (ten Hoopen et al., 2003; Adebola and Amadi, 2010). It should be noted that isolates identified within the *Moniliophthora* and *Fusarium* genera were detected in studies of cultivable microorganisms associated with *Theobroma cacao* plants in the same region (Cano, 2023).

Regarding the metabarcoding analysis of eukaryotic communities, the limitations of this approach were particularly evident in this study. In spite of the availability of sequencing protocols and bioinformatic pipelines for biodiversity assessment, taxonomic assignment for obtained ASVs remains challenging. Reference databases are incomplete and often tailored for certain taxonomic groups, and the use of only the 18S biomarker was not enough to
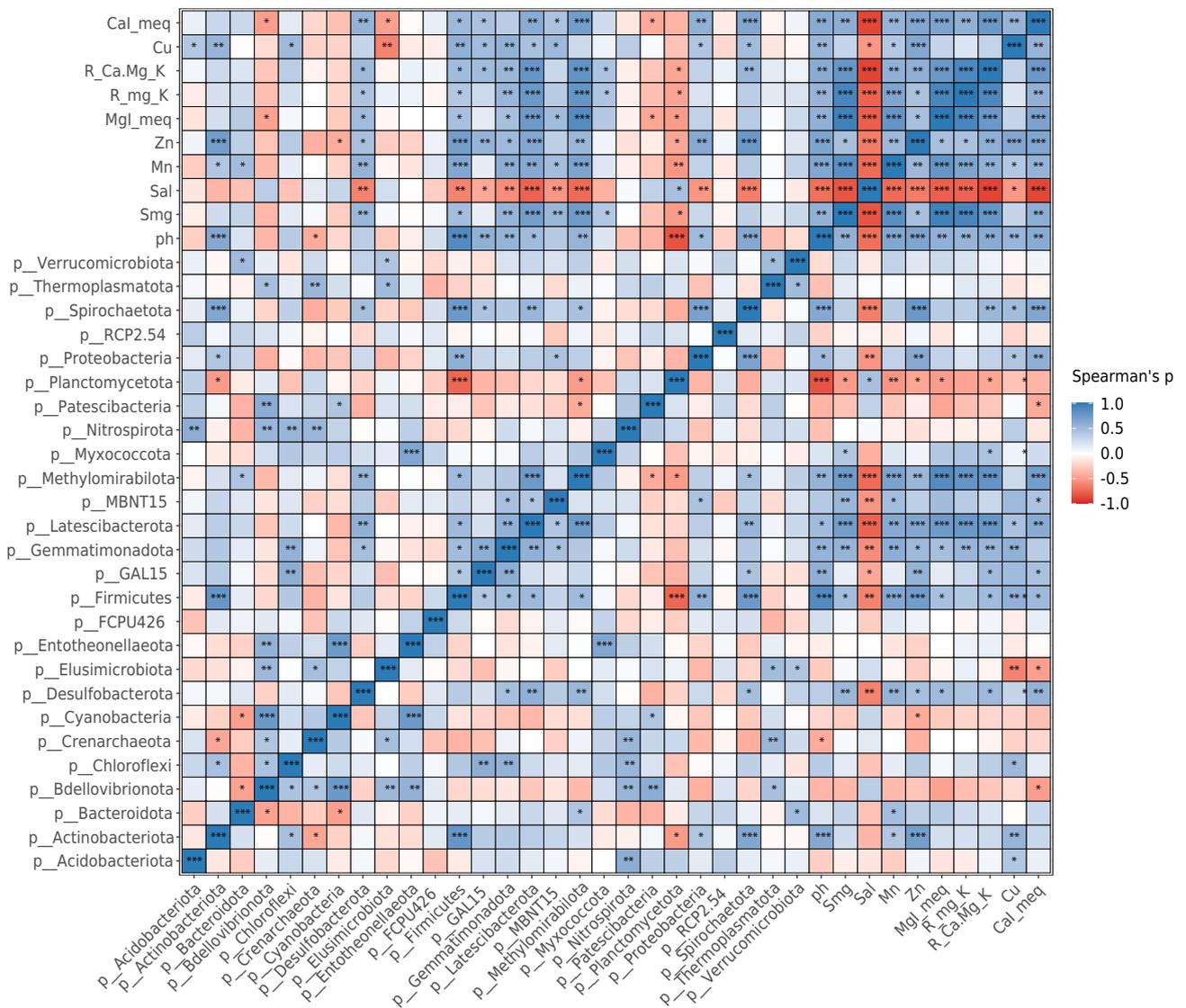
Figure 7. Correlation plot of the bacterial abundance at phylum level and the values of the most important physicochemical factors, given the correlation with both axes on the ordination plots. Correlation was made using the Spearman's test, and the significance levels are labeled with stars.

enable the detection of a broader range of taxa. To obtain valuable taxonomic data from CCWR and other crop wild relatives from the use of metabarcoding, a multidisciplinary effort that involves experts, researchers, bioinformaticians in the database building process is necessary. For example, a specialized database with specific biomarkers for insects and other arthropods can help to identify new potential pollinators, pests or biocontrolers for cacao crops.

β-diversity analysis determined for the bacterial communities revealed that Cd, Al saturation, and exchangeable acidity appear to be clustering drivers for some of the samples, while other samples tend to cluster by the effect of Ca saturation and pH. No clustering associated with CCWR species or localities was found. It is noteworthy that Cd content in the soil and its translocation to the fruit are variables that significantly affect cacao quality, have serious effects on health, and impact its commercialization (Maddela et al., 2020). Although in recent years, exhaustive studies have been conducted on

Cd-tolerant bacteria and their potential application as soil amendment in commercial cacao crops (Bravo et al., 2018, 2021; Bravo and Braissant, 2022), studies on the identification of Cd-tolerant microorganisms have not been conducted for CCWR. Cadmiun dynamics in regions such as the Pacific have not been investigated either because they are regions with no agricultural vocation or because they are not classified as suitable for cultivation. However, the microbial diversity present in soils associated with CCWR can be a source not only for isolating microorganisms useful in Cd bioremediation, but also for elucidating the relationships between CCWR and soil microorganisms in relation to this variable. This study not only suggests that the microbiota associated with *Teobroma glaucum* may be of interest in the search for Cd-tolerant strains, but also reveals that further studies on the dynamics between soil parameters and Cd metabolism in *T. glaucum* are needed. Understanding the role of genes related to Cd uptake or translocation in this CCWR would be valuable to elucidate the molecular mechanisms contributing to Cd tolerance in cacao plants, with particular attention to decreasing Cd accumulation in edible parts such as pods and beans.

CONCLUSION

This study demonstrates the importance of studying CCWR microbiota and, therefore, constitutes an additional reason to conserve areas that host high diversities of these species. Anthropogenic activities that are currently taking place in areas previously inaccessible due to armed conflicts in areas such as Chocó - Colombia, related to deforestation for road construction, illegal mining activities, and the establishment of crops with no technical assistance, among others, jeopardize the conservation of areas where a high diversity of CWR of economically important crops such as cacao has been reported, losing valuable information that can be useful in understanding the biodiversity for crop improvement programs. As an example of the urgent need to preserve the biodiversity of this region, we present the following experience: in May 2022, we re-visit the location only *T. hylaeum* tree found during the expedition; our intention was to obtain a picture of the flowers since, to the best of our knowledge, no register of the fresh, flowering structure is available for this species. The tree was no longer present at the geographical localization.

Further studies are essential to understand this ecosystem and should be integrated with activities that involve the local community in a participatory action research model that allows them not only to gain in-depth knowledge of the biodiversity in the area, but also to become aware of the importance of the CRWs for their conservation and sustainable use. Studies of soil diversity in areas such as La Victoria can reveal how CWR diversity and soil metabolic potential can improve crop soils' quality and health, especially by integrating microorganisms and some of their metabolic functions in strategies that contribute to nutrient cycling. Nutrient cycling, in turn, stimulates plant growth and production, bioremediation of soils contaminated with heavy metals, and biocontrol of pests that affect crops. This integrated approach will help to reduce the use of fertilizers and pesticides, and reverse soil degradation.

APPENDIX A. SUPPLEMENTARY DATA

Supplementary data to this article can be found online at https://www.rbcsjournal.org/wp-content/uploads/articles_xml/1806-9657-rbcs-48-e0230069/1806-9657-rbcs-48-e0230069-suppl01.pdf

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

Conceptualization: ⓘ Martha Josefina Vives Flórez (supporting) and ⓘ Martha Lucía Cepeda Hernández (lead).

Data curation: ⓘ Luis Alberto Chica Cárdenas (lead).

Formal analysis: ⓘ Luis Alberto Chica Cárdenas (lead).

Funding acquisition: ⓘ Martha Josefina Vives Flórez (equal) and ⓘ Martha Lucía Cepeda Hernández (equal).

Investigation: ⓘ Monica Arias Peñaloza (lead).

Methodology: ⓘ Monica Arias Peñaloza (lead).

Project administration: ⓘ Martha Josefina Vives Flórez (equal) and ⓘ Martha Lucía Cepeda Hernández (equal).

Software: ⓘ Luis Alberto Chica Cárdenas (lead).

Supervision: ⓘ Martha Josefina Vives Flórez (lead) and ⓘ Martha Lucía Cepeda Hernández (supporting).

Visualization: ⓘ Luis Alberto Chica Cárdenas (lead).

Writing – original draft: ⓘ Luis Alberto Chica Cárdenas (equal), ⓘ Martha Josefina Vives Flórez (equal), ⓘ Martha Lucía Cepeda Hernández (equal) and ⓘ Monica Arias Peñaloza (equal).

Writing – review & editing: ⓘ Luis Alberto Chica Cárdenas (equal), ⓘ Martha Josefina Vives Flórez (equal), ⓘ Martha Lucía Cepeda Hernández (equal) and ⓘ Monica Arias Peñaloza (equal).

REFERENCES

Adebola MO, Amadi JE. Screening three *Aspergillus* species for antagonistic activities against the cocoa black pod organism (*Phytophthora palmivora*). Agric Biol J N Am. 2010;1:362-5. https://doi.org/10.5251/abjna.2010.1.3.362.365

Adesemoye AO, Torbert HA, Kloepper JW. Plant growth-promoting rhizobacteria allow reduced application rates of chemical fertilizers. Microb Ecol. 2009;58:921-9. https://doi.org/10.1007/s00248-009-9531-y

Andrews JH. Biological control in the phyllosphere. Annu Rev Phytopathol. 1992;30:603-35. https://doi.org/10.1146/annurev.py.30.090192.003131

Bakker MG, Manter DK, Sheflin AM, Weir TL, Vivanco JM. Harnessing the rhizosphere microbiome through plant breeding and agricultural management. Plant Soil. 2012;360:1-13. https://doi.org/10.1007/s11104-012-1361-x

Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, Walters WA, Knight R, Fierer N. The under-recognized dominance of *Verrucomicrobia* in soil bacterial communities. Soil Biol Biochem. 2011;43:1450-5. https://doi.org/10.1016/j.soilbio.2011.03.012

Bloemberg GV, Lugtenberg BJ. Molecular basis of plant growth promotion and biocontrol by rhizobacteria. Curr Opin Plant Biol. 2001;4:343-50. https://doi.org/10.1016/S1369-5266(00)00183-7

Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114-20. https://doi.org/10.1093/bioinformatics/btu170

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37:852-7. https://doi.org/10.1038/s41587-019-0209-9

Boudjeko T, Tchinda RA, Zitouni M, Nana JA, Lerat S, Beaulieu C. *Streptomyces cameroonensis* sp. nov., a geldanamycin producer that promotes *Theobroma cacao* growth. Microbes Environ. 2017;32:24-31. https://doi.org/10.1264/jsme2.ME16095

Bravo D, Braissant O. Cadmium-tolerant bacteria: Current trends and applications in agriculture. Lett Appl Microbiol. 2022;74:311-33. https://doi.org/10.1111/lam.13594

Bravo D, Leon-Moreno C, Martínez CA, Varón-Ramírez VM, Araujo-Carrillo GA, Vargas R, Quiroga-Mateus R, Zamora A, Rodríguez EAG. The first national survey of cadmium in cacao farm soil in Colombia. Agronomy. 2021;11:761. https://doi.org/10.3390/agronomy11040761

Bravo D, Pardo-Díaz S, Benavides-Erazo J, Rengifo-Estrada G, Braissant O, Leon-Moreno C. Cadmium and cadmium-tolerant soil bacteria in cacao crops from northeastern Colombia. J Appl Microbiol. 2018;124:1175-94. https://doi.org/10.1111/jam.13698

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581-3. https://doi.org/10.1038/nmeth.3869

Callejas R. Generalidades del Departamento de Antioquia. In: Posada RC, Piedrahíta AI, editors. Flora de Antioquia: Catálogo de las plantas vasculares. Vol. I: Introducción. Bogotá, Colombia: Editorial ÓVinni; 2011. p. 11-8.

Cano Y. Expedición Cacao Colombia-BIO. v1.1. 2023. Available from: https://ipt.biodiversidad.co/permisos/resource?r=0359_cacaobio_20210719&v=1.1

Colombian Ministry of Environment and Sustainable Development. Política Nacional para la gestión integral de la biodiversidad y sus servicios ecosistémicos. Bogotá, Colombia: Ministerio de Ambiente y Desarrollo Sostenible; 2012. Available from: https://www.minambiente.gov.co/direccion-de-bosques-biodiversidad-y-servicios-ecosistemicos/politica-nacional-para-la-gestion-integral-de-la-biodiversidad-y-sus-servicios-ecosistemicos/.

Colombian Ministry of Environment and Sustainable Development. Plan Integral de Cambio Climático del Departamento del Chocó (PICC-Chocó). Bogotá, Colombia: Ministerio de Ambiente y Desarrollo Sostenible; 2016. Available from: https://archivo.minambiente.gov.co/images/BosquesBiodiversidadyServiciosEcosistemicos/pdf/nodo_pacifico/CC.compressed.pdf

Dávalos LM, Bejarano AC, Hall MA, Correa HL, Corthals A, Espejo OJ. Forests and drugs: Coca-driven deforestation in tropical biodiversity hotspots. Environ Sci Technol. 2011;45:1219-27. https://doi.org/10.1021/es102373d

Delgado-Ospina J, Molina-Hernández JB, Chaves-López C, Romanazzi G, Paparella A. The role of fungi in the cocoa production chain and the challenge of climate change. J Fungi. 2021;7:202. https://doi.org/10.3390/jof7030202

Diamond S, Lavy A, Crits-Christoph A, Carnevali PBM, Sharrar A, Williams KH, Banfield JH. Soils and sediments host Thermoplasmata archaea encoding novel copper membrane monooxygenases (CuMMOs). ISME J. 2022;16:1348-62. https://doi.org/10.1038/s41396-021-01177-5

Epp LS, Boessenkool S, Bellemain EP, Haile J, Esposito A, Riaz T, Brochmann C. New environmental metabarcodes for analysing soil DNA: Potential for studying past and present ecosystems. Mol Ecol. 2012;21:1821-33. https://doi.org/10.1111/j.1365-294X.2012.05537.x

Estensmo ELF, Maurice S, Morgado L, Martin-Sanchez PM, Skrede I, Kauserud H. The influence of intraspecific sequence variation during DNA metabarcoding: A case study of eleven fungal species. Mol Ecol Resour. 2021;21:1141-8. https://doi.org/10.1111/1755-0998.13329

González-Orozco CE, Sanchez-Galán AA, Ramos PE, Yockteng R. Exploring the diversity and distribution of crop wild relatives of cacao (Theobroma cacao L.) in Colombia. Genet Resour Crop Ev. 2020;67:2071-85. https://doi.org/10.1007/s10722-020-00960-1

Hackl E, Zechmeister-Boltenstern S, Bodrossy L, Sessitsch A. Comparison of diversities and compositions of bacterial populations inhabiting natural forest soils. Appl Environ Microb. 2004;70:5057-65. https://doi.org/10.1128/AEM.70.9.5057-5065.2004

Hansen AJ. Fusaria as agents of cacao green point cushion gall in the Caribbean and in Latin America. Plant Dis Rep. 1966;50:229-33.

Janssen PH. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. Appl Environ Microb. 2006;72:1719-28. https://doi.org/10.1128/AEM.72.3.1719-1728.2006

Kalam S, Basu A, Ahmad I, Sayyed RZ, El-Enshasy HA, Dailin DJ, Suriani NL. Recent understanding of soil Acidobacteria and their ecological significance: A critical review. Front Microbiol. 2020;11:580024. https://doi.org/10.3389/fmicb.2020.580024

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microb. 2013;79:5112-20. https://doi.org/10.1128/AEM.01043-13

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550. https://doi.org/10.1186/s13059-014-0550-8

Maddela NR, Kakarla D, García LC, Chakraborty S, Venkateswarlu K, Megharaj M. Cocoa-laden cadmium threatens human health and cacao economy: A critical view. Sci Total Environ. 2020;720:137645. https://doi.org/10.1016/j.scitotenv.2020.137645

Nagler M, Insam H, Pietramellara G, Ascher-Jenull J. Extracellular DNA in natural environments: Features, relevance and applications. Appl Microbiol Biot. 2018;102:6343-56. https://doi.org/10.1007/s00253-018-9120-4

Navarrete AA, Soares T, Rossetto R, van Veen JA, Tsai SM, Kuramae EE. Verrucomicrobial community structure and abundance as indicators for changes in chemical factors linked to soil fertility. Anton Leeuw. 2015;108:741-52. https://doi.org/10.1007/s10482-015-0530-3

Nilsson RH, Larsson KH, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Saar I, Kõljalg I, Abarenkov K. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Res. 2019;47:D259-64. https://doi.org/10.1093/nar/gky1022

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, et al. Package 'vegan' Community ecology package. Version, 2.6-2. Cran R Project; 2013. Available from: https://cran.r-project.org/web/packages/vegan/vegan.pdf.

Orgiazzi A, Dunbar MB, Panagos P, Groot GA, Lemanceau P. Soil biodiversity and DNA barcodes: opportunities and challenges. Soil Biol Biochem. 2015;80:244-50. https://doi.org/10.1016/j.soilbio.2014.10.014

Panke-Buisse K, Poole AC, Goodrich JK, Ley RE, Kao-Kniffin J. Selection on soil microbiomes reveals reproducible impacts on plant function. ISME J. 2015;9:980-9. https://doi.org/10.1038/ismej.2014.196

Ploetz RC. Fusarium-induced diseases of tropical, perennial crops. Phytopathology. 2006;96:648-52. https://doi.org/10.1094/PHYTO-96-0648

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2012;41:D590-6. https://doi.org/10.1093/nar/gks1219

Richardson JE, Whitlock BA, Meerow AW, Madriñan S. The age of chocolate: A diversification history of *Theobroma* and *Malvaceae*. Front Ecol Evol. 2015;3:120. https://doi.org/10.3389/fevo.2015.00120

Robeson MS, O'Rourke DR, Kaehler BD, Ziemski M, Dillon MR, Foster JT, Bokulich NA. RESCRIPt: Reproducible sequence taxonomy reference database management for the masses. PLoS Comput Biol. 2021;17:e1009581. https://doi.org/10.1371/journal.pcbi.1009581

Rofner NF. Cadmium in soil and cacao beans of Peruvian and South American origin. Rev Fac Nac Agron Medellín. 2021;74:9499-515. https://doi.org/10.15446/rfnam.v74n2.91107

Schmidt JE, DuVal A, Isaac ME, Hohmann P. At the roots of chocolate: Understanding and optimizing the cacao rootassociated microbiome for ecosystem services. A review. Agron Sustain Dev. 2022;42:14. https://doi.org/10.1007/s13593-021-00748-2

Singh BK, Bardgett RD, Smith P, Reay DS. Microorganisms and climate change: Terrestrial feedbacks and mitigation options. Nat Rev Microbiol. 2010;8:779-90. https://doi.org/10.1038/nrmicro2439

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol Ecol. 2012;21:2045-50. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Tanaka Y, Matsuzawa H, Tamaki H, Tagawa M, Toyama T, Kamagata Y, Mori K. Isolation of novel Bacteria including rarely cultivated phyla, *Acidobacteria* and *Verrucomicrobia*, from the roots of emergent plants by simple culturing method. Microbes Environ. 2017;32:288-92. https://doi.org/10.1264/jsme2.ME17027

ten Hoopen GM, Rees R, Aisa P, Stirrup T, Krauss U. Population dynamics of epiphytic mycoparasites of the genera *Clonostachys* and *Fusarium* for the biocontrol of black pod (*Phytophthora palmivora*) and moniliasis (*Moniliophthora roreri*) on cocoa (*Theobroma cacao*). Mycol Res. 2003;107:587-96. https://doi.org/10.1017/S095375620300772X

Valencia PM, Franco-Sierra ND, González MC, Baena-Bejarano N, Pulido-Santacruz P, Villegas AS, Luque ET, Herrera MG. 3.1 Protocolos de genómica para monitoreo ambiental asociado a acciones de respuesta por impacto o contingencia ambiental formalizados y listos para ser transferidos a usuarios interesados. Bogotá: Instituto de Investigación de Recursos Biológicos Alexander von Humboldt; 2021. Available from: http://repository.humboldt.org.co/handle/20.500.11761/35649.

Wagner MR, Lundberg DS, Coleman-Derr D, Tringe SG, Dangl JL, Mitchell-Olds T. Natural soil microbes alter flowering phenology and the intensity of selection on flowering time in a wild *Arabidopsis* relative. Ecol Lett. 2014;17:717-26. https://doi.org/10.1111/ele.12276

White T, Bruns T, Lee S, Taylor JW. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ, editors. PCR Protocols: A guide to methods and applications New York: Press Inc.; 1990. p. 315-22. https://doi.org/10.1016/b978-0-12-372180- 8.50042-1