ANYTHING BUT BLAND: UNCOVERING THE HIDDEN DIVERSITY AND GENOMIC ORIGIN OF THE THREATENED VANILLA SPICE

by

Paige Ellestad

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Ecology, Evolution, and Behavior

Boise State University

August 2023

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the dissertation submitted by

Paige Ellestad

Dissertation Title: ANYTHING BUT BLAND: UNCOVERING THE HIDDEN DIVERSITY AND GENOMIC ORIGIN OF THE THREATENED VANILLA SPICE

Date of Final Oral Examination: 02 March 2023

The following individuals read and discussed the dissertation submitted by student Paige Ellestad, and they evaluated the student's presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Sven Buerki, Ph.D.	Chair, Supervisory Committee
Stephen Novak, Ph.D.	Member, Supervisory Committee
Marcelo Serpe, Ph.D.	Member, Supervisory Committee
Félix Forest, Ph.D.	Member, Supervisory Committee

The final reading approval of the dissertation was granted by Sven Buerki, Ph.D., Chair of the Supervisory Committee. The dissertation was approved by the Graduate College.

DEDICATION

I would like to dedicate this document to my family and friends for their support during my academic journey at Boise State University and to my previous managers who have pushed me to believe in myself and pursue my dreams: Maria Blatter, Amanda Staton Mason, and Paul Emerson. Additionally, I would like to dedicate this to my advisor, Dr. Sven Buerki, who has been an inspiration and mentor to my professional development.

ACKNOWLEDGMENTS

I would like to thank Dr. Miguel Angel Perez Farrera and Marco Antonio Dominguez Vazquez for collaboration and sample collection and maintenance, as well as all farmers, botanists, and conservationists who helped to acquire material for this study. I would also like to thank Jaya LittleWing, Michael Wojahn, Carlos Dumaguit, Peggy Martinez, Anthony Melton, Felix Forest, Stephen Novak, Marcelo Serpe, and Sven Buerki for assistance and guidance on this project. Additionally, I thank LUSH Cosmetics Ltd. for funding and the Department of Biology at Boise State University for support.

From the following herbaria, I would like to thank Germán Carnevali Fernández-Concha (CICY), Nicolalde Morejón Edison Fernando (CIB), Louis Nusbaumer (G), Maria Regina de Vasconcellos Barbosa (JPB), Silvia Lobo (INB), Andres Alberto Barona Colmenares (COAH), Paula Leitman (ALCB), Lauren Raz and Henry Dario Agudelo Zamora (ICN), Norbert Holstein (BM), Miguel Angel Perez Farrera (HEM) and Alicia Rojas for providing images of herbarium specimens. Additionally, I thank O. M. Montiel and M. Rooney (MO) for helping to supply plant material.

ABSTRACT

Conserving the genetic diversity of crop species and their wild relatives has become a mounting concern as the detrimental effects of climate change, habitat destruction, and genetic erosion are being realized. In this epoch of unprecedented biodiversity loss, the genetic resources needed to improve crops may be at risk of extinction. Even one of the most iconic spices, vanilla, is threatened. Wild populations of the main vanilla producing species, *Vanilla planifolia* Andrews (Orchidaceae), are being rapidly extirpated due to deforestation and illegal harvesting in their native range. On top of that, clonal propagation methods within cultivated plants are hypothesized to have limited their genetic diversity and decreased their ability to cope with changing environmental conditions and respond to pathogens. Although the vanilla spice is so well-known, there is an unexpected lack of knowledge on its natural history and the ecological and evolutionary processes that have shaped its genetic resources, overall hindering its effective preservation and sustainability. To mitigate this gap of knowledge and help ensure the sustainability of this globally important spice, this thesis aimed to unravel the cultivation and domestication processes that have affected *V. planifolia* in its cultivated center of origin, Mexico, by answering the fundamental questions: What is the native distribution of *V. planifolia?* What are its crop-wild relatives and how should their conservation be prioritized? How many vanilla species are cultivated in its center of origin? What domestication processes have shaped its genetic resources? and What is the genomic origin of cultivated vanilla? By answering these questions, this thesis aimed to

distinguish between cultivated wild populations, regionally domesticated landraces, and globally domesticated cultivars using a range of approaches, from ecological to taxonomic to comparative phylogenetic to genomic. Analyses used samples collected from vanilla's cultivated center of origin, Mexico, along with publicly available genetic sequences of *Vanilla spp.* and a haplotype-phased reference genome of the global "Daphna" cultivar. Results indicated that *V. planifolia* occurs within a larger distribution than previously expected, from Mexico to northern Brazil, along with ten crop-wild relatives. Occurrences from Mexico encompassed the range of climatic niches exhibited by all occurrences within the entire distribution. Due to this high climatic variability, along with recorded morphological variability in *V. planifolia*, Mexico was used as a focal region to assess vanilla's genetic resources. In addition to the predominantly cultivated *V. planifolia,* two other crop-wild relatives, *V. pompona* and *V. insignis,* were found to be cultivated in Mexico based on DNA barcoding of ITS sequences. Ten haplotypes were identified within Mexican accessions of *V. planifolia* and two were identified within *V. pompona*. Genetic variability and high levels of genome-wide heterozygosity found within Mexican *V. planifolia* and the "Daphna" cultivar revealed the occurrence of multiple domestication events and past hybridization within cultivated vanilla. Signatures of introgressive hybridization between *V. planifolia* and *V. pompona* were discovered in the "Daphna" cultivar based on comparative chromosomal analyses (e.g. incongruence along the terminal region of chromosome two). A parental origin for the highly heterozygous Mexican accessions, however, has yet to be identified. Considering the high levels of crop-wild relative diversity and the long history of cultivation by different cultural groups in Mexico, these results might provide evidence

for regionally cultivated landraces produced from regional domestication events. These may provide important sources of genetic diversity to potentially increase crop resilience in the face of climate change. Findings from this thesis provide a clearer illustration of vanilla's genetic resources and support the urgent prioritization of biodiversity within this important region through the conservation of *V. planifolia's* crop-wild relatives and landraces. These recommendations will help to benefit the livelihoods of farmers, encourage the protection of biological and cultural diversity in Mexico, and ultimately help to ensure the sustainable cultivation of this iconic spice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ALCB	Universidade Federal da Bahia Herbario, Campus Universitário de Ondina
AOO	Area of Occurrence
BM	The Natural History Museum Herbarium, London
BPP	Bayesian posterior probability
BS	Bootstrap
BSU	Boise State University
CIB	Universidad Veracruzana Herbario
CICY	Centro de Investigación Científica de Yucatán, A.C. Herbario
COAH	Instituto Amazónico de Investigaciones Científicas SINCHI Herbario
CR	Critically endangered
CWR	Crop-wild Relative
EN	Endangered
EOO	Extent of Occurrence
G	Geneva Herbarium
GBIF	Global Biodiversity Information Facility
GWH	Genome-wide heterozygosity
HEM	Universidad de Ciencias y Artes de Chiapas Herbarium

ICN	Universidade Federal do Rio Grande do Sul Herbario
INB	Herbario Instituto Nacional de Biodiversidad
ITS	Internal Transcribed Spacer
IUCN	International Union for the Conservation of Nature
JPB	Universidade Federal da Paraíba, Cidade Universitária Herbario
K	Royal Botanical Gardens, Kew
K2P	Kimura two-parameter model
LC	Least concern
MO	Missouri Botanical Garden
PCA	Principal component analysis
PCR	Polymerase chain reaction
PGR	Plant Genetic Resources
rbcL	RuBisCo large subunit
SNP	Single nucleotide polymorphism
SRA	Sequence read archive
SRP	Snake River Plains Herbarium

CHAPTER ONE: HARNESSING LARGE-SCALE BIODIVERSITY DATA TO INFER THE CURRENT DISTRIBUTION OF *VANILLA PLANIFOLIA* (ORCHIDACEAE)

The final version of this article has undergone full peer review and has been published. Please see:
Ellestad, P., Forest, F., Serpe, M., Novak, S. J., & Buerki, S. (2021). Harnessing large-scale biodiversity data to infer the current distribution of *Vanilla planifolia* (Orchidaceae). *Botanical Journal of the Linnean Society*, *XX*, 1–16. https://doi.org/10.1093/botlinnean/boab005

Abstract

Although vanilla is one of the most popular flavours in the world, there is still uncertainty concerning the native distribution of the species that produces it, *Vanilla planifolia*. To circumscribe the native geographical extent of this economically important species more precisely, we propose a new landscape-based approach to incorporate information from open-source databases and validate occurrences. In this approach, we include metrics to account for habitat suitability and population sustainability in terms of the biotic (co-occurrence of pollinators and dispersers) and abiotic (habitat quality) factors limiting plant distributions. To further validate occurrences within the resulting distribution, we compare the presence of morphologically similar wild relatives, assess the heterogeneity of ecological niches and verify the correct identification of herbarium specimens. Results from this approach suggest that *V. planifolia* has a larger geographical distribution than previously recognized; we hypothesize that populations naturally dispersed from Mesoamerica and became established in South America (with a south-eastern limit in Brazil). The recognition of an improved estimate of the distribution of

this species will increase the accuracy of predictive models, promote further species circumscription, improve the efficacy of conservation strategies, and help to ensure the sustainability of a valuable, sought-after spice

Introduction

Accurate knowledge of the extent of the geographical distribution of a species is an essential component of effective conservation. However, a lack of this knowledge, known as the Wallacean shortfall (Lomolino, 2004), is prevalent among our understanding of global biodiversity, most notably among tropical species (Bini et al., 2006; Whittaker et al., 2005). Anthropogenic changes in land use, coupled with human-induced climate change, compound this challenge to biodiversity conservation by altering the composition of biological communities through population declines, extirpations, distribution shifts and species extinctions (Ceballos et al., 2017; Ellis et al., 2012; Sala et al., 2000; Walther et al., 2002). Even our knowledge of one of the most iconic spices worldwide, vanilla, is affected by the Wallacean shortfall. To mitigate this deficiency and help to ensure the sustainability of this globally important and valuable spice under future landscape conditions, we present an approach to resolve its current distribution and shed light into the extent of its geographical variation.

Derived from the cured seed pod of the tropical orchid *Vanilla planifolia* Andrews (Orchidaceae), vanilla is the second most valuable spice in the world (second only to saffron). In 2018 it was valued at $515/ kg, close to the price of silver (Baker, 2018). With a purported origin in Mesoamerica, vanilla has been spread across the globe to be cultivated for use in the culinary, cosmetic, and medicinal industries (Bruman, 1948; Lubinsky, Bory, et al., 2008). Historical records indicate that vanilla was used as a

flavoring and medicinal beverage by multiple cultures in Mesoamerica, including the Totonacs, the Mayans and the Aztecs. After the Spanish conquest of the Aztecs in 1520 AD, it was transported to Europe. Vanilla was not cultivated globally until 1832, when Edmond Albius, an enslaved Frenchman born on La Réunion, developed a technique for manually pollinating its flowers (Rain, 2004). Currently, Madagascar is the largest producer of vanilla, followed by Indonesia and Mexico. Many tropical countries rely heavily on vanilla production to sustain their economy and provide a livelihood for agricultural workers (Food and Agriculture Organization of the United Nations (FAO), 2020).

Vanilla planifolia is only able to reproduce naturally in the presence of specific pollinators. In cultivation outside its native range, however, it must be hand-pollinated to produce fruit (Arditti et al., 2009). In addition to the putative single origin of the cultivated variety and limited genetic variation, clonal propagation methods have severely constrained the genetic diversity of global cultivars (Lubinsky, Bory, et al., 2008). Thus, we hypothesize that the capacity of globally cultivated vanilla to rapidly adapt to new environmental conditions and community shifts created by human-induced climate change and their ability to respond to new pathogens is limited. Evidence for such a hypothesis already exists as demonstrated by large-scale loss of vanilla plantations due to fungal pathogen outbreaks. For instance, vanilla stem rot disease due to *Fusarium* spp. threatened vanilla production in Indonesia and has caused significant economic losses over the last decade (Pinaria et al., 2010). Because of these threats, we advocate for the study of vanilla populations within its native range to identify genotypes that would be potentially adapted to new human-induced environmental conditions. Many of

these populations, however, are being extirpated by land-use change, habitat fragmentation and illegal harvesting (Hinsley et al., 2018; Soto Arenas & Dressler, 2010). Conservation of the remaining populations within its native range is essential to ensure the sustainability of global vanilla production.

Achieving the ambitious goal detailed above and identifying genotypes adapted to future climates relies on accurate knowledge of the native range of vanilla, which remains elusive. Currently, two contrasting hypotheses exist concerning the native distribution of *V. planifolia* (Figure 1.1). According to two prominent researchers on the systematics of *Vanilla* Plum. ex Mill., the distribution of *V. planifolia* extends from southern Mexico to Panama, encompassing most of Mesoamerica (Figure 1.1A; Soto Arenas & Cribb, 2010). In contrast, the IUCN reports a much more narrow and fragmented distribution for the species, extending from southern Mexico to Belize (Figure 1.1B; Vega et al., 2017). Biodiversity occurrence data from the Global Biodiversity Information Facility (GBIF) show that *V. planifolia* occurs all over the world, but this includes specimens collected from cultivated material in addition to living collections mostly grown at botanical institutions (Figure 1.1C). To make matters even more difficult to disentangle, the type specimen of *V. planifolia* is a drawing of a plant that was in cultivation in the West Indies (Soto Arenas & Cribb, 2010). The absence of a physical herbarium specimen, and the fact that the drawing was from a plant in cultivation, greatly hinders the morphological delimitation of this species and contributes to another challenge, known as the Linnaean shortfall, which further complicates an understanding of the taxonomy of *V. planifoia* and its native distribution.

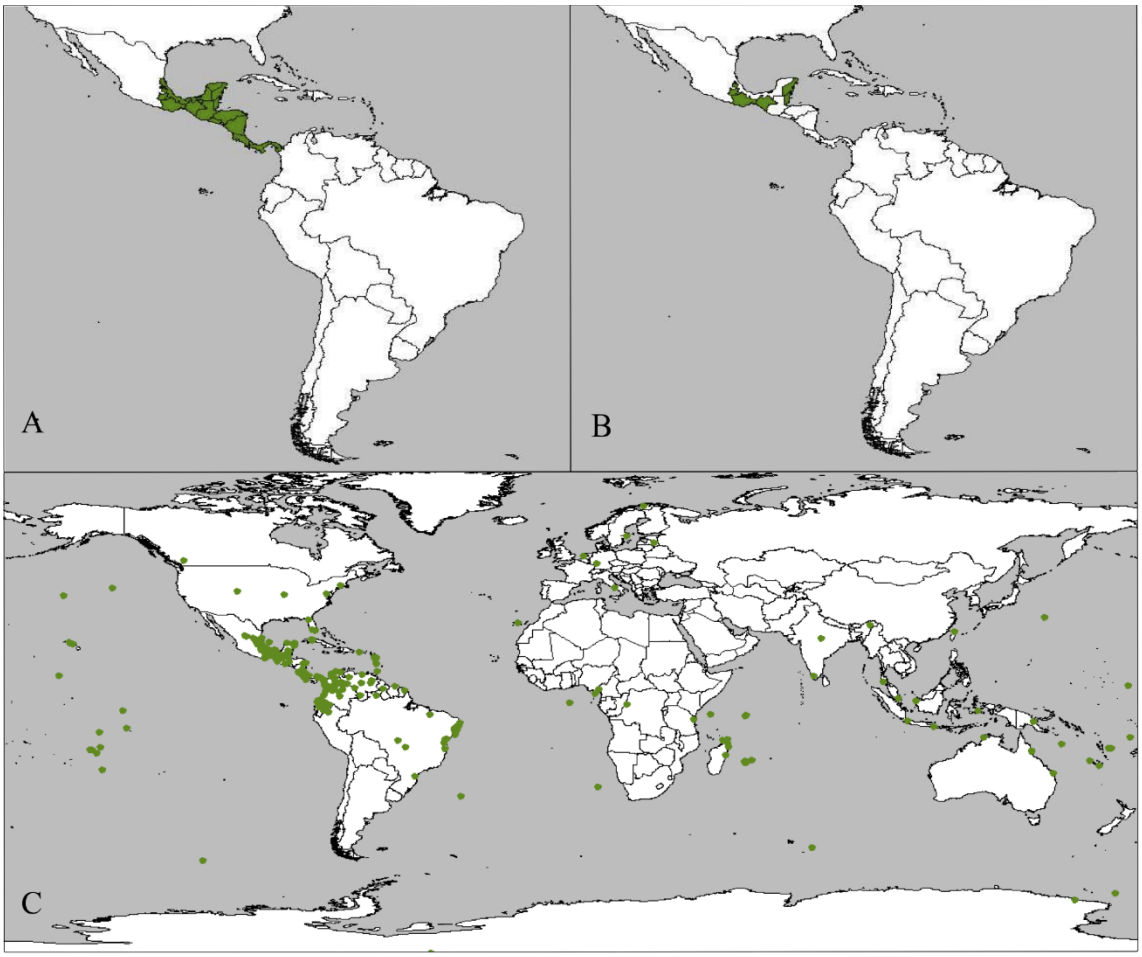


Figure 1.1 Previous distribution hypotheses of *Vanilla planifolia*. A) Distribution based on Soto Arenas & Cribb (2010). B) Distribution produced by Vega et al., (2017) for the IUCN. C) Un-curated occurrence data from GBIF

Here, we contend that accounting for the biological and ecological factors necessary to support sustainable vanilla populations could contribute to reconciling the uncertainty associated with its distribution. Interspecific interactions play important roles in creating a suitable habitat and are vital for reproductive success and genetic recombination. More than 70% of angiosperms rely mainly or exclusively on plant-animal interactions to complete their life cycle (Fontaine et al., 2005), and some estimates increase this value to almost 90% (Ollerton et al., 2011). Such biotic interactions should be taken into account when assessing the landscape in which vanilla occurs. *Vanilla*

planifolia is self-compatible, but incapable of self-fertilization without a pollinator (Bory, Grisoni, et al., 2008). Although we still have limited knowledge of the pollination syndrome of *V. planifolia*, most research indicates that bees from the genera *Eulaema* and *Euglossa* (tribe Euglossini; Apidae), which occur only in the New World, serve as pollinators (Ackerman, 1983; Rodolphe et al., 2012; Soto Arenas & Dressler, 2010). There are few recorded observations of the dispersal of vanilla fruits, but the strong aroma emitted by these fruits suggests dispersal by frugivorous bats (Schlüter et al., 2007; see next for more details). In addition to natural interactions, human activities such as habitat destruction and degradation have tremendous impacts on local populations. Approximately 1 × 106 km2 of rainforest, which provides habitat for over two-thirds of all species on the planet, is cleared every five to ten years (Pimm & Raven, 2000). Consequently, it may be inaccurate to infer the current distribution of *V. planifolia* based solely on historical occurrence data. Metrics reflecting habitat quality at a fine scale, e.g. the human influence index that records metrics of human disturbance across landscapes (WCS & University, 2005), are likely to improve inferences of the current distribution of *V. planifolia.*

With this study, we offer an integrative approach to better circumscribe the current geographical distribution of *V. planifolia* by accounting for the co-occurrence of pollinators and dispersers and habitat quality/disturbance. We take advantage of open-source biodiversity occurrence databases deposited on GBIF (GBIF.Org, 2019), which do have limitations, such as incorrect georeferencing, lack of voucher specimens and species misidentification. Our approach attempts to correct for these flaws, filter validated occurrences and characterize the ecological niche of *V. planifolia* populations based on

biotic and abiotic factors. Results from this study can be used to conduct targeted fieldwork to assess and hopefully preserve the phenotypic and genotypic variation in this economically important species (Brummitt et al., 2015). By better understanding the extent of its distribution, this study will contribute the raw data to facilitate the search for populations of *V. planifolia* capable of tolerating future climate change (especially drought and outbreaks of pathogens and pests predicted to occur with these new environmental conditions; see previous), thereby ensuring the maintenance of this crop to sustain local cultures and economies.

Materials and Methods

A three-step data exclusion analysis was implemented in R (R Core Team, 2017) to determine the extent of the current distribution of *V. planifolia* using biodiversity occurrence data from GBIF. Starting with all available biodiversity occurrences, we successively eliminated unsuitable occurrences and identified suitable occurrences based on a set of criteria. Our three steps included tests for the validity of biodiversity occurrences based on plant life-cycle requirements and states of human disturbance of the landscape: step (1) presence of pollinators; step (2) presence of dispersers and step (3) habitat quality (Figure 1.2). Yearly sampling effort was assessed by plotting two cumulative frequency curves of biodiversity occurrences using basic R functions: the first curve for all occurrences, and the second curve based on occurrences remaining after our three-step data exclusion analysis. This approach allows for an assessment of sampling efforts over time and the inclusion of older and more recent occurrences within the final distribution. In addition, the record type (observation, living specimen, preserved

specimen or unknown) of initial and final occurrences by country was compared to assess legitimacy among types and regional variation of occurrences.

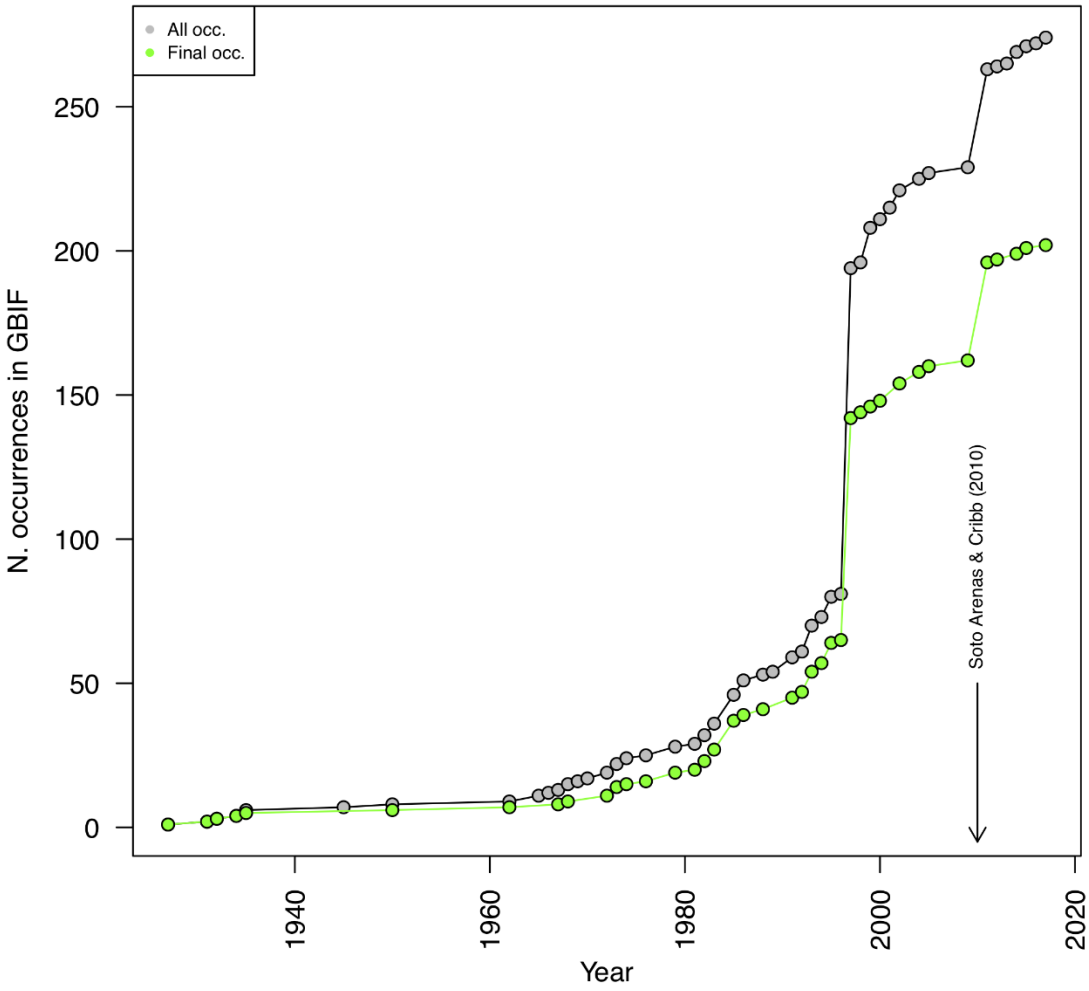


Figure 1.2 Cumulative curve of years that *V. planifolia* occurrences were recorded. The grey line and symbols represent the occurrence dataset after curation, and the green line and symbols represent the final occurrence dataset. See the text for an explanation of these two datasets. Additional occurrence data has been made available since the publication of the distribution hypothesized by Soto Arenas & Cribb (2010).

Biodiversity occurrence data cleaning procedure

We downloaded three datasets from GBIF that matched the search under 'Scientific Name' for the target species (*Vanilla planifolia*), pollinator species (species of *Eulaema* and *Euglossa*) and disperser species (frugivorous species in Phyllostomidae). A list of bat species recorded to eat fruit was created using information from Reid (2009), López-Baucells *et al.* (2016) and the IUCN Red List of Threatened Species (Appendix A).

Due to the amount of georeferencing errors within occurrence datasets, we manually curated the *V. planifolia* dataset so that coordinates reflected the location description included in the occurrence data. Where location coordinates and descriptions did not match, a new coordinate was geo-referenced using satellite imagery from Google Earth v.7.3.3 based on these descriptions. We added two columns to the dataset to account for adjusted latitude/longitude occurrences and these were used in subsequent analyses.

We cleaned the pollinator and disperser datasets by removing all occurrences without geographical coordinates. We then used the function *SpatialPoints* in the R package *rgeos* to transform coordinates into spatial points from the occurrence data and excluded points falling into the oceans by using a shapefile of the oceans available at http://www.naturalearthdata. com (Bivand & Rundel, 2019).

Step1: presence of pollinators

The R package *raster* was used to create a blank raster to assess co-occurrence of pollinator species within cells (Hijmans, 2019a). Insect faunas have been notoriously under sampled and poorly described (only *c*. 20% of global species have been named;

Stork, 2018). Despite obstacles within landscapes, bees from tribe Euglossini (including *Eulaema* and *Euglossa*) have been found to travel large distances to forage, sometimes up to 50 km (Pokorny et al., 2015). Therefore, we used a raster resolution of 30 minutes (*c*. 55 km2) to act as a buffer around *V. planifolia* occurrences and minimize the effect of under sampled pollinators. The *SpatialPoints* function implemented in the R package *sp* was used to overlay occurrences of *V. planifolia* and pollinator genera onto the species distribution raster (Pebesma & Bivand, 2018). The presence of a pollinator in the same cell as the target species was preliminarily considered suitable for pollination to occur. Thus, all *V. planifolia* occurrences that did not co-occur with a pollinator were excluded from further analysis because such populations were deemed unable to reproduce naturally (i.e. they would require human intervention to reproduce due to the absence of suitable pollinators).

Step2: presence of dispersers

The approach used for pollinators was also applied to disperser occurrence data. Although literature on fruit dispersal by bats is sparse, frugivorous bats (Phyllostomidae) have been recorded to travel long distances for foraging and migration (> 100 km for *Artibeus lituratus*; Arnone et al., 2016). As with the pollinator analysis, we used a resolution of 30 minutes to act as a buffer around *V. planifolia* occurrences and minimize the effect of under sampled dispersers. Therefore, the presence of dispersers in the same cell as the target species was judged suitable to ensure seed dispersal and to contribute to recruitment within a population and gene flow among populations. All *V. planifolia* occurrences that did not co-occur with a disperser were deemed unable to be naturally

dispersed due to the absence of suitable dispersers and were excluded from further analysis.

Step 3: habitat quality

We used the human influence index (HII) as a proxy for the current impact of human activities on the natural landscape. The HII is a global dataset that incorporates human population pressure, land use and infrastructure, and human access to formulate an index between 0 and 100, where 0 indicates no human disturbance and 100 indicates habitats that are completely disturbed (WCS & University, 2005). Using the R package *raster*, we extracted HII values for each 1-km geographic cell that contained a *V. planifolia* occurrence (Hijmans, 2019a). We applied the majority rule criterion to HII values; therefore, landscapes with an HII of < 50 were presumed to be minimally impacted by human activities and those with an HII > 50 were presumed to be highly impacted. Occurrences that fell within minimally impacted areas were assumed to represent habitats currently occupied by *V. planifolia*. Conversely, occurrences that fell within highly impacted areas were assumed to be either introduced by humans (e.g. vanilla plantations) or natural populations extirpated by habitat destruction. These occurrences were excluded from the analysis.

Co-occurrence of closely related species

We assessed the feasibility of resulting occurrences by comparing our resulting distribution to that of occurrences of closely related species in the *Vanilla planifolia* group. Of the 106 *Vanilla* spp., Soto Arenas & Cribb (2010) identified 16 species belonging to the *V. planifolia* group based on similarities of flower morphology and which occur in tropical America (including the West Indies). This group includes: *V.*

appendiculata Rolfe, *V. bahiana* Hoehne, *V. cristagalli* Hoehne, *V. denticulata* Pabst, *V. dubia* Hoehne, *V. dungsii* Pabst, *V. fimbriata* Rolfe, *V. helleri* A.D. Hawkes, *V. insignis* Ames, *V. odorata* C. Presl, *V. phaeantha* Rchb.f., *V. planifolia*, *V. ribeiroi* Hoehne, *V. schwackeana* Hoehne, *V. tahitiensis* J.W. Moore and *V. uncinata* Huber ex Hoehne. Flower morphology in this group is deceptive, and the same insect(s) could potentially pollinate different species, mediating hybridization (Soto Arenas & Cribb, 2010). With the exception of *V. tahitensis*, which is not found in the wild (Lubinsky *et al.*, 2008), overlapping distributions of these species reinforce the possibility of an accurate native distribution and identify regions where gene flow may occur. We downloaded GBIF data for all occurrences of these closely related 15 species and cleaned the data using the methods described above. To follow the same approach as applied with the target pollinators and dispersers, we also used a blank raster with a resolution of 30 minutes to overlay filtered occurrences of *V. planifolia* and its wild relatives and to identify cells that contained both.

Inferring the ecological niche of *Vanilla planifolia*

The ecological niche of *V. planifolia* was inferred from filtered occurrence data (see above) based on their biotic and abiotic environment. Within the final distribution raster overlaid by resulting *V. planifolia* occurrences, geographical distances between the closest co-occurring pollinator, disperser and wild relatives were assessed. Using the *DistHaversine* function in the R package *geosphere*, the minimum distance between points was calculated (Hijmans, 2019b). A histogram was created to visualize the frequency and extent of distances between resulting *V. planifolia* occurrences and their proximate biotic community. This approach allowed for estimating at a smaller scale, the

likelihood of co-occurring species to participate in essential life-cycle processes, such as pollination, dispersal and recruitment, to ensure sustainable populations of *V. planifolia.*

The ecological niche of *V. planifolia* was also inferred from abiotic factors by using climatic variables following a previously published procedure by Bone et al., (2015) and by extracting biome data using the shapefile from EcoRegions2017 (Dinerstein et al., 2017). First, 19 bioclimatic temperature and precipitation variables were retrieved for each occurrence from WorldClim at a resolution of 2.5 minutes (Fick & Hijmans, 2017). Because the bioclimatic variables were highly correlated with each other, a principal component analysis (PCA) was performed to summarize them as highly explanatory eigenvalues using the R package *vegan* (Oksanen et al., 2019). The PCA was performed after standardizing the bioclimatic variables (i.e. subtraction of mean followed by variance division, as implemented in R). Next, additional ecological data was gathered by assessing the biomes of each resulting *V. planifolia* occurrence. Biome shapefiles were downloaded from Ecoregions2017 and overlaid with occurrence points using the R package *rgeos* (Bivand & Rundel, 2019). Biome polygons containing occurrences were extracted using the package *raster* and variation among the dataset was assessed. Climate and biome data associated with each resulting occurrence allowed for the comparison of populations within three regions: Mexico (the purported origin of *V. planifolia* and the only region with Belize to be included in all three hypotheses; Figure 1.1); Central America (from Guatemala and Belize to Panama) and South America (from Colombia to Peru and Brazil).

Sampling efforts of plants in South America

Gaps in biodiversity collections are common in tropical countries and have been noted as a significant issue for species collection data in biodiversity hotspots (Buerki et al., 2015; Kier et al., 2005; Nelson et al., 1990). Based on these well-documented challenges, we predicted that a sampling bias may affect the resulting distribution of our target species. To identify and visualize this potential bias, we constructed a specimen richness map of all plants in Central and South America following methods described in Buerki *et al.* (2015). We used all occurrences from GBIF that matched the query for 'Plantae', which included a preserved specimen and cleaned the dataset using the methods outlined for previous biodiversity occurrence datasets.

Results

We downloaded 1,262 occurrences for *V. planifolia*, 73,097 occurrences for pollinators and 249,905 occurrences for dispersers from GBIF. Processes to clean GBIF occurrence data reduced pollinator and disperser occurrences on average by 39%, and curation of *V. planifolia* occurrences reduced their number by 54%. After these processes, occurrence numbers were reduced to 578 (*V. planifolia),* 44,634 (pollinators) and 155,093 (dispersers). Curated occurrences of *V. planifolia* were recorded between the years 1862 and 2017, with collection efforts increasing substantially in 1997 (Figure 1.2). Following the year that the hypothesized distribution was published by Soto Arenas & Cribb (2010), 45 additional occurrences of *V. planifolia* were recorded (Figure 1.2). Most occurrence data (61%) were recorded from Mexico but included occurrences from 32 additional countries around the globe. The basis of record of most occurrences was

documented as preserved specimens, meaning they can be found in an herbarium, although other types were also recorded.

Cleaned and curated occurrences were used to initiate the three-step analysis. After each step, unsuitable occurrences of *V. planifolia* were excluded (Table 1.1). Most occurrences were excluded after step 1, which constrained occurrences to the Americas. Resulting occurrences were located in ten countries and, reflecting the initial dataset, the majority of occurrences were from Mexico and recorded as preserved specimens (Table 1.2). Occurrences remaining after all three steps in the analysis depict a broader distribution than previously recognized. Based on 210 *V. planifolia* occurrences, our revised estimate of the current distribution of the species extends from Mexico southward through South America, to the eastern coast of Brazil, with an apparent gap in northern Brazil (Figure 1.3).

Table 1.1 Summary of GBIF biodiversity occurrences used in this study. The number of occurrences of *Vanilla planifolia* after each step in the pipeline are also displayed. See text for details.

	Mexico	**Central America**	**South America**	**Total**
Downloaded from GBIF				1,262
Curation	353	48	113	578
Step 1: Presence of pollinators	199	39	60	300
Step 2: Presence of dispersers	199	39	58	298
Step 3: Habitat availability	132	23	55	212
Verification of herbarium specimen	11	1	3	15

Table 1.2 ***Vanilla planifolia*** **GBIF occurrence record types by country. The numbers in front of the lines indicate the curated dataset before the analysis and the numbers after the line indicate the final occurrence dataset. Only countries containing final occurrence data are shown here.**

		Observation	Living specimen	Preserved specimen	Unknown	Total
Mexico	**Mexico**	4|1	0|0	169|116	26|15	199|132
Central America	**Belize**	0|0	0|0	1|1	0|0	1|1
	Costa Rica	2|2	1|1	18|10	1|1	22|14
	Nicaragua	0|0	0|0	2|2	0|0	2|2
	Panama	0|0	0|0	14|6	0|0	14|6
South America	**Colombia**	0|0	34|34	7|6	0|0	41|40
	Ecuador	0|0	0|0	11|11	0|0	11|11
	Venezuela	0|0	0|0	2|2	0|0	2|2
	French Guiana	0|0	0|0	1|1	0|0	1|1
	Brazil	0|0	0|0	3|1	0|0	3|1
Total		6|3	35|35	228|156	27|16	296|210

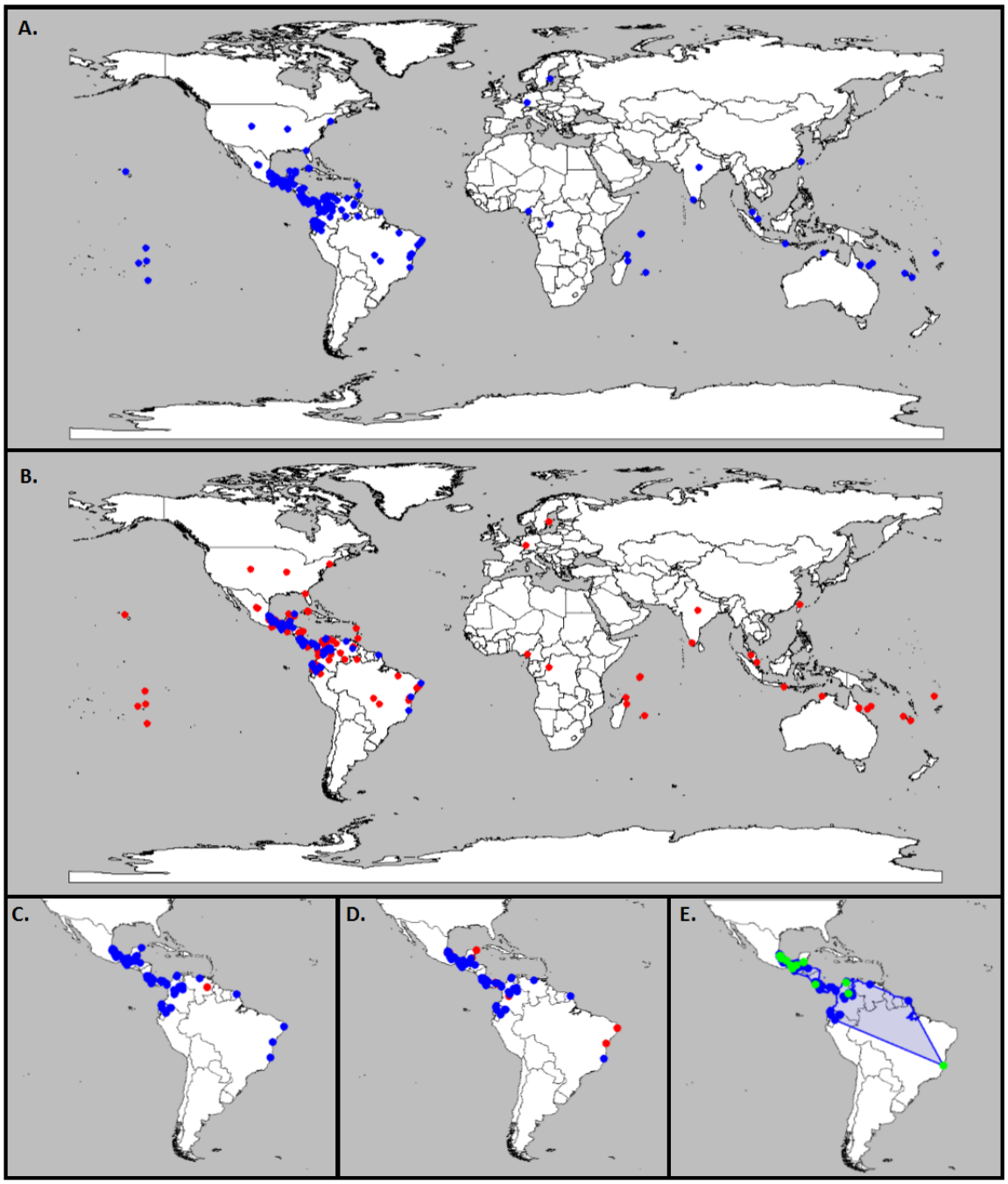


Figure 1.3 Occurrences of *Vanilla planifolia* throughout the landscape-based approach to infer geographical distribution. In maps A–D, blue circles represent occurrences of *V. planifolia* that have not been excluded from a particular step of the analysis, and red circles represent occurrences that have been excluded. A) All occurrences; B) Step 1, presence of pollinators; C) Step 2, presence of dispersers; D) Step 3, habitat quality and E) distribution of final occurrences. Green circles in E represent occurrences that have been taxonomically verified by herbarium images.

Co-occurrence of closely related species

To reinforce the possibility of this extended distribution, we plotted the coordinates of 569 occurrences of the 15 species that are morphologically closely related to *V. planifolia*. Within this raster, we identified seven cells that contained both *V. planifolia* and one of its wild relatives (Figure 1.4).

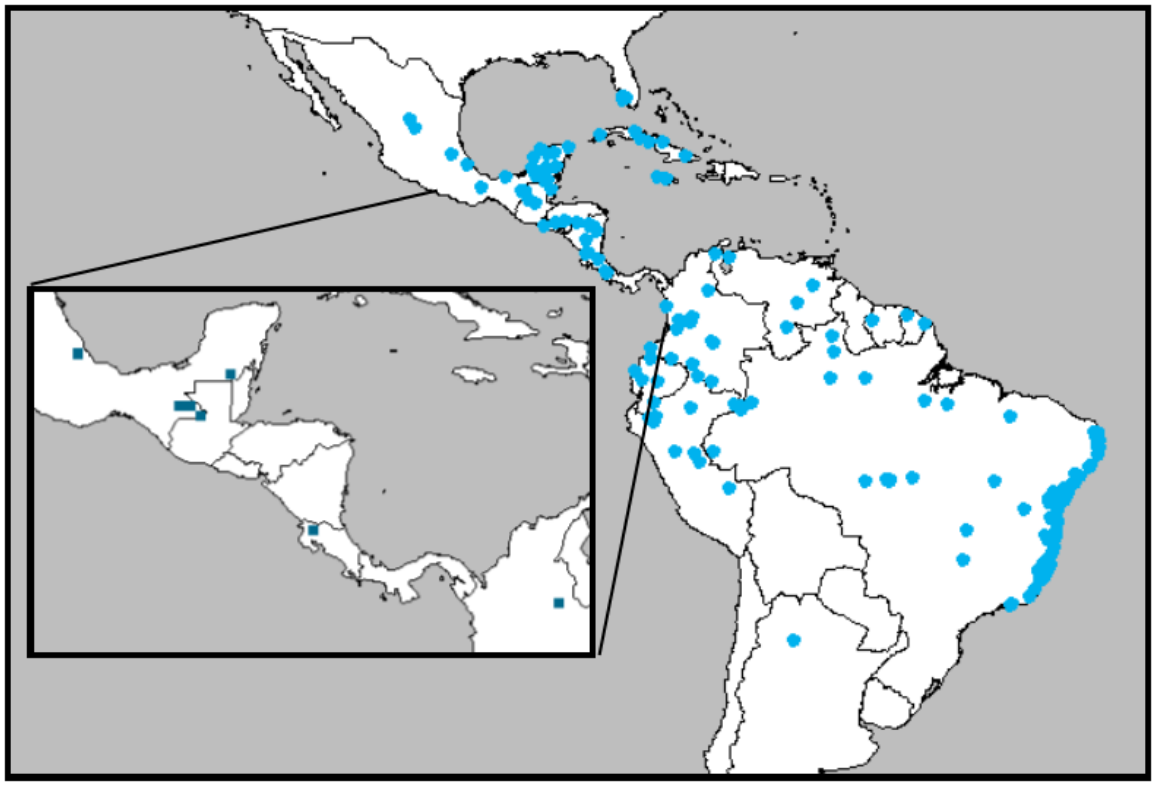


Figure 1.4 Co-occurrence of closely related species. Light blue circles represent wild relative occurrences based on GBIF dataset. In the insert, dark blue squares represent grid cells (0.50°) where *Vanilla planifolia* co-occurs with a wild relative.

The ecological niche of *Vanilla planifolia*

Histograms of the geographical distances between resulting *V. planifolia* occurrences and their closest co-occurring (within the same cell) pollinator, disperser and wild relative are shown in Figure 1.5. The frequency of distances to the closest recorded pollinator peaked at 35–45 km, with a secondary peak at 5–10 km. The majority of *V. planifolia* occurrences were closer to a disperser with the frequency of distances peaking

at 10–15 km and even closer to relatives with the frequency of distances peaking at 5–10 km.

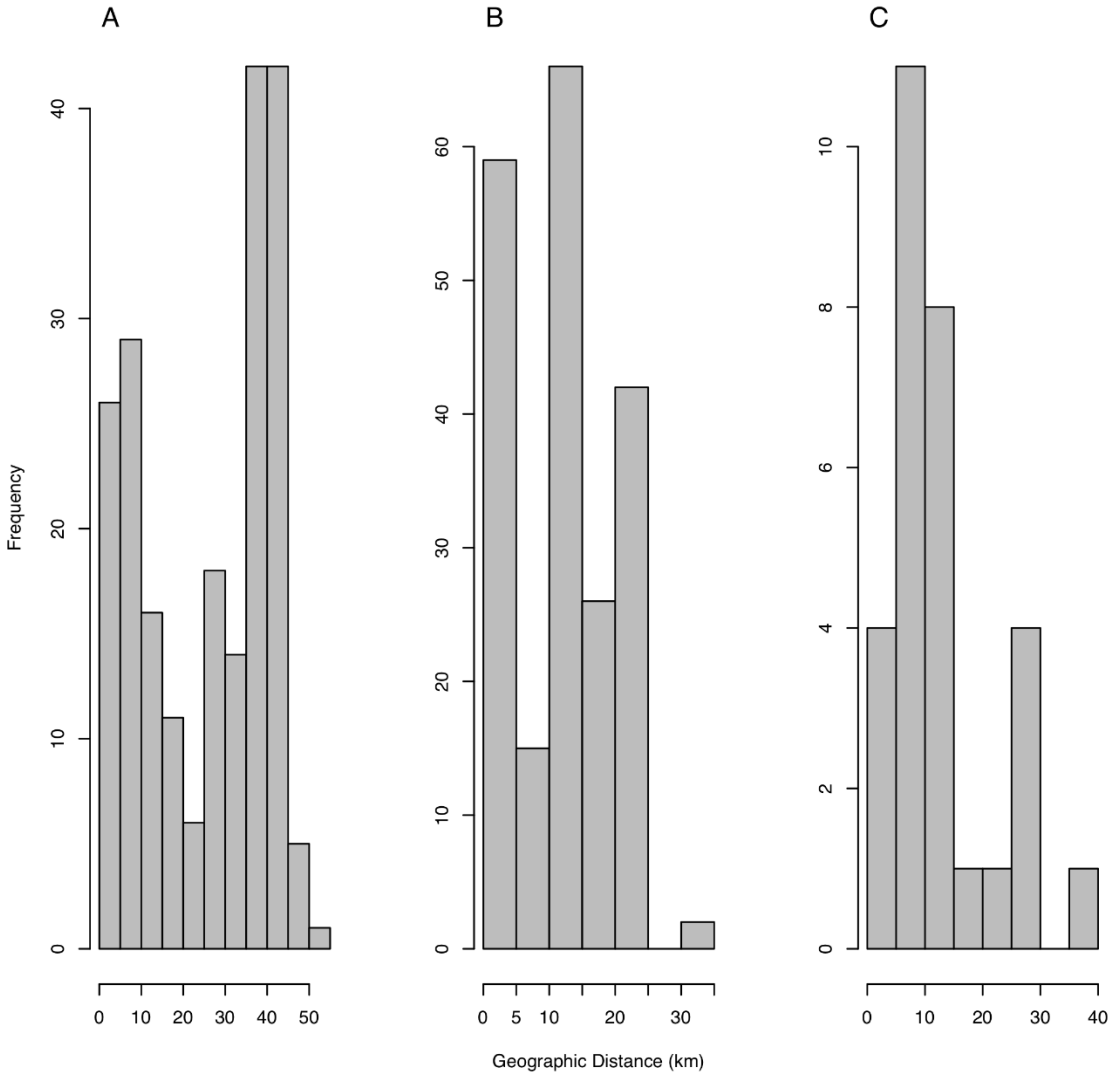


Figure 1.5 Geographic distances between resulting *V. planifolia* occurrences and A) a pollinator, B) a disperser and C) 15 wild relatives (from the *V. planifolia* group). See the text for a list of these 15 species.

The climatic niche of *V. planifolia* inferred using the 19 bioclimatic variables and the filtered occurrence dataset is displayed in Figure 6. The first two axes of the PCA explained 64.6% of the variance in the data. The most important variables contributing to

PC1 are BIO3 (isothermality), BIO4 (temperature seasonality), BIO6 (minimum temperature of coldest month), BIO7 (temperature annual range), BIO9 (mean temperature of driest quarter), BIO11 (mean temperature of coldest quarter) and BIO19 (precipitation of coldest quarter), and the main variables contributing to PC2 are BIO8 (mean temperature of wettest quarter) and BIO10 (mean temperature of warmest quarter). A figure representing the contribution of each variable to the first two principal components is available in Appendix B. Occurrences located in Mexico exhibited the most climatic variation, occurring in all quadrants, with most variation observed along PC2 (Figure 1.6). Central American occurrences group in the fourth quadrant, for which BIO1 (annual mean temperature) is the most discriminative variable. South American occurrences were grouped together in the first quadrant with BIO3 (isothermality) being the main bioclimatic factor.

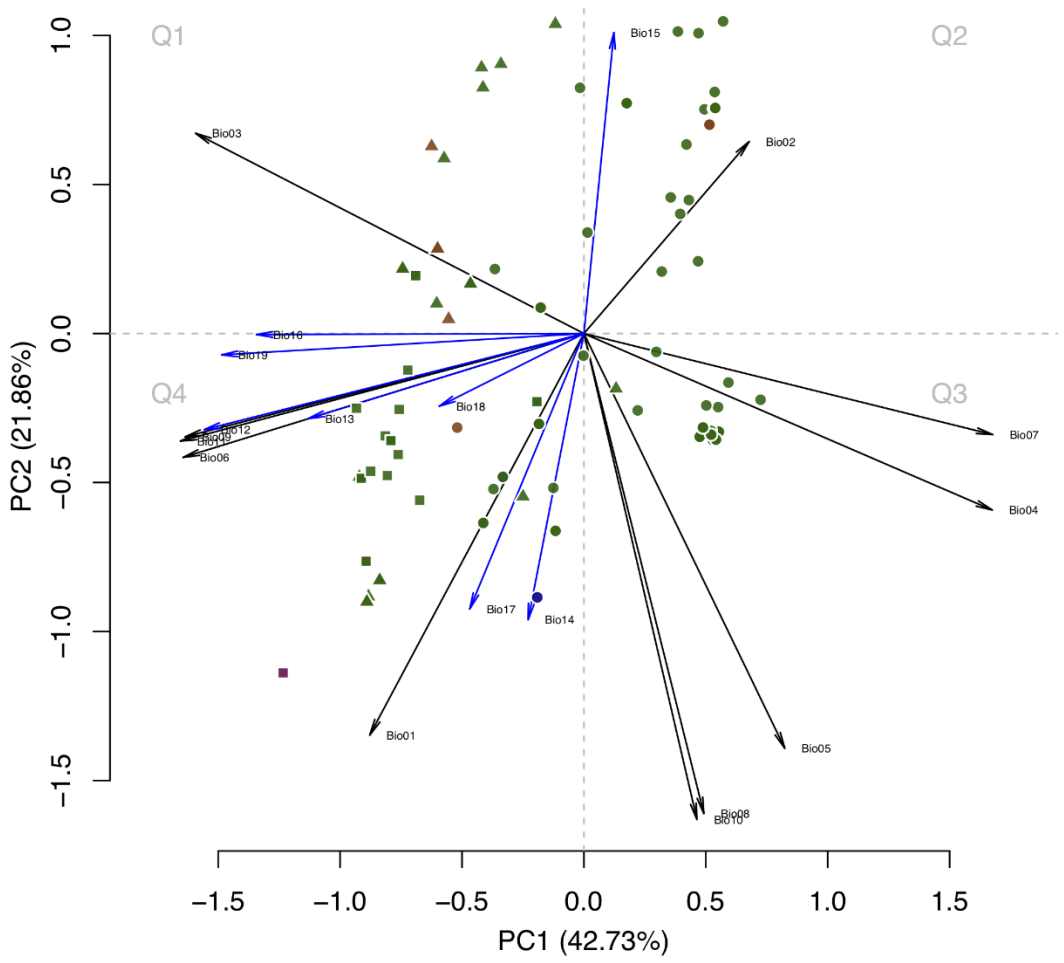


Figure 1.6 Climatic niche of filtered *Vanilla planifolia* occurrences as determined by a PCA of 19 bioclimatic variables. Occurrences from Mexico exhibit the most climatic variation, whereas occurrences from Central and South America are grouped together in the ordination space. The climatic niche of Mexican occurrences encompasses the climatic niches of Central and South American occurrences. Arrows represent contribution of individual bioclimatic variables to each axis and quadrant. Blue arrows correspond to variables associated with precipitation, and black arrows correspond to variables associated with temperature. Point colours correspond to occurrence biomes.

We found *V. planifolia* to occur in the following biomes: tropical and subtropical moist broadleaf forests; tropical and subtropical dry broadleaf forests; tropical and subtropical coniferous forests; deserts and xeric shrublands; and mangroves (Figure 1.7).

Most occurrences (77%) were located in tropical and subtropical moist broadleaf forests, and 20% occurred in tropical and subtropical dry broadleaf forests. Occurrences in the tropical and subtropical moist broadleaf forest biome were found in all three regions (Mexico, Central America and South America). The remaining biomes constituted <3% of occurrences (Table 1.3). When the climate variables of occurrences within each biome type of occurrences were evaluated using a PCA, no distinct grouping of biomes was apparent (Figure 1.6).

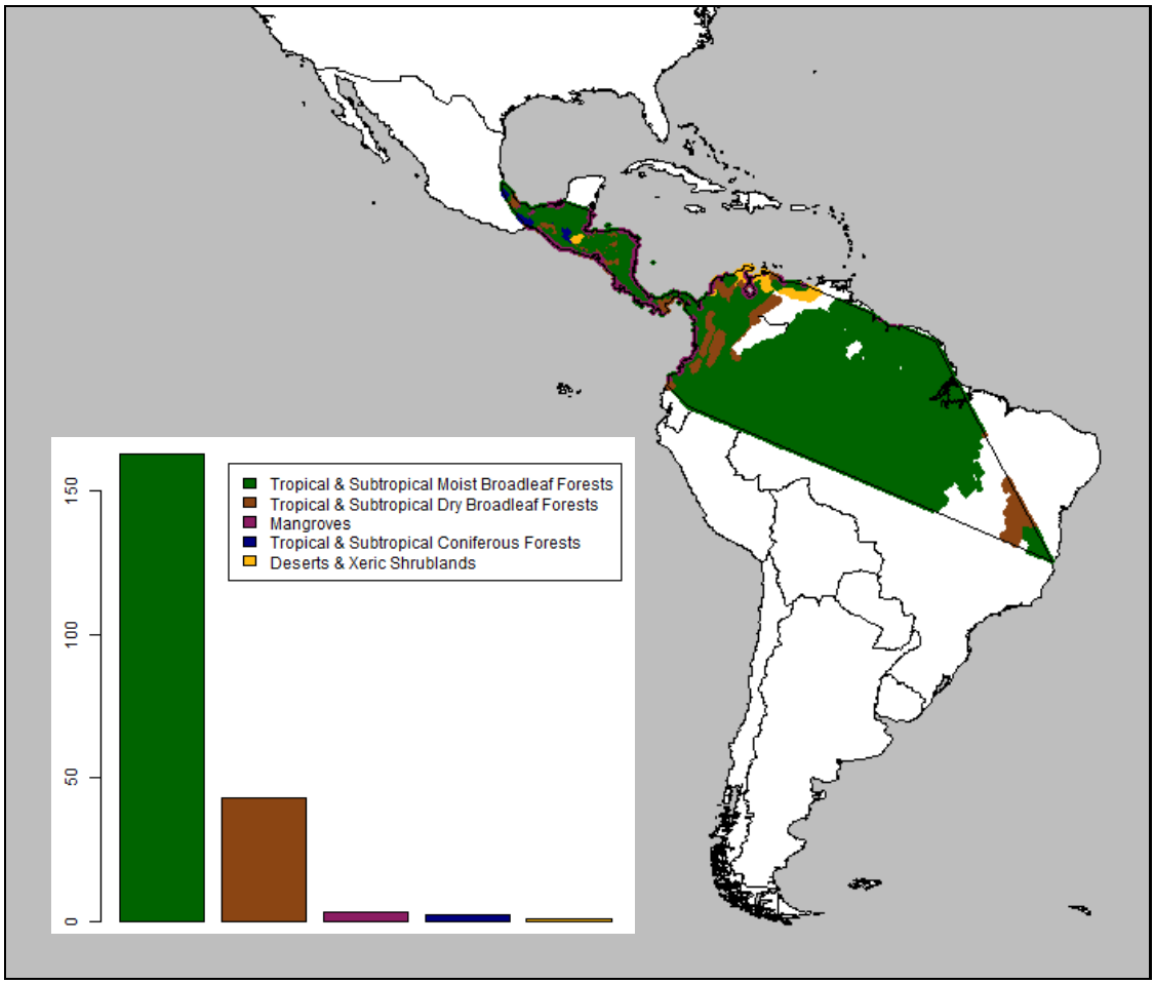


Table 1.3 Biomes in which resulting *V. planifolia* occurrences were located, within the three regions.

	Mexico	Central America	South America	Total
Tropical and subtropical moist broadleaf forests	123	22	18	163
Tropical and subtropical dry broadleaf forests	7	0	36	43
Tropical and subtropical coniferous forests	2	0	0	2
Deserts and xeric shrublands	0	0	1	1
Mangroves	0	3	0	3

Validating taxonomic identifications of biodiversity occurrences

We requested images of 102 specimens from 26 herbaria around the world. After one year, we received images from only 21 specimens. With these images, it became apparent that accurate identification of *V. planifolia* and its relatives relying solely on morphology remains difficult, especially when several species occur in sympatry as emphasized by Soto Arenas & Cribb (2010). Of these 21 herbarium specimens, two were recorded as being similar to *V. planifolia*, two appeared to be identified incorrectly, and the determination of one had been changed to *Vanilla insignis*. Specimens with incorrect or unknown identification, however, did not affect the overall extent of *V. planifolia* occurrence (when investigated at country level). From occurrences outside of previously hypothesized distributions, we verified the correct identification of four herbarium specimens from Colombia, Brazil and French Guiana (Figure 1.3E).

Sampling efforts of plants in South America

The resulting specimen richness map of plant occurrences in South America indicates low specimen richness in the center of the Amazon Basin as compared to areas

bordering this region (Appendix C). This sampling gap was similar to the one observed for *V. planifolia* (Figure 1.3).

Discussion

Mitigating the Wallacean shortfall to propose a new distribution

By incorporating plant life-cycle requirements and habitat quality into our assessment, we aimed to mitigate many of the effects of the Wallacean shortfall on this valuable tropical species. Results from our approach suggest that *V. planifolia* occurs from southern Mexico southward to Brazil (with a sampling gap in the Amazon Basin; Figure 1.3). Thus, the hypothesized distributions of *V. planifolia* inferred based on previously acquired point data (Figure 1.1A, B) were too restrictive and did not account for additional factors contributing to habitat suitability or for more recent occurrence data available on GBIF (Figure 1.2). Moreover, a hypothesized distribution based solely on all available data from GBIF (Figure 1.1C) is obviously too broad, because it does not differentiate between introduced, natural and extirpated populations.

We propose this current landscape-based assessment of *V. planifolia* occurrences to reconcile flaws within large biodiversity datasets and resolve the incongruences of previous distribution hypotheses. Furthermore, we identify the limitations of our approach, such as uncertainties in the precise ecology and taxonomy, that often hinder scientific studies of *V. planifolia* so that they may be better addressed. Additional field work will be required to verify our new hypothesis on the distribution of *V. planifolia*, but our approach has the advantage of identifying locations that are likely to sustain populations of the species due to the presence of pollinators and dispersers as well as the required climatic conditions and the relatively undisturbed habitat (see next).

Conservation strategies, which might have been restricted to the narrow distribution suggested in previous studies, should be reassessed to include populations within this extended range so that the full potential extent of genetic variation in *V. planifolia* may be protected.

Closely related species, many of which have similar life-cycle and habitat requirements to *V. planifolia*, also co-occur within this extended distribution, even along the eastern coast of Brazil (Figure 1.4). Their natural dispersal into this region supports our hypothesis that *V. planifolia* may also occur there. Similarities in morphology and habitat between *V. planifolia* and closely related species provide strong support for these taxa to be included into *V. planifolia* conservation strategies. Crop wild relatives provide an important source of genetic diversity that may help to increase the resilience of crop species in the face of environmental and climatic changes (Jump et al., 2009; Viruel et al., 2020). These genetic resources may be used to identify desirable traits that may be transferred to the crop plant or they can be used to increase genetic diversity through hybridization. Inclusion of this extended distribution and occurrences of the wild relatives of *V. planifolia* into conservation strategies will expand the protection of suitable habitat for these species and the preservation of their genetic variability.

In addition, abiotic factors associated with resulting occurrences support our finding of a larger distribution than previously proposed. Consistent with earlier literature, resulting occurrences were predominantly located in the same biome type (tropical and subtropical moist broadleaf forest; Figure 1.7), and exhibit similar climatic niches (Hernández-Ruíz et al., 2016; Vega et al., 2017; Figure 1.6). Six occurrences were found to be located in uncharacteristic biomes: tropical and subtropical coniferous

forests, deserts and xeric shrublands and mangroves. Although at first glance these biomes seem unsuitable to sustain populations of *V. planifolia*, our analysis indicates that the climatic niches of these occurrences appear to be similar to that of occurrences located in tropical and subtropical moist broadleaf forests (Figure 1.6). These data suggest that biomes do not exhibit an exclusively homogeneous climate; therefore, they may contain microclimates exhibiting habitat suitable for *V. planifolia*.

Results from the climate analysis also provide support for the out-of-Mexico hypothesis for the origin of *V. planifolia*, as proposed by Bory *et al.* (2008). Occurrences located in Mexico exhibit considerable variation in climatic niche and encompass the climatic niches of occurrences found outside of this region (Fig.6). The species, therefore, is likely to have dispersed from pre-adapted populations in Mexico, to Central America and, subsequently, to South America, most likely sometime after the formation of the Isthmus of Panama, *c*. 2.8 Mya (Bacon et al., 2013; O'Dea et al., 2016). Although the occurrence on the east coast of Brazil is located in the tropical and subtropical moist broadleaf forest, along with most resulting occurrences, it is disconnected from other occurrences in South America by one ecoregion called the Cerrado, characterized by open grasslands intermixed with patches of forest (Figure 1.7). Approximately 2.3 Mya during the Quaternary climatic-vegetational fluctuation, the flora of adjacent ecoregions (the Amazon and the Atlantic Forest) expanded into the cerrado (Chaves et al., 2015; Silva, 1995). This expansion may have connected both biomes allowing for dispersal to occur. Although mechanisms of dispersal are unclear, it is important to note the historical importance of this species for many cultures throughout the Americas and also to acknowledge the possibility of multiple dispersal events through pre-Colombian or recent

trade. Phylogeographical analyses, coupled with additional sampling, is required to verify populations in this extended distribution and to further assess where *V. planifolia* may have originated and the major dispersal events, either natural or human-mediated, required to shape its current biogeography.

The results of our analysis suggest that additional regions should be considered in conservation strategies to preserve the phenotypic and genotypic variation within *V. planifolia.* Variation in geography and climate may have led to natural adaptations in the species. Therefore, research on populations, such as those occurring in less frequent biomes and that occurring in Brazil, should be prioritized. Further research and monitoring are needed in regions of South America outside of previously hypothesized distributions, especially in the Amazon Basin, where there is a conspicuous lack of biodiversity data (Appendix C). Therefore, within our resulting distribution, we believe that sampling bias is the main factor affecting the lack of recorded occurrences of *V. planifolia* from the Amazon Basin.

Most resulting occurrences were recorded from living or preserved specimens (Table 1.2). Validation of the taxonomic identification of these specimens additionally provided support for our hypothesis of a larger distribution of *V. planifolia* (Figure 1.3E). Although we were unable to validate the occurrences from multiple countries, all countries containing resulting occurrences of *V. planifolia* included records of preserved specimens. Electronic images of specimens were assessed from four countries (Mexico, Costa Rica, Colombia and Brazil), and correct identification was confirmed. Support, however, was severely limited due the inaccessibility of the majority of specimen images from herbaria (this situation will probably not improve during the worldwide COVID-19

pandemic). Preserved herbarium specimens offer valuable information for botanical and ecological studies but are often not used to their full potential due to challenges such as this one. Although recent efforts to digitize herbarium specimens have substantially increased the accessibility of many specimens, several obstacles still remain that limit the dissemination and use of natural history collections.

The Linnaean shortfall poses additional challenges to the conservation of *Vanilla planifolia*

Morphological similarities between *V. planifolia* and its sympatric wild relatives mandate the need for increased species delimitation efforts in the *V. planifolia* group. Lack of formal taxonomic descriptions of species, known as the Linnaean shortfall, hinders an understanding of species richness and the conservation of biodiversity (Bini et al., 2006). This effect is exhibited among *V. planifolia* and its relatives by the apparent uncertainty in taxonomic descriptions. Species delimitation analyses using modern genomic techniques will help to distinguish taxa in this group and ensure accurate taxonomic descriptions. For instance, the Angiosperms353 bait set for targeted enrichment could be applied to infer species delimitations in the *V. planifolia* group. Such an approach has been effective for other groups (Johnson et al., 2018; Larridon et al., 2020; Murphy et al., 2020) and enables the recovery of a high fraction of genes from historical collections (Brewer et al., 2019). This advantage would allow tapping into old collections deposited in major herbaria (Buerki & Baker, 2016, and references therein).

Conclusions

The new approach presented here to infer current geographical distribution of species is tailored to address the challenges of *V. planifolia* conservation, which have

been exacerbated by climate change and habitat destruction. It accounts for (1) the co-occurrence of species that sustain gene flow within and among populations and (2) the dynamics of landscape usage that can be updated to reflect the most current influence of human-induced disturbances on natural habitats. The development of an evidence-based method to estimate the current distribution of a species such as *V. planifolia* has important and positive societal and ecological implications. First, it necessitates a reassessment and expansion of current conservation strategies to better ensure the survival of *V. planifolia* populations under future landscape conditions. Second, improvements in our understanding of the landscapes in which *V. planifolia* occurs provides insight into the variation in climate and habitat conditions that can support this species and increases knowledge of the community composition and biotic interactions taking place. Third, a revised distribution may help to establish new opportunities for sustainable agriculture of this valuable crop and increase economic development in the regions where it currently occurs. Last, awareness of the current distribution of *V. planifolia* may assist in the discovery of genotypes within the species that can be used for genetic improvements in cultivated varieties through traditional plant breeding. In the Anthropocene, the conservation of vanilla populations in the native range of the species is urgent and essential to ensure the sustainability of global vanilla production, the livelihoods of millions and the future supply of this important spice.

CHAPTER TWO: IDENTIFYING AND CONSERVING THE CROP-WILD RELATIVES OF *THEOBROMA CACAO* AND *VANILLA PLANIFOLIA*

Abstract

Crop wild relatives (CWRs) provide an important source of genetic diversity that may increase the resilience of many crop species and help to ensure future food security. Challenges to characterize CWRs have limited the identification of many species that could potentially hybridize with crops to offer desirable traits and have prevented the implementation of conservation strategies needed to ensure their protection under future climate and habitat conditions. Using two important crops (*Theobroma cacao* and *Vanilla planifolia*) as a model, this review aims to identify CWRs and assess their respective conservation priority through the integration of available data on hybridization, genetics (DNA sequences), cytogenetics (chromosome counts and genome size), occurrence, and human influence on the landscape. Overall, there is surprisingly very little known of the relatives of these two important crops; only 17% and 24% of candidate CWRs of *T. cacao* and *V. planifolia* were represented among ITS accessions on GenBank. Using available data, results identify seven *T. cacao* CWRs and 17 *V. planifolia* CWRs that should be categorized as GP-2 relatives, meaning that they may be cross-compatible with their crop relative. Preliminary assessments of conservation status from the area of occurrence (AOO) of these identified species reveals that 69% of all *T. cacao* candidate CWRs and 67% of all *V. planifolia* candidate CWRs are categorized as endangered.

Introduction

Domestication processes, dating back to 10,000 years ago, have shaped all major crops that are consumed today (Doebley et al., 2006). Beginning with the wild collection of fruits and seeds, individual plants that exhibited the most desirable phenotypes were selected to form the next generation of cultivated plants. These processes increased the frequency of desirable phenotypes within cultivated populations and over time decreased the levels of genetic variation through genetic bottlenecks due to population size reductions (Doebley, 1989; Tanksley & McCouch, 1997). For each crop, the extent of its genetic bottleneck differs depending on various factors, such as the originating population size, the duration of the domestication period, the number of domestication events, and the species' reproductive biology (Eyre-Walker et al., 1998). Compared to their wild progenitors, crops today exhibit decreased genetic diversity and have often lost traits such as disease resistance and drought tolerance (Hyten et al., 2006; Koziol et al., 2012; Rosenthal & Dirzo, 1997).

Climate change is a major threat to global biodiversity and the global food system, especially for domesticated crops with limited genetic diversity. Changing regional conditions result in decreased crop productivity as they become less suited to their cultivated environment (Jägermeyr et al., 2021). Biodiversity conservation helps provide functioning ecosystems that supply humans with necessary resources. As one component of biodiversity, plant genetic resources (PGR), the genetic material of plants which are of present or future value, acts as a foundation for agricultural development by providing avenues for genetic adaptability and future crop sustainability. The human population is expected to reach 9.7 billion by 2050 (*Population | United Nations*, n.d.).

To feed such an immense population, it is an imperative and urgent task to utilize and conserve the resources necessary to ensure diversity within crops. In this epoch of unprecedented biodiversity loss (Brondizo et al., 2019), however, many of the related species that contain the genetic resources need to improve crops may be at risk of extinction.

Wild ancestors or close relatives of domesticated plants, called crop wild relatives (CWRs), provide an important source of genetic diversity that may help to increase the resilience of many crop species in the face of environmental changes (Jump et al., 2009). They have been used successfully to introduce traits that increase tolerances to biotic and abiotic stresses. Conventional breeding is typically used for interspecific crossing between wild and crop species, but synthetic hybridization is now often supplemented with newer processes such as marker-assisted selection, embryo rescue, and chromosomal manipulation (Ford-Lloyd et al., 2011; Katche et al., 2019). For example, through synthetic hybridization and backcrossing, the wild potato *Solanum demissum* Lindl. was used to overcome the late blight (Black, 1970); wild wheat, *Aegilops tauschii* Coss., was used for its resistance to stem rot (Kilian et al., 2010); and a wild rice relative, *Porteresia coarctata* (Roxb.) Tateoka, has been used to increase salt tolerance (Majee et al., 2004). To ensure future food security, it is essential to understand and protect the resources of genetic variability provided by crop wild relatives.

Challenges to characterize CWRs

Proposing a definition that bridges both theory and application for what constitutes a CWR has been a challenging feat. Two main concepts have traditionally been used to define and identify CWRs: the gene pool concept and the taxon group

concept. As described by Harlan & Wet (1971), the gene pool concept aims to identify and categorize CWRs based on their biological species concepts, i.e. their ability to reproduce with the target crop species to produce fertile offspring. Within this system, species are categorized into three gene pools: the primary gene pool, GP-1, consisting of the crop itself along with wild types that will cross to produce fertile offspring; the secondary gene pool, GP-2, consisting of distinct species that could potentially overcome reproductive barriers (e.g. geographic isolation), to cross with the crop species; and the tertiary gene pool, GP-3, consisting of relatives that are unable to be crossed with the crop species to produce viable progeny. Where interspecific crossing experiments have been conducted and hybridization data is available, the gene pool concept provides an efficient and valid classification system for CWRs. This is, however, far from the case for the majority of species. To mitigate this lack of hybridization data, Maxted et al. (2006) proposed the taxon group concept to identify CWRs based on their taxonomic species concept. This system of classification categorized CWRs into five taxon groups: the first consisting of the cultivated and the wild form of the crop; the second consisting of species within the same taxonomic section, the third consisting of species within the same subgenus, the fourth consisting of species within the same genus, and the fifth consisting of species within the same tribe as the crop. Within this system, species belonging to taxonomic groups 1-4 are considered CWRs.

While there may exist more taxonomic data than cross-compatibility data for global plant species, the taxon group concept is still limited due to incomplete or flawed taxonomic data. Between 10 to 20 percent of plants have yet to be described (Joppa et al., 2011). In addition, many taxonomic classifications only reflect a categorization based on

morphological similarities or the historical taxonomic treatment of the group, therefore often do not represent true evolutionary histories and phylogenetic relationships. Unresolved taxonomic data or data that only incorporates morphology may inhibit the identification of cross-compatible species. As genetic methods become more advanced and affordable, phylogenetic evidence from DNA data may provide insights for taxonomic revision and further facilitate the identification of CWRs through the taxon group concept.

To account for these recent technological advances and datasets, Viruel et al. (2020) have proposed a new system to identify and classify CWRs based on their phylogenetic species concepts. Combining data on cross-compatibility, phylogenetic distance, cytogenetic compatibility, and breeding systems, they aim to predict inter-specific cross-compatibility and identify "crop wild phylorelatives". In accordance with the gene pool concept, this system categorize species as GP-1, GP-2, and GP-3, where GP-1 consists of the crop species, GP-2 consists of species that may be used to potentially hybridize with the crop species, and GP-3 consists of species that are not cross-compatible. To categorize species into gene pools, three levels of criteria are assessed. First, whether or not hybridization experiments have been conducted and cross-compatibility has been successful. In this case, species are categorized as GP-1 or GP-2, depending on whether the species is the same as the crop and hybridization is successful, or not. Where hybridization data is not available for a species, DNA sequence data is used to predict cross-compatibility. Phylogenetic distances, calculated using patristic distances of ultrametric trees, are compared between the species and crop to assess relatedness. Distantly related species are predicted to be unable to hybridize with the crop

and therefore placed into GP-3. For closely related species, additional data is collected on ploidy level and breeding system. Closely related species with similar cytogenetics and breeding systems are then categorized as GP-2 and are predicted to be able to hybridize with the crop. Although sequence data may not be available for many species, this system overcomes many of the challenges of previous CWR concepts to propose a heuristic pipeline to identify crop wild phylorelatives and predict interspecific cross-compatibility.

Challenges to conserve CWRs

In the Anthropocene, human activities have directly and indirectly caused unprecedented global changes in biodiversity (Ellis et al., 2012). For instance, in the Amazon basin, 58% of tree species are expected to go extinct in the next 30 years due to deforestation and climate change (Gomes et al., 2019). In addition to species extinctions; population declines, extirpations, distribution shifts and an increase in invasive species pose a major threat to the stability of biological communities and the preservation of CWR genetic resources. In a study on CWRs in South America, Jarvis et al. (2008) predicted that almost half of the current ranges of CWRs of peanuts in South America, cowpeas in Africa, and potatoes in Central and South America will be lost, and that 16% to 22% of these species would go extinct by 2055. Another analysis using bioclimatic modeling gravely revealed that the distributions of eight CWRs in the Cucurbitaceae family will severely contract under future climate scenarios and most taxa are unlikely to survive (Lira et al., 2009).

Assessing threats to CWRs and their habitat is of utmost importance to the conservation of their genetic diversity. The IUCN Red List of Threatened Species provides a critical resource assessing the global extinction risk status of biodiversity,

including CWRs (IUCN, 2019). However, most CWRs are lacking assessments. From an extensive yet incomprehensive inventory of CWR taxa provided by https://www.cwrdiversity.org, only around 20% of taxa have been assessed through the IUCN and of these 23% were categorized as being Data Deficient (DD) (Viruel et al., 2020). As deforestation, habitat fragmentation, and climate change continue to affect global biodiversity, there exists an urgent need to identify CWRs for a wide range of crop plants and assess their habitat requirements and risks of extinction.

Objectives

Within this review, I aim to infer the identification of CWRs and assess their respective conservation priority through the integration of available data on hybridization, genetic diversity (DNA sequences), cytogenetics (chromosome counts and genome size), occurrence, and human influence on the landscape. I use an adapted version of the pipeline of Viruel et al. (2020) to identify crop wild phylorelatives by taking advantage of the large resource of publicly available genomic data provided by GenBank to compare genetic distances of crops and their relatives. The GenBank database contains DNA sequences for over 105,000 different species and is growing at an exponential rate (Benson et al., 2002). Within this database, standardized DNA regions (also known as DNA barcodes; (Hollingsworth et al., 2009) may be used to assess differences between target species and act as a proxy for cross-compatibility following the methods of Viruel et al. (2020). Using DNA barcoding as a tool to identify species and assess evolutionary relationships among plants, however, has inherent challenges due to factors such as: lack of a universal gene region, low sequence variability within plastid genomes, inability to infer hybridization processes using plastid genomes, horizontal

gene transfer, hybridization, and homoplasy. Nonetheless, DNA barcoding has been used extensively to identify species and address many ecological, evolutionary and conservation issues. For most applications involving plants, standard DNA barcodes include: *rbcL, matK*, *trnh-psbA*, and nuclear ribosomal ITS region (Hollingsworth et al., 2009; X. Li et al., 2015).

I use two economically important crop species, *Theobroma cacao* L. and *Vanilla planifolia* Andrews, as models for this approach. Both crops have conflicting hypotheses explaining their origin in either Central or South America and have a long history of domestication in both regions. Insight into the identification and vulnerability of their crop wild relatives will help to guide research into cross-compatibility and prioritize conservation strategies to protect these important genetic resources.

The intertwined history of chocolate and vanilla

Although chocolate and vanilla are two of the most well-known and beloved flavors, there is surprisingly very little known of their genetic resources especially their crop-wild relatives. Chocolate is made from the seeds of the fruit of *Theobroma cacao*, a tree native to Central and South America (Thomas et al., 2012). Vanilla is produced from the cured seed pod of *Vanilla planifolia*, a tropical climbing orchid native to Mexico and Central and South America (Ellestad et al., 2021). It has been reported that both crops were first domesticated in Mexico but may have origins in other regions of Central and South America (Bruman, 1948; Lubinsky, Bory, et al., 2008; Rain, 2004; van Hall, 1914). Together, they were used by the Mayans to flavor a ceremonial beverage called “choclatl’ (Rain, 2004). Both were traded among the Mayans and the Aztecs, and after

the Spanish conquered the Aztecs in the 1500's, they were transported to Europe. Today, *T. cacao* and *V. planifolia* are cultivated in tropical regions around the world.

Vanilla planifolia

Vanilla planifolia is a tropical vine in the family Orchidaceae naturally occurring in Mexico and Central and South America (Bruman, 1948; Ellestad et al., 2021). Although *V. planifolia* is self-fertile, it is not self-compatible because its rostellum prevents contact between the stamen and stigma. This floral structure favors outcrossing, however, vegetative propagation is also common. For natural sexual reproduction to occur, it must be fertilized by a pollinator (Bory, Grisoni, et al., 2008). In the wild, little is known of pollinator interactions, however, some research suggests a relationship with bee from the genera *Eulaema* and *Euglossa* (tribe Euglossini; Apidae) serve as pollinators (Ackerman, 1983; Rodolphe et al., 2012; Soto Arenas & Dressler, 2010). Usually *V. planifolia* is diploid (2n = 2x = 32), however recent polyploidization has been recorded among cultivated individuals exhibiting a variation in chromosome numbers from n = 16 to 54. Additionally, genome size has been shown to vary from around 5 pg to around 10 pg (2C-value) (Donini et al., 2008).

The genus *Vanilla* consists of over 100 pantropical species (Soto Arenas & Cribb, 2010). *Vanilla planifolia* is the predominant source of global vanilla, however, three other species have also been used in cultivation: *V. tahitensis, V. pompona,* and *V. insignis* (Ellestad, Pérez-Farrera, Forest, et al., 2022). Following the most recent taxonomic revision of Soto Arenas & Cribb (2010), the genus *Vanilla* is formed by two subgenera: *Vanilla* and *Xanata*, of which the subgenus *Xanata* is comprised of two sections: *Xanata* and *Tethya*. The predominantly cultivated species, *V. planifolia,* belongs

to the subgenus *Xanata* in section *Xanata* and the group *V. planifolia*, along with 16 other morphologically similar species as described by Soto Arenas & Cribb (2010).

Hybridization has been recorded between *V. planifolia* and six congeneric species: *V. pompona* (Delassus, 1963), *V. aphylla* (Divakaran et al., 2006), *V. odorata, V. tahitensis* (Lubinsky, Cameron, et al., 2008), *V. phaeantha* (Y. Hu et al., 2019), and *V. palmarum* (J. Li et al., 2020). All species proven to hybridize with *V. planifolia* belong to the subgenus *Xanata*. Of these species, *V. pompona, V. insignis, V. phaeantha* and *V. odorata* co-occur with *V. planiofolia* in Central America and northern South America. One species, *V. aphylla*, has a geographically distant distribution in Southeast Asia (*Plants of the World Online | Kew Science*, n.d.). *Vanilla tahitensis* occurs only in cultivated and feral stands in Papua New Guinea and French Polynesia. It is hypothesized to be the product of recent human-mediated hybridization events between *V. planifolia* and *V. odorata* (Lubinsky, Cameron, et al., 2008).

Theobroma cacao

As a small understory tree, *T. cacao* grows naturally in Neotropical lowland forests. It is highly outcrossing and mostly self-incompatible, although self-compatibility has been recorded among wild and cultivated individuals (Chumacero de Schawe et al., 2013). Hermaphroditic flowers grow on its trunk and are pollinated almost exclusively by insects, particularly midges in the family Ceratopogonidae (Claus et al., 2018). Usually, flower production peaks twice annually after increases in precipitation and temperature (Claus et al., 2018; Lahive et al., 2019). *Theobroma cacao* is diploid (2n = 2x= 20) and has a 1C genome size estimated to be between 0.40 pg and 0.43 pg (Figueira et al., 2019; Lanaud et al., 1992).

The genus *Theobroma* consists of 22 species distributed in Neotropical lowland forests from southern Mexico to the Amazon basin (Cuatrecasas, 1964). Within the family Malvaceae, it belongs to the tribe Theobromeae along with 3 other genera: *Herrania* Goudot (18 species), *Glossostemon* Desf. (1 species), and *Guazuma* Mill.(3 species) (Richardson et al., 2015). All occur in parts of Central and South America except for the genus *Glossostemon*, which occurs in the Middle East (*Plants of the World Online | Kew Science*, n.d.). Of the 47 species in tribe Theobromeae, only *T. cacao* and *T. grandiflorum* have been cultivated: *T. cacao* provides the raw material for the production of chocolate as well as some cosmetics and pharmaceuticals, and the pulp of *T. grandiflorum* is used for flavoring and cosmetics. Studies have shown that these two species have the same number and similar structure of chromosomes, as well as similar genome sizes (Azevedo da Silva et al., 2017). Hybridization between the species has been reported, however, it has not resulted in viable hybrids (Martinson, 1966).

Methods

To facilitate CWR identification and assess their respective conservation priority, candidate CWR were assessed using data on hybridization, genetic distances, and cytogenetics. As with the taxon group concept, taxonomic information was used as a heuristic attempt to fill in the gaps of phylogenetic understanding and identify a wide array of candidate CWRs which may then be assessed for cross-compatibility. Where hybridization data were available (see above), the next higher taxonomic level (section, subgenus, genus, tribe) were used to create a list of candidate species. For *T. cacao,* hybridization data were available for *T. grandiflorum*, within the same genus, therefore,

all taxa within the tribe Theobromateae were included as candidate CWRs. For *V. planifolia*, hybridization data were available for six species within the same subgenus *Xanata*, therefore, all species within the entire genus were included as candidate CWRs (Table 2.1). Identified CWRs, which were more closely related than species recorded to hybridize successfully and had similar cytogenetic characteristics, were predicted to be cross-compatible with their respective crop. Their occurrences (downloaded from GBIF on 10 October 2020) were used to obtain data on their distributions and risk of global extinction. Occurrence locations were used to identify regions of high CWR biodiversity, assess species' extent of occurrences (EOO) and area of occurrences (AOO) and to understand the influence of human activities on their landscapes.

Table 2.1 Cross-compatibility references for *Vanilla* species. No cross-compatibility data was available for *Theobroma* species nor relatives.

Cross-compatibility		
Crop	**Relative**	**Reference**
Vanilla planifolia	*Vanilla aphylla*	Minoo et al. 2006b
	Vanilla odorata	Lubinsky et al 2008
	Vanilla phaeantha	Hu et al 2019, Li et al 2020
	Vanilla pompona	Delassus 1960; Dequaire 1976; FOFIFA 1990, bory et al 2008
	Vanilla palmarum	Li et al 2020
	Vanilla tahitensis	Delassus 1960; Dequaire 1976; FOFIFA 1990, bory et al 2008

Identifying CWRs using DNA sequences

To identify and categorize CWRs, pairwise K2P genetic distances were calculated for related species with available sequence data on GenBank. Of the four most commonly used plant DNA barcodes (*rbcL, matK, trnH-psbA*, and ITS), the biparentally inherited nuclear ribosomal ITS genetic marker was used due to its level of polymorphism supporting superior resolution of inter- and intra- specific relationships compared to plastid markers, *rbcL, matK,* and *trnH-psbA* (Cheng et al., 2015). Additionally, this

nuclear region was chosen so that it may potentially reflect processes of natural hybridization that has been shown to occur among sympatric taxa. Due to the uncertainty of taxonomic groupings within *V. planifolia* (Soto Arenas & Cribb, 2010), candidate CWRs included all 110 species that comprise the genus *Vanilla.,* including the large subgenus *Xanata* and the smaller subgenus *Vanilla*. For *T. cacao*, candidate CWRs included the 47 species within the genus *Theobroma* and its three sister genera, *Herrania, Guazuma,* and *Glossostemon* (Table 2.2; Richardson et al., 2015). A search for ITS sequences on GenBank for the related species and crops was conducted using the *rentrez* package in R (D. J. Winter, 2017). For each crop, resulting sequences were aligned using AliView (Larsson, 2014) and pairwise genetic distances were calculated using Kimura's 2-parameter (K2P) model as implemented in the dist.dna function within the 'ape' R package (Paradis & Schliep, 2019). Inter- and intra-specific variation of ITS sequences were assessed for taxa with more than one DNA accession and plotted using ridgeline plots (the R code associated to these analyses is available on GitHub: https://github.com/svenbuerki/RidgelinesCWRdist).

Table 2.2 List of candidate CWR taxa for *T. cacao* and *V. planifolia*.

Vanilla planifolia		***Theobroma cacao***	
Subgenus	**Species**	**Genus**	**Species**
Vanilla	*Vanilla angustipetala*	*Theobroma*	*Theobroma angustifolium*
Vanilla	*Vanilla bertoniensis*	*Theobroma*	*Theobroma bernoullii*
Vanilla	*Vanilla bradei*	*Theobroma*	*Theobroma bernoullii*
Vanilla	*Vanilla costaricensis*	*Theobroma*	*Theobroma bernoullii*
Vanilla	*Vanilla guianensis*	*Theobroma*	*Theobroma bicolor*
Vanilla	*Vanilla martinezii*	*Theobroma*	*Theobroma cacao*
Vanilla	*Vanilla methonica*	*Theobroma*	*Theobroma canumanense*
Vanilla	*Vanilla organensis*	*Theobroma*	*Theobroma cirmolinae*
Vanilla	*Vanilla oroana*	*Theobroma*	*Theobroma duckei*
Vanilla	*Vanilla ovata*	*Theobroma*	*Theobroma gileri*
Vanilla	*Vanilla parvifolia*	*Theobroma*	*Theobroma glaucum*

Vanilla	*Vanilla edwallii*	*Theobroma*	*Theobroma grandiflorum*
Vanilla	*Vanilla inodora*	*Theobroma*	*Theobroma hylaeum*
Vanilla	*Vanilla mexicana*	*Theobroma*	*Theobroma mammosum*
Xanata	*Vanilla abundiflora*	*Theobroma*	*Theobroma microcarpum*
Xanata	*Vanilla acuminata*	*Theobroma*	*Theobroma nemorale*
Xanata	*Vanilla africana*	*Theobroma*	*Theobroma obovatum*
Xanata	*Vanilla aphylla*	*Theobroma*	*Theobroma simiarum*
Xanata	*Vanilla appendiculata*	*Theobroma*	*Theobroma sinuosum*
Xanata	*Vanilla bahiana*	*Theobroma*	*Theobroma speciosum*
Xanata	*Vanilla barbellata*	*Theobroma*	*Theobroma stipulatum*
Xanata	*Vanilla bicolor*	*Theobroma*	*Theobroma subincanum*
Xanata	*Vanilla calopogon*	*Theobroma*	*Theobroma sylvestre*
Xanata	*Vanilla calyculata*	*Theobroma*	*Theobroma velutinum*
Xanata	*Vanilla chalotii*	*Herrania*	*Herrania albiflora*
Xanata	*Vanilla cribbiana*	*Herrania*	*Herrania amazonica*
Xanata	*Vanilla columbiana*	*Herrania*	*Herrania balaensis*
Xanata	*Vanilla coursii*	*Herrania*	*Herrania breviligulata*
Xanata	*Vanilla crenulata*	*Herrania*	*Herrania camargoana*
Xanata	*Vanilla dressleri*	*Herrania*	*Herrania cuatrecasana*
Xanata	*Vanilla cristagalli*	*Herrania*	*Herrania dugandii*
Xanata	*Vanilla cucullata*	*Herrania*	*Herrania kanukuensis*
Xanata	*Vanilla decaryana*	*Herrania*	*Herrania kofanorum*
Xanata	*Vanilla dilloniana*	*Herrania*	*Herrania laciniifolia*
Xanata	*Vanilla dubia*	*Herrania*	*Herrania lemniscata*
Xanata	*Vanilla dungsii*	*Herrania*	*Herrania nitida*
Xanata	*Vanilla hartii*	*Herrania*	*Herrania nitida*
Xanata	*Vanilla espondae*	*Herrania*	*Herrania nycterodendron*
Xanata	*Vanilla fimbriata*	*Herrania*	*Herrania pulcherrima*
Xanata	*Vanilla francoisii*	*Herrania*	*Herrania purpurea*
Xanata	*Vanilla gardneri*	*Herrania*	*Herrania tomentella*
Xanata	*Vanilla giulianettii*	*Herrania*	*Herrania umbratica*
Xanata	*Vanilla grandifolia*	*Glossostemon*	*Glossostemon bruguieri*
Xanata	*Vanilla helleri*	*Guazuma*	*Guazuma crinita*
Xanata	*Vanilla hallei*	*Guazuma*	*Guazuma longipedicellata*
Xanata	*Vanilla imperialis*	*Guazuma*	*Guazuma ulmifolia*
Xanata	*Vanilla insignis*	*Theobroma*	*Theobroma angustifolium*
Xanata	*Vanilla heterolopha*	*Theobroma*	*Theobroma bernoullii*
Xanata	*Vanilla hostmannii*		
Xanata	*Vanilla humblotii*		
Xanata	*Vanilla kinabaluensis*		
Xanata	*Vanilla odorata*		

Xanata	*Vanilla phaeantha*	
Xanata	*Vanilla kaniensis*	
Xanata	*Vanilla kempteriana*	
Xanata	*Vanilla pompona*	
Xanata	*Vanilla pompona*	
Xanata	*Vanilla nigerica*	
Xanata	*Vanilla ochyrae*	
Xanata	*Vanilla tahitensis*	
Xanata	*Vanilla ovalis*	
Xanata	*Vanilla palmarum*	
Xanata	*Vanilla penicillata*	
Xanata	*Vanilla perrieri*	
Xanata	*Vanilla trigonocarpa*	
Xanata	*Vanilla phalaenopsis*	
Xanata	*Vanilla planifolia*	
Xanata	*Vanilla platyphylla*	
Xanata	*Vanilla poitaei*	
Xanata	*Vanilla polylepis*	
Xanata	*Vanilla claviculata*	
Xanata	*Vanilla pompona*	
Xanata	*Vanilla ramificans*	
Xanata	*Vanilla ramosa*	
Xanata	*Vanilla ribeiroi*	
Xanata	*Vanilla savannarum*	
Xanata	*Vanilla schwackeana*	
Xanata	*Vanilla seranica*	
Xanata	*Vanilla seretii*	
Xanata	*Vanilla madagascariensis*	
Xanata	*Vanilla sprucei*	
Xanata	*Vanilla sumatrana*	
Xanata	*Vanilla roscheri*	
Xanata	*Vanilla siamensis*	
Xanata	*Vanilla utteridgei*	
Xanata	*Vanilla vellozoi*	
Xanata	*Vanilla walkeriae*	
Xanata	*Vanilla wariensis*	
Xanata	*Vanilla wightii*	
Xanata	*Vanilla zanzibarica*	
	Vanilla albida	
	Vanilla andamanica	
	Vanilla annamica	

	Vanilla borneensis	
	Vanilla chamissonis	
	Vanilla diabolica	
	Vanilla dietschiana	
	Vanilla griffithii	
	Vanilla hamata	
	Vanilla havilandii	
	Vanilla somae	
	Vanilla montana	
	Vanilla moonii	
	Vanilla palembanica	
	Vanilla ruiziana	
	Vanilla sanjappae	
	Vanilla sarapiquensis	
	Vanilla yersiniana	

Genomic and cytogenetic data

To further investigate hybridization possibility between the crop and its wild relatives, genomic and cytogenetic data were assessed for similarity and used to proxy cross-compatibility. Differences in parental chromosome number or structure causing failed pairing during meiosis or genetic incompatibilities between genomes, such as polyploidization, may cause pre-zygotic barriers inhibiting interspecific reproduction (Sbilordo et al., 2012). Genome sizes for available species were accessed from the Genomic Plant DNA C-values Database (https://cvalues.science.kew.org/) and chromosome numbers were accessed from the Chromosome Counts Database (http://ccdb.tau.ac.il/home/).

Species conservation assessments

All occurrence data associated with preserved specimens were downloaded from the Global Biodiversity Information Facility (GBIF) for each candidate CWR. Although many species lacked enough data to infer cross-compatibility, preliminary conservation

assessments were conducted on all taxa (except cultivated species) to reinforce the need for protection of plant genetic resources and enable future genomic and hybridization research. The predominant cultivated species (*T. cacao, T. grandiflorum, V. planifolia,* and *V. tahitensis)* were not included due to the difficulty of determining whether occurrences represented cultivated or wild individuals. For remaining species, datasets were cleaned; all occurrences without geographical coordinates and those which occurred in oceans were removed. Occurrence coordinates were then transformed into spatial points using the *SpatialPoints* function in the R package *rgeos* (Bivand & Rundel, 2019). To assess the threat status of taxa, the extent of occurrence (EOO) and area of occupancy (AOO) were calculated for each species, as implemented by the IUCN (Le Breton et al., 2019). The *ConBatch* function in the *rCAT* package was used, with a grid cell size of 2 km (as recommended by the IUCN), to calculate EOO and AOO (Moat, 2017).

Habitat vulnerability and protection status was also assessed using data on protected areas and the Global Human Influence Index (Dinerstein et al., 2017; WCS & University, 2005). The Human Influence Index (HII) was used as a proxy for the current impact of human activities on the natural landscape. The HII is a global dataset that incorporates human population pressure, land use and infrastructure, and access to formulate an index between 0 and 100, where 0 indicates no human disturbance and 100 indicates total disturbance (WCS & University, 2005). Using the R package *raster*, HII values were extracted for each 1-kilometer geographic cell that contained an occurrence (Hijmans, 2019a).

Identifying biodiversity hotspots for CWRs

GBIF occurrence data of CWR taxa were used to build a species richness map. The R package *raster* was used to create a blank raster with a resolution of 1.50 degrees and to record the number of species per cell based on the cleaned GBIF dataset (Hijmans, 2019a). See Buerki et al. (2015) for more details on producing the species richness map. This map served to identify regions of high CWR biodiversity.

Results

Identifying CWRs using DNA sequences

From the GenBank search for ITS sequences of the 47 species related to *T. cacao*, DNA sequences from 29 accessions encompassing nine species were downloaded (Table 2.3). Intra-specific K2P genetic distances were calculated for species with multiple accessions (*T. cacao, T. speciosum, T. microcarpum, T. subincanum, T. obovatum,* and *T. bicolor*) revealing varying levels of genetic variation within all taxa; *T. cacao* exhibited the highest number of polymorphism and *T. bicolor* the lowest (Figure 2.1). Intraspecific genetic distances among *T. cacao* accessions ranged from 0 to 0.51. Interspecific genetic distances (between *T. cacao* and relatives) ranged from 0 to 0.50. No successful cross-compatibility studies have been reported between *T. cacao* and other species, therefore, the genetic distance of 0.51 was used as the threshold to determine CWR levels (GP-1, GP-2, and GP-3) and used as a proxy for cross-compatibility with *T. cacao*. Genetic distances of all species analyzed were less than this threshold, therefore, all were categorized as GP-2, indicating that cross-compatibility may be possible.

Table 2.3 Levels of crop wild relatives based on relative species with available ITS data on GenBank.

Crop	Relative	CWR Level
Theobroma cacao	*Theobroma bicolor*	GP-2
	Theobroma grandiflorum	GP-2
	Theobroma microcarpum	GP-2
	Theobroma obovatum	GP-2
	Theobroma speciosum	GP-2
	Theobroma subincanum	GP-2
	Theobroma sylvestre	GP-2
	Guazuma ulmifolia	GP-2
V. planifolia	*Vanilla bahiana*	GP-2
	Vanilla calyculata	GP-2
	Vanilla cribbiana	GP-2
	Vanilla dressleri	GP-2
	Vanilla hartii	GP-2
	Vanilla helleri	GP-2
	Vanilla imperialis	GP-2
	Vanilla insignis	GP-2
	Vanilla kinabaluensis	GP-2
	Vanilla odorata	GP-2
	Vanilla phaeantha	GP-2
	Vanilla pompona	GP-2
	Vanilla tahitensis	GP-2
	Vanilla trigonocarpa	GP-2
	Vanilla griffithi	GP-2
	Vanilla somae	GP-3
	Vanilla edwallii	GP-3
	Vanilla inodora	GP-3
	Vanilla Mexicana	GP-3
	Vanilla Africana	GP-3
	Vanilla barbellata	GP-3
	Vanilla claviculata	GP-3
	Vanilla madagascariensis	GP-3
	Vanilla roscheri	GP-3
	Vanilla siamensis	GP-3
	Vanilla borneensis	GP-3

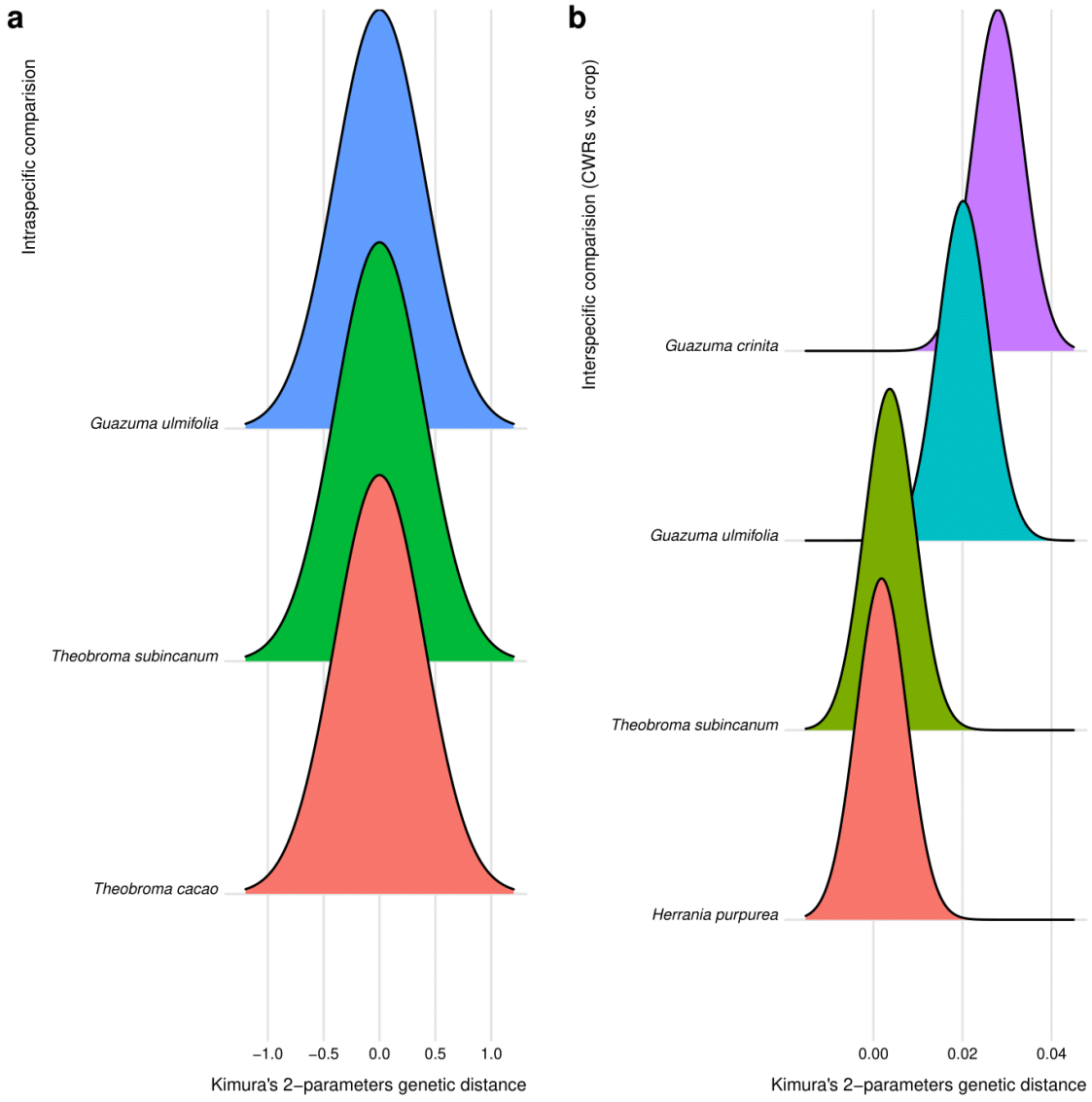


Figure 2.1 On the left are density plots illustrating intraspecific K2P genetic distances using ITS sequence data of candidate CWRs of *T. cacao* with multiple GenBank accessions. *Theobroma cacao* exhibited the highest range of intraspecific genetic distances. On the right are density plots illustrating interspecific (between *T. cacao* and relatives) K2P genetic distances using ITS sequence data of candidate all CWRs of *T. cacao* with available GenBank data. Candidate CWRs exhibited interspecific genetic distances within the range of 0 to 0.52.

From the GenBank search for ITS sequences of the 110 species related to *V. planifolia,* sequences from 115 accessions incorporating 32 species were downloaded (Table 2.3). Intra-specific genetic distances were calculated for 19 species with multiple accessions revealing varying levels of polymorphism within all taxa except *V. cribbiana* and *V. trigonocarpa* (Figure 2.2). Intraspecific genetic distances among *V. planifolia* accessions ranged from 0 to 0.07. Interspecific genetic distances (between *V. planifolia* and relatives) ranged from 0 to 0.41 (Figure 2.2). Successful hybridization studies have been carried out between *V. planifolia* and five other species: *V. odorata, V. pompona, V. tahitensis, V. phaeantha,* and *V. aphylla* (Table 2.2). Of these species, the highest genetic distance was used as the threshold to determine CWR levels (GP-1, GP-2, and GP-3) and used as a proxy for cross-compatibility with *V. planifolia*. From the analyzed taxa, 20 species with a mean pairwise distance of less than that of *V. aphylla* (0.16) were categorized as GP-2, indicating that cross-compatibility may be possible. Ten species had sequences with mean genetic distances higher than this threshold, therefore, they were

tentatively categorized as GP-3; not cross-compatible.

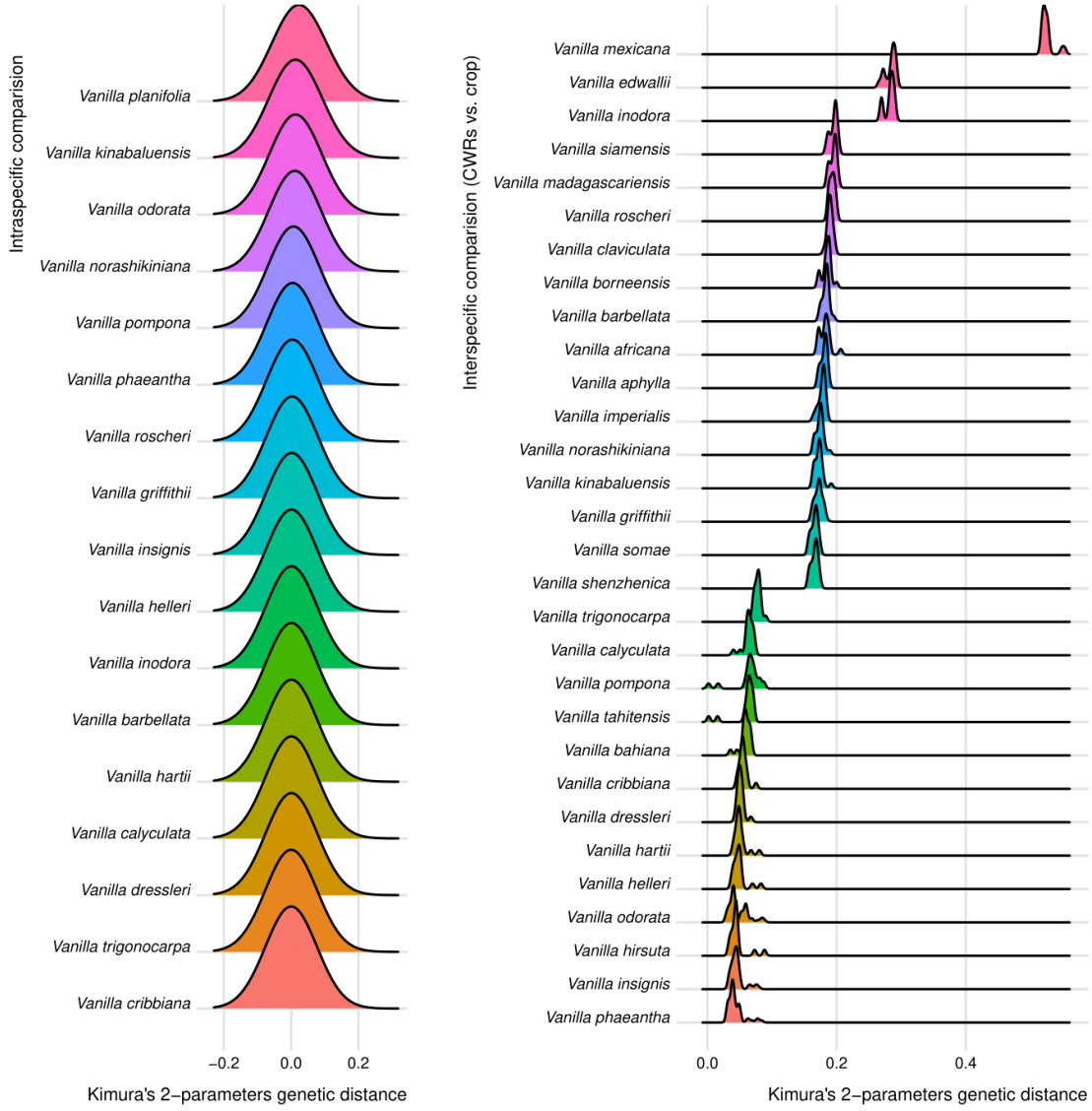


Figure 2.2 On the left are density plots illustrating intraspecific K2P genetic distances using ITS sequence data of candidate CWRs of *V. planifolia* with multiple GenBank accessions. On the right are density plots illustrating interspecific (between *V. planifolia* and relatives) K2P genetic distances using ITS sequence data of candidate CWRs with available GenBank data.

Genomic and cytogenetic data

Data on genome size and chromosome numbers were acquired for all available relative species. For *T. cacao* (2n = 2x= 20), relatives with available data showed slight

variation among genome sizes and many similarities among haploid chromosome counts (Table 2.4). Cultivated *T. cacao* is reported to have a base chromosome number of 10, along with *T. grandiflorum, T. angustifolium*, *T. bicolor, H. albiflora,* and *H. purpurea* (Table 2.5). The small genome sizes of these species may be correlated with the overall stability of karyotypes within this group. Karyotypes of these species were found to have small, symmetric, diploid chromosomes. The only differences were reported between *T. cacao* and *T. grandiflorum,* where small heterochromatic bands were observed on the centromere/pericentromere of *T. cacao* chromosomes after staining, while none were observed on *T. grandiflorum* chromosomes (Azevedo da Silva et al., 2017). Of the species with available cytogenetic data, only *G. ulmifolia* differed in base chromosome number (n = 8; Table 2.5).

Table 2.4 DNA amount in the un-replicated gametic nucleus of crop wild relatives with available data from the Plant DNA C-Values database: https://cvalues.science.kew.org/.

Crop	Relative	Phenotype	Chrom # (2n)	Ploidy Level (x)	DNA Amount 1C (pg)	Reference
Theobroma cacao	*Theobroma cacao*		20	2	0.4	Figueira et al., 2019
	Theobroma grandiflorum				0.51	Argout et al., 2011
	Theobroma microcarpum				0.37	Argout et al., 2011
	Theobroma speciosum				0.51	Argout et al., 2011
	Herrania albiflora				0.35	Argout et al., 2011
	Herrania balaensis				0.41	Argout et al., 2011
	Herrania breviligulata				0.4	Argout et al., 2011
	Herrania camargoana				0.53	Argout et al., 2011
	Herrania kanukuensis				0.38	Argout et al., 2011
	Herrania nitida				0.35	Argout et al., 2011
Vanilla planifolia	*Vanilla aphylla*				2.75	Trávníček et al., 2015
	Vanilla phaeantha		32		7.6	Jones et al., 1998
	Vanilla planifolia			2	2.52	Bory, Catrice, et al., 2008
	Vanilla planifolia	"Sterile"		3	3.84	Bory, Catrice, et al., 2008

	Vanilla planifolia	"Grosse"		4	5	Bory, Catrice, et al., 2008
	Vanilla pompona		32		7.3	Jones et al., 1998
	Vanilla tahitensis		22/31/43	2/3/4	2.62/3.91/5.14	Lepers-Andrzejewski et al., 2011

Genome sizes were accessed for four relatives of *V. planifolia: V. aphylla, V. pompona, V. tahitensis,* and *V. phaeantha.* Sizes ranged considerably from 1C 2.52 pg for *V. planifolia* to 7.6 pg for *V. phaeantha* (Table 2.4). Phenotypes of *V. planifolia* also varied in reported genome size and ploidy level. Polyploidization has also been reported within cultivated vanilla and multiple, varied genome sizes have been reported for both *V. planifolia* and *V. tahitensis.* Recorded base chromosome counts (n = 16) were the same for the diploid *V. planifolia* and *V. tahitensis* and relatives with available cytogenetic data: *V. bahiana, V. barbellata, V. dilloniana, V. griffithii, V. hartii, V. imperialis, V. mexicana, V. moonii, V. palmarum, V. phaeantha, V. pompona, V. roscheri, V. siamensis,* and *V. somae* (Table 2.5).

Table 2.5 Chromosome counts of crop wild relatives with available data from the Chromosome Counts Database: http://ccdb.tau.ac.il/home/.

Crop	Relative	Chromosome Count (n)
Theobroma cacao	*Theobroma angustifolium*	10
	Theobroma bicolor	10
	Theobroma cacao	10
	Theobroma grandiflorum	10
	Herrania albiflora	10
	Herrania purpurea	10
	Guazuma ulmifolia	8
Vanilla planifolia	*Vanilla bahiana*	16
	Vanilla barbellata	16
	Vanilla dilloniana	16
	Vanilla griffithii	16
	Vanilla hartii	16
	Vanilla imperialis	16
	Vanilla mexicana	16
	Vanilla moonii	16

	Vanilla palmarum	16
	Vanilla phaeantha	16
	Vanilla planifolia	16
	Vanilla pompona	16
	Vanilla roscheri	16
	Vanilla siamensis	16
	Vanilla somae	16
	Vanilla tahitensis	16

Species conservation assessments

The majority (80%) of *T. cacao* relatives were categorized as Least Concern when calculating the extent of occurrence (EOO), however, when calculating the area of occurrence, the majority (69%) were categorized as Endangered (EN; Figure 2.3; Table 2.6). Of the GP-2 CWRs identified in this study, the species with the highest risks of global extinction were *T. microcarpum* and *T. sylvestre,* which were categorized as Endangered from their AOO. Neither of these species was previously associated with an IUCN Red List Assessment (Table 2.7). The majority (58%) of *V. planifolia* relatives were categorized as Least Concern (LC) when calculated for their EOO, but 67% were categorized as Endangered from their AOO (Figure 2.3). The GP-2 CWR species with the highest risk was *V. helleri,* which was categorized as critically endangered (CR; Table 2.6). This species was not previously associated with an IUCN Red List Assessment (Table 2.7).

Table 2.6 Preliminary assessments of CWR conservation status following criterion B of the International Union for Conservation of Nature (IUCN).

taxa	NOP	MER	EOOkm2	AOO2km	EOOcat	AOOcat
Vanilla acuminata	3	12470.09	20.81053	12	CR	EN
Vanilla capixaba	1	0	0	4	CR	CR
Vanilla cristagalli	2	1256334	0	8	CR	CR

Vanilla dietschiana	1	0	0	4	CR	CR
Vanilla gardneri	1	0	0	4	CR	CR
Vanilla giulianettii	2	3393.748	0	8	CR	CR
Vanilla hallei	1	0	0	4	CR	CR
Vanilla havilandii	1	0	0	4	CR	CR
Vanilla helleri	2	65799.89	0	8	CR	CR
Vanilla humblotii	1	0	0	4	CR	CR
Vanilla kempteriana	2	5284.587	0	8	CR	CR
Vanilla labellopapillata	1	0	0	4	CR	CR
Vanilla martinezii	1	0	0	4	CR	CR
Vanilla moonii	3	0	0	4	CR	CR
Vanilla ochyrae	4	10.22534	0	8	CR	CR
Vanilla oroana	2	41042.83	0	8	CR	CR
Vanilla ovata	1	0	0	4	CR	CR
Vanilla parviflorum	1	0	0	4	CR	CR
Vanilla ribeiroi	1	0	0	4	CR	CR
Vanilla sarapiquensis	1	0	0	4	CR	CR
Vanilla savannarum	1	0	0	4	CR	CR
Vanilla seranica	2	0	0	8	CR	CR
Vanilla somai	1	0	0	4	CR	CR
Vanilla sumatrana	1	0	0	4	CR	CR
Vanilla tahitensis	1	0	0	4	CR	CR
Vanilla uncinata	1	0	0	4	CR	CR
Vanilla utteridgei	8	0	0	4	CR	CR
Vanilla walkeriae	1	0	0	4	CR	CR
Vanilla wariensis	1	0	0	4	CR	CR
Vanilla dubia	6	756.9859	409.711	24	EN	EN
Vanilla francoisii	4	486821.6	1104.327	16	EN	EN
Vanilla ruiziana	9	9549.445	3120.048	12	EN	EN

Vanilla somae	5	10534.93	4796.938	20	EN	EN
Vanilla africana	89	6269863	3284064	224	LC	EN
Vanilla albida	25	11962093	6207627	52	LC	EN
Vanilla angustipetala	17	3640054	1842846	56	LC	EN
Vanilla annamica	12	533650.3	206992.5	44	LC	EN
Vanilla aphylla	11	57196170	22738549	32	LC	EN
Vanilla appendiculata	17	3232221	2072098	68	LC	EN
Vanilla bahiana	113	2809014	1654119	328	LC	EN
Vanilla barbellata	15	1210295	443350.9	60	LC	EN
Vanilla bicolor	52	17315384	9329579	152	LC	EN
Vanilla borneensis	8	2524398	1116833	32	LC	EN
Vanilla bosseri	12	111615.5	42610.23	44	LC	EN
Vanilla calyculata	8	2004673	768775.2	32	LC	EN
Vanilla chamissonis	51	8867471	5359281	184	LC	EN
Vanilla columbiana	6	178218.7	79841.72	20	LC	EN
Vanilla crenulata	31	2643650	956903.2	92	LC	EN
Vanilla cribbiana	14	15113999	5689714	48	LC	EN
Vanilla cucullata	14	4079277	1122865	24	LC	EN
Vanilla decaryana	19	175868.1	92692.6	68	LC	EN
Vanilla dilloniana	8	1191572	330030.9	28	LC	EN
Vanilla dressleri	19	1225750	487707.3	52	LC	EN
Vanilla edwallii	47	4500917	2536845	156	LC	EN
Vanilla grandifolia	33	1783665	1181628	56	LC	EN
Vanilla griffithii	14	6029318	2243140	56	LC	EN
Vanilla guianensis	8	2314686	813079.5	24	LC	EN
Vanilla hartii	11	5650676	2117600	36	LC	EN
Vanilla heterolopha	8	107487.4	65802.18	24	LC	EN
Vanilla hostmannii	24	12793113	4593610	64	LC	EN
Vanilla imperialis	33	77879054	10991832	92	LC	EN

Vanilla inodora	54	17060268	8100328	144	LC	EN
Vanilla insignis	19	1194419	652037.4	56	LC	EN
Vanilla kinabaluensis	9	1359763	671380.6	20	LC	EN
Vanilla madagascariensis	47	908829.1	541165.2	152	LC	EN
Vanilla methonica	3	374080.6	175535.4	12	LC	EN
Vanilla mexicana	48	81105272	39317222	156	LC	EN
Vanilla odorata	91	38841926	14955989	308	LC	EN
Vanilla ovalis	5	146652.6	60032.68	16	LC	EN
Vanilla palmarum	277	21712571	11329186	796	LC	VU
Vanilla parvifolia	12	334230.4	241000.5	20	LC	EN
Vanilla penicillata	5	229808.8	90583.23	12	LC	EN
Vanilla perrieri	9	273308.5	140344.1	32	LC	EN
Vanilla phaeantha	16	21178506	5166077	60	LC	EN
Vanilla planifolia	304	3E+08	2.16E+08	836	LC	VU
Vanilla poitaei	21	20221490	2949774	80	LC	EN
Vanilla polylepis	15	75370812	4853519	52	LC	EN
Vanilla pompona	89	48684350	22308551	288	LC	EN
Vanilla ramificans	9	189986.7	86386.33	24	LC	EN
Vanilla ramosa	24	5163666	1447797	64	LC	EN
Vanilla roscheri	12	99471172	26088067	32	LC	EN
Vanilla schwackeana	4	31637426	6015352	16	LC	EN
Vanilla seretii	24	3163693	2162922	68	LC	EN
Vanilla siamensis	21	356304.7	129019.2	24	LC	EN
Vanilla sprucei	16	462435.9	182055.5	44	LC	EN
Vanilla trigonocarpa	40	2514046	1114957	80	LC	EN
Vanilla organensis	6	393045.3	29248.85	20	NT	EN
Vanilla chalotii	9	34346.28	15744.69	20	VU	EN
Vanilla claviculata	7	24297.22	17991.39	24	VU	EN
Vanilla costaricensis	7	51283.12	13888.53	16	VU	EN

Vanilla hamata	5	1462513	5143.77	12	VU	EN
Herrania tomentella	3	10081.51	0	8	CR	CR
Herrania dugandii	3	136813.2	0	8	CR	CR
Theobroma canumanense	2	1702353	0	8	CR	CR
Glossostemon bruguieri	1	0	0	4	CR	CR
Theobroma aspera	1	0	0	4	CR	CR
Herrania amazonica	1	0	0	4	CR	CR
Theobroma duckei	1	0	0	4	CR	CR
Theobroma hylaeum	4	14972.21	517.8727	12	EN	EN
Guazuma ulmifolia	7079	2.97E+08	1.88E+08	17816	LC	LC
Guazuma crinita	153	11597471	4739844	288	LC	EN
Theobroma cacao	1488	2.99E+08	2.15E+08	2388	LC	NT
Theobroma subincanum	1091	7492469	5686983	1920	LC	VU
Theobroma speciosum	433	36635580	16356874	960	LC	VU
Theobroma obovatum	332	8396148	5069695	656	LC	VU
Herrania mariae	123	7422402	5746862	392	LC	EN
Theobroma microcarpum	103	8937357	4457017	204	LC	EN
Theobroma grandiflorum	180	31975346	18418187	444	LC	EN
Theobroma glaucum	174	10786005	4376731	400	LC	EN
Theobroma bicolor	217	29566599	10680805	544	LC	VU
Herrania nitida	218	7757326	5055426	488	LC	EN
Herrania nycterodendron	46	2778529	1990625	144	LC	EN
Theobroma cirmolinae	24	150383	73997.48	24	LC	EN
Theobroma sylvestre	115	4388762	2914148	220	LC	EN
Herrania albiflora	43	36790645	7770431	112	LC	EN
Theobroma mammosum	67	74029.66	33514.69	108	LC	EN
Theobroma simiarum	108	2695295	1149587	232	LC	EN
Theobroma angustifolium	99	6709858	1344488	188	LC	EN
Herrania lemniscata	34	1900245	1070071	88	LC	EN

Herrania cuatrecasana	36	1504247	863622.5	92	LC	EN
Theobroma velutinum	27	2737975	512507.2	44	LC	EN
Herrania camargoana	8	1011265	239836.5	24	LC	EN
Herrania kanukuensis	22	882284.4	478769.6	48	LC	EN
Theobroma bernouillii	47	2508887	609538.3	104	LC	EN
Herrania purpurea	242	8046835	2942093	588	LC	VU
Guazuma invira	141	30546168	15728506	384	LC	EN
Theobroma gileri	41	286532.9	70482.65	72	LC	EN
Theobroma nemorale	26	154448.7	86361.82	36	LC	EN
Herrania laciniifolia	16	92884.6	34528.16	44	LC	EN
Theobroma bernoullii	47	2584927	516459.5	140	LC	EN
Herrania pulcherrima	40	897247.1	581616.3	100	LC	EN
Herrania balaensis	37	865480.1	491891.8	116	LC	EN
Theobroma chocoense	7	160412	102294	20	LC	EN
Theobroma sinuosum	7	463160.9	228119.4	16	LC	EN
Theobroma sylvestris	4	1258908	624429.9	16	LC	EN
Herrania breviligulata	11	1268272	5892.156	16	VU	EN

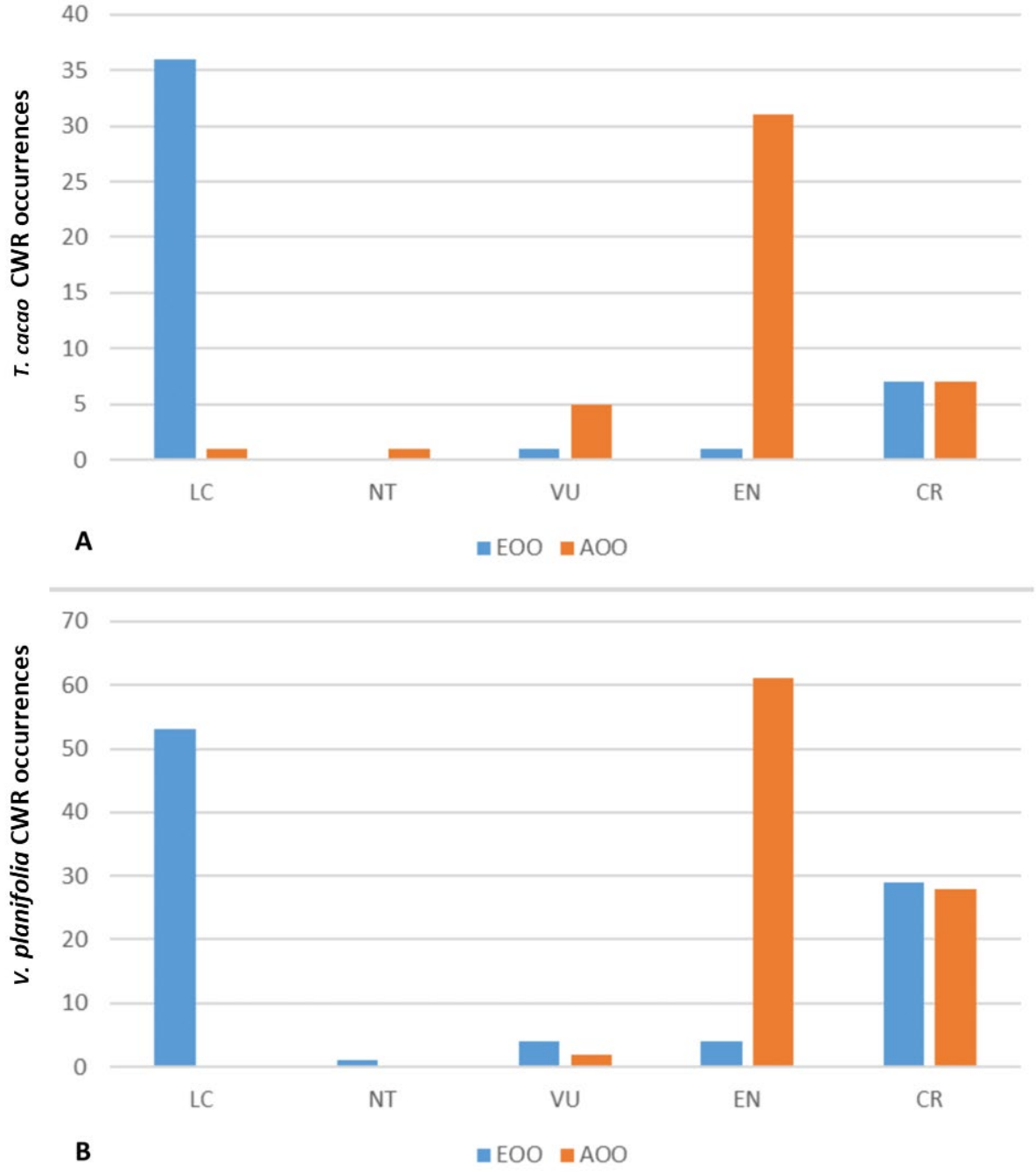


Figure 2.3 AOO and EOO categorization of A) 45 relatives of *T. cacao* from genera *Theobroma, Herrania, Guazuma,* and *Glossostemon* and the B) 91 relatives of *V. planifolia* from the genus *Vanilla*. IUCN categories are: Least Concerned (LC), Near Threatened (NT), Vulnerable (VU), Endangered (EN), and Critically Endangered (CR).

Table 2.7 Extinction risk status for crop wild relatives with available data from the Red List of Threatened Species. https://www.iucnredlist.org/.

Crop	Relative	Red List Category
Theobroma cacao	*Theobroma angustifolium*	LC
	Theobroma bernoullii	LC
	Theobroma cirmolinae	LC
	Theobroma nemorale	LC
	Theobroma simiarum	LC
	Herrania balaensis	EN
	Herrania laciniifolia	LC
	Herrania nycterodendron	LC
	Herrania pulcherrima	LC
	Herrania purpurea	LC
	Herrania umbratica	EN
	Guazuma crinita	LC
	Guazuma ulmifolia	LC
Vanilla planifolia	*Vanilla cribbiana*	CR
	Vanilla hartii	EN
	Vanilla inodora	EN
	Vanilla insignis	EN
	Vanilla odorata	EN
	Vanilla phaeantha	EN
	Vanilla phalaenopsis	LC
	Vanilla planifolia	EN
	Vanilla pompona	EN
	Vanilla somae	EN

Extracting human influence index (HII) values from CWR occurrence revealed that the majority of CWR occurrences were located in minimally impacted areas (HII between 0 and 40). For *T. cacao* CWRs, the mean HII was 31 and 83% of all occurrences were located in areas with a HII value under 50. For *V. planifolia* CWRs, the mean HII was 33 and 70% were located in areas with a HII value under 50 (Figure 2.4).

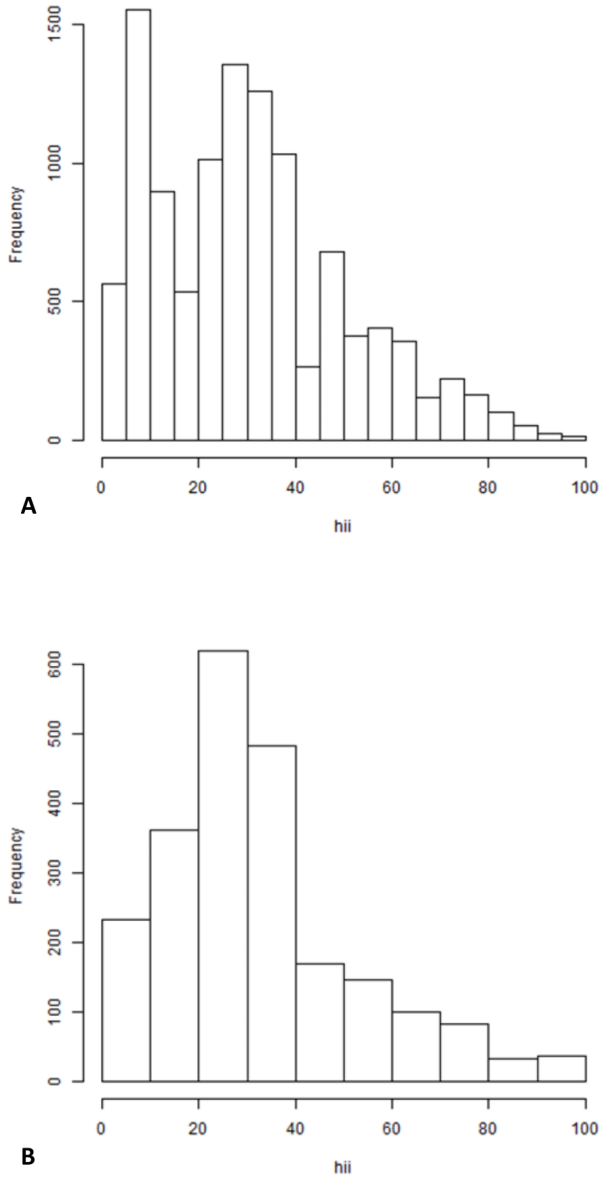


Figure 2.4 Human Influence Index of A) *T. cacao* CWR occurrences and B) *V. planifolia* CWR occurrences.

Identifying biodiversity hotspots for CWRs

The produced species richness maps of *T. cacao* and *V. planifolia* wild relatives illustrate regions of importance for the conservation of plant genetic resources for these two crops (Figure 2.5). For relatives of *T. cacao,* biodiverse regions include Colombia, Ecuador, and northern Peru. For relatives of *V. planifolia,* biodiversity is more widespread, in regions such as: Mexico, Guatemala, Belize, Nicaragua, Costa Rica, Cuba,

Puerto Rico, Colombia, Ecuador, Peru, Brazil, Cameroon, and the Democratic Republic of Congo.

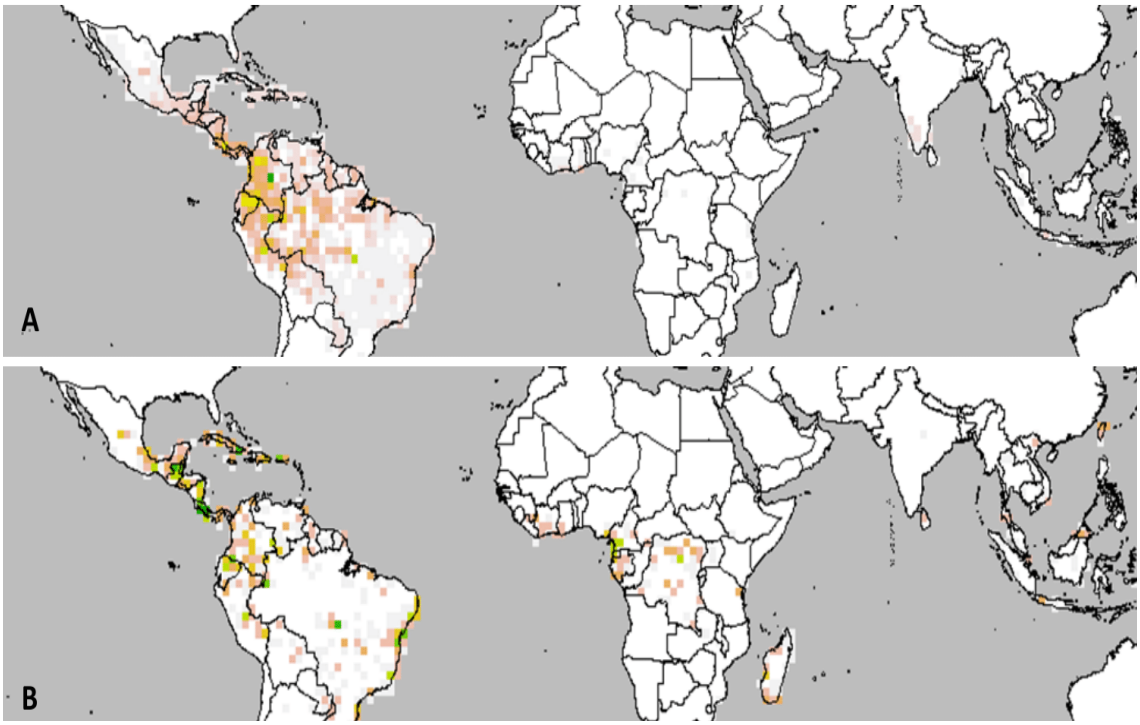


Figure 2.5 Species richness maps of CWRs of A) *T. cacao* and B) *V. planifolia*. Green are areas of high species richness and red are areas of low species richness

Discussion

Overall, there is surprisingly very little known of the relatives of these two important crops, *Theobroma cacao* and *Vanilla planifolia.* Of all 47 candidate CWRs belonging to the tribe Theobromateae, 17% were represented among ITS sequences on GenBank, 21% were represented on the Plant DNA C-values Database, 15% were represented on the Chromosome Counts Database, and the global conservation status of 28% had been assessed by the IUCN Red List of Threatened Species. Of the 110 candidate CWRs belonging to the genus *Vanilla*, 24% were represented among ITS sequences on GenBank, 15% were represented on the Plant DNA C-values Database, 5% were represented on the Chromosome Counts Database, and the global conservation

status of 9% had been assessed by the IUCN Red List of Threatened Species. This study exemplifies the urgent need for additional collecting, sequencing, and research on the species closely related to crops.

CWRs of *Theobroma cacao*

This review identifies seven species that should be categorized as GP-2 CWRs of *T. cacao* based on previous hybridization data and similarities of ITS sequences and cytogenetics: *T. bicolor, T. microcarpum, T. grandiflorum, T. obovatum, T. speciosum T. subincanum,* and *T. sylvestre* (Table 3). While sequences of *G. ulmifolia* did have genetic distances within the threshold that was used to determine cross-compatibility, the species' recorded base chromosome numbers (n=8) were dissimilar to the crop species (n=10; Table 5), therefore it is unlikely to be able to hybridize with *T. cacao* to produce fertile offspring and was categorized as GP-3. The seven GP-2 CWRs should be prioritized for conservation and hybridization experiments. Only 20% of the candidate CWRs were able to be assessed for cross-compatibility using available ITS sequence data. Additional collecting and sequencing of taxa within the Tribe Theobromateae are needed to increase our knowledge of cross-compatibility with *T. cacao*. Results showing that the majority of CWRs occur in landscapes minimally affected by human activities illustrate the necessity for protected areas and intact habitat within CWR distributions (Figure 2.4). Species such as *T. microcarpum* and *T. sylvestre* exhibit high risks of extinction as revealed by their Endangered status, therefore, species-specific conservation strategies should prioritize higher risk CWRs such as these. Additionally, regional strategies should be implemented in the regions identified within this study to have a high biodiversity of CWR occurrence: Mexico, Peru, and Colombia (Figure 2.5).

CWRs of *Vanilla planifolia*

Within this review, 17 GP-2 CWRS were identified for *V. planifolia* and ten congeneric species were categorized as GP-3 (Table 3). Because *V. planifolia* has been shown to hybridize with a distantly related (and geographically distant) *V. aphylla*, it is probable that species that are less genetically distant and that exhibit the same chromosome numbers, ploidy level, and genome size, are also able to hybridize with *V. planifolia* to produce fertile offspring. These 17 species should be prioritized for additional research on cross-compatibility. Only 25% of all candidate CWRs had ITS sequence data available on GenBank, and fewer than this had genome size (Table 4) and cytogenetic data (Table 5). Due to the high amount of labor, time and money necessary to conduct hybridization experiments, it is highly valuable to use genetic distances and cytogenetic compatibility as a proxy for cross-compatibility. Within this large genus, additional collecting and sequencing of candidate CWRs could benefit our overall understanding of relationships and reproductive similarities to *V. planifolia*. As with *T. cacao* CWRs, most CWRs of *V. planifolia* occur in landscape minimally impacted by human activities (Figure 2.4). Therefore, protection of suitable intact habitat is essential to preserves CWRs and protect these resources of genetic variability. Specifically, species such as *V. helleri*, which was determined to be critically endangered through this study (Table 7), should be prioritized within conservation efforts. Increased protection of natural areas is essential to conserving vanilla's plant genetic resources especially within the CWR hotspots in Mexico, Central America, South America, and western Africa (Figure 2.5).

Species delimitation to increase understanding of crop relatives

Pairwise genetic distances reveal intraspecific variation among both *T. cacao* and *V. planifolia* ITS sequences. Accessions of *T. cacao* vary between 0% and 51% and accessions of *V. planioflia* vary between 0% and 7% (Figures 2.1, 2.2). Although there is no standardized threshold for delimiting species based on ITS genetic distances, one study reports that the mean average intraspecific genetic distance among angiosperms is 1.20%, and the mean average genetic distance between sister species is 3.98% (Qin et al., 2017). Results from this study show intraspecific genetic distances of *T. cacao* vary much greater than the reported average, up to 51%. While these findings may indicate an insufficiency in the lengths of ITS sequences that were analyzed, it is possible that they reveal an inaccuracy in the delimitation of the species. Further research needs to be conducted on the genetic variation of *T. cacao* to assess delimitation and possibly identify cryptic species.

Intraspecific variation of *V. planifolia* accessions, while still high (up to 7%), was more congruent with the averages reported by Qin et al (2017). However, delimitation of this species should also be further investigated, especially because congeneric cryptic species have been reported (Soto Arenas & Cribb, 2010). For example, genetic distances between *V. planifoli*a and *V. pompona*, which exhibit very similar vegetative morphology, vary between two distinct peaks (Figure 2.2). These findings may indicate unresolved species delimitations, the misidentification of cryptic species, or hybridization between both species. This is also the case for interspecific genetic distances between *V. planifolia* and *V. tahitensis,* however, these results are expected as *V. tahitensis* has been shown to be the result of a hybrid origin (Lubinsky, Cameron, et al., 2008). Genomic

research assessing intra- and inter- specific variation is essential to further assess relationships and help to delimit species within this genus.

Risk of genetic pollution from crops to CWRs

On top of landscape use and climatic changes, the genetic erosion of many CWRs is being compounded through the genetic pollution of nearby crops. The term "genetic pollution" refers to the gene flow and introgression from a domesticated species into its wild, native species or relatives. Crops cultivated in their center of origin can have severe impacts on surrounding CWR populations which are often already stressed by limited population sizes, habitat destruction, and climate change. Genetic pollution from crops has severely disrupted the genetic structure of sympatric wild ryegrass in the United Kingdom (Sackville Hamilton, 1999) and has affected many wild rice species and even caused the extinction of *Oryza perennis,* a wild rice relative in Taiwan (Kiang et al., 1979). Many concerns arise from the large-scale cultivation of genetically modified (GM) crops around smaller CWR populations. Although it is illegal to plant GM maize in Mexico, traditional varieties grown in remote regions have been found to be contaminated by up to 35% with genetic material from GM maize (Maxted & Guarino, 2004). The consequences of gene introgression from large, cultivated populations to small wild populations are still not fully realized and may have enduring effects on PGR, especially in regions of high CWR biodiversity, such as the ones identified within this study. Efforts to conserve CWR should take into account proximity to crops, especially GM crops, to protect against undesired introgression and genetic pollution.

Mediating species concepts to identify CWRs

The concept of a species is an important aspect of biology and species delimitation is crucial to many biological disciplines. Uniformly defining what constitutes a species, however, has been a supreme challenge. The incongruence of species concepts, and often a lack of available data, greatly complicates the identification of crop wild relatives and inhibits their conservation. The biological species concept is predominantly used to define CWRs as taxa that can successfully breed and produce fertile offspring. Species may exhibit reproductive barriers in the form of habitat isolation, pollinator isolation, gametic incompatibilities, hybrid inviability, etc While this concept may be the most practical when the objective is to assess hybridization for agricultural purposes, assessing reproductive barriers through experimentation can be time-consuming and costly. Due to ongoing threats to many wild plant populations, amending this concept is necessary. Yet the taxonomic species concept, defined by morphological characteristics which might not denote true evolutionary relatedness, may be ineffective to accurately identify CWRs. Appending the biological species concept using genetic data as a proxy for cross-compatibility can aid in a more comprehensive and rapid identification of CWRs. Therefore, where hybridization data is unavailable, the genetic species concept, which uses genetic differences to infer reproductive isolation, may be a more appropriate route to rapidly identifying CWRs and prioritizing their conservation.

CHAPTER THREE: UNCOVERING HAPLOTYPE DIVERSITY IN CULTIVATED MEXICAN VANILLA SPECIES

The final version of this article has undergone full peer review and has been published. Please see:
Ellestad, P., Pérez-Farrera, M. A., Forest, F., & Buerki, S. (2022). Uncovering haplotype diversity in cultivated Mexican vanilla species. *American Journal of Botany*. https://doi.org/10.1002/ajb2.16024

Abstract

Premise: Although vanilla is one of the best-known spices, there is only a limited understanding of its biology and genetics within Mexico, where its cultivation originated and where phenotypic variability is high. This study aims to augment our understanding of vanilla's genetic resources by assessing species delimitation and genetic, geographic, and climatic variability within Mexican cultivated vanilla.

Methods: Using nuclear and plastid DNA sequence data from 58 Mexican samples collected from three regions and 133 ex situ accessions, we assessed species monophyly using phylogenetic analyses and genetic distances. Intraspecific genetic variation was summarized through the identification of haplotypes. Within the primarily cultivated species, *Vanilla planifolia*, haplotype relationships were further verified using plastome and rRNA gene sequences. Climatic niche and haplotype composition were assessed across the landscape.

Results: Three species (*V. planifolia*, *V. pompona*, and *V. insignis*) and 13 haplotypes were identified among Mexican vanilla. Within *V. planifolia* haplotypes, hard

phylogenetic incongruences between plastid and nuclear sequences suggest past hybridization events. Eight haplotypes consisted exclusively of Mexican samples. The dominant *V. planifolia* haplotype occurred throughout all three regions as well as outside of its country of origin. Haplotype richness was found to be highest in regions around Papantla and Chinantla.

Conclusions: Long histories of regional cultivation support the consideration of endemic haplotypes as landraces shaped by adaptation to local conditions and/or hybridization. Results may aid further genomic investigations of vanilla's genetic resources and ultimately support the preservation of genetic diversity within the economically important crop.

Introduction

Conserving the genetic diversity of crop species and their wild relatives has become a mounting concern as the detrimental effects of climate change and the genetic erosion of more crops in their center of origin are documented, notably within staple crops (e.g. rice, maize, soybean; Gao, 2003; Mammadov et al., 2018). These genetic resources offer an important source of natural variation, which may be used to increase crop resilience and help ensure sustainability in a changing climate. In this study, we aim to augment the current understanding of the genetic resources and species boundaries of an economically important and beloved spice, vanilla (*Vanilla* Plum. ex Mill., Orchidaceae), by shedding light into the genetic, phenotypic, and climatic variability of cultivated plants within their purported country of origin, Mexico (Lubinsky, Bory, et al.,

2008). Although vanilla is one of the most well-known and valuable spices worldwide, a clear understanding of the genetic variability within its native range is lacking.

Vanilla extract is produced from the cured seed pods of the tropical orchid genus *Vanilla* and is the second most valuable spice in the world, after saffron (Baker, 2018). In 2020, its global market was valued at 866.6 million USD and is expected to grow by 13.4% by 2026 (MarketWatch, 2020). Originating from Mexico, it has been introduced across the globe to be cultivated for use in the culinary, cosmetic, and medicinal industries (Bruman, 1948; Lubinsky, Bory, et al., 2008). Currently, Madagascar is the largest producer of vanilla, followed by Indonesia and Mexico (Food and Agriculture Organization of the United Nations, 2020). While the species *Vanilla planifolia* Andrews and *Vanilla* x *tahitensis* J.W. Moore are the primary global sources of vanilla, other species such as *Vanilla pompona* Schiede*, Vanilla odorata* C. Presl, and *Vanilla insignis* Ames can also be found in cultivation in certain regions that correspond to their respective ranges (Schlüter et al., 2007; Soto Arenas & Dressler, 2010). Their use in commercial vanilla production, however, is prohibited within United States and European markets (CBI, 2018; U.S. FDA, 2022). Cultivated in French Polynesia and New Guinea, *V.* x *tahitensis* is presumed to have human-mediated origins through hybridization between *V. odorata* and *V. planifolia* (Hasing et al., 2020; Lubinsky, Cameron, et al., 2008) or *V. sotoarenasii* M.Pignal, Azof.-Bolaños & Grisoni (Favre et al., 2022), although recent evidence supports the possibility of a rare natural hybridization event (Chambers et al., 2021). This latter hypothesis could be confirmed since the distributions of multiple congeneric species, including *V. pompona, V. insignis* and *V. odorata,* overlap

with *V. planifolia* in parts of Mexico, and Central and South America (Ellestad et al., 2021).

Vanilla is only able to reproduce naturally within its native range, where it co-occurs with specialized pollinators and dispersers (Cameron, 2011; Lozano Rodríguez et al., 2022). Outside of its native range, a manual and labor-intensive technique is required to pollinate the flower and produce fruits (Cameron, 2011). Due to the absence of natural genetic recombination through manual pollination and clonal vegetative propagation from a single origin (Lubinsky, Bory, et al., 2008), the genetic diversity of globally cultivated vanilla has become severely constrained (Y. Hu et al., 2019; Schlüter et al., 2007; Soto Arenas & Dressler, 2010). Therefore, its capacity to cope with changing environmental conditions is consequently limited (Armenta-Montero et al., 2022). Large-scale loss of vanilla plants, due to fungal outbreaks associated with a changing climate, have caused significant losses over the past decade (Pinaria et al., 2010). On top of increasing drought conditions (Varela et al., 2021), the rapid loss of wild populations due to land-use change, habitat fragmentation, and illegal harvesting poses an immediate and irreversible threat to the preservation of the genetic variation within this crop (Hinsley et al., 2018; Soto Arenas & Dressler, 2010).

Due to the importance of cultivated vanilla on the international market, multiple molecular studies have been conducted to characterize genetic diversity levels within *V. planifolia* cultivated outside of the native distribution of the genus. Congruent results have identified a substantial lack of genetic diversity within globally cultivated vanilla (Besse et al., 2004; Y. Hu et al., 2019; Lubinsky, Cameron, et al., 2008) and intraspecific morphological variation as well as genetic structuring within native wild populations

(Ramos-Castellá et al., 2017; Schlüter et al., 2007; Villanueva-Viramontes et al., 2017). More recent studies using genomic data have identified a trend in genomic heterozygosity levels of *V. planifolia* with cultivated accessions exhibiting higher levels than wild accessions (Chambers et al., 2021; Favre et al., 2022). Additionally, congeneric relationships have been assessed (although many remain unresolved) and hybridization has been shown to occur between *V. planifolia* and six species: *V. pompona* (Delassus, 1963)*, V. aphylla* Blume (Divakaran et al., 2006)*, V. odorata, V.* x *tahitensis* (Lubinsky, Cameron, et al., 2008), *V. phaeantha* Rchb.f. (Y. Hu et al., 2019)*,* and *V. palmarum* Salzm. ex Lindl (J. Li et al., 2020). Although previous research has provided a broad representation of vanilla's genetic resources, additional work is needed to adequately characterize the genetic resources of a highly important subset of vanilla's gene pool found within locally cultivated populations in Mexico, the center of origin of globally cultivated *V. planifolia*.

In Mexico alone, at least ten other species of *Vanilla* occur, many of which exhibit cryptic vegetative morphology (Soto Arenas & Cribb, 2010), throughout a range of environmental conditions. Climatic requirements vary for species such as *V. planifolia*, *V. pompona, V. insignis, V. odorata,* and *V. inodora*; of these *V. pompona* has been found to tolerate comparatively low annual precipitation, and *V. planifolia* has been shown to exhibit the widest climatic niche breadth in terms of precipitation and temperature (Ellestad et al., 2021; Flores Jiménez et al., 2017). Within *V. planifolia,* morphological variations of leaf size, shape, and color complicate species delimitation based on the morphological species concept (Cronquist, 1978). Three phenotypes have been well recorded among Mexican vanilla: 'Mansa', 'Variegata', and 'Oreja de Burro' (Soto

Arenas & Dressler, 2010). Species misidentification has been a prevalent obstacle to the reliability of *Vanilla* assessments (Karremans et al., 2020). Although recent genomic advances have increased the resolution of evolutionary relationships within the genus, species delimitations are still largely unresolved (Bouetard et al., 2010). Furthermore, the type specimen of *V. planifolia* is an illustration depicting a cultivated individual from the West Indies, therefore, a genetic reference for this species is not available (Andrews et al., 1797). These challenges limit our understanding of species delimitations and phylogenetic relationships within *Vanilla*, which in turn hamper effective conservation.

Our main objectives were to investigate i) how many species of vanilla are cultivated in Mexico, ii) what levels of genetic variation (haplotypes) exist within each taxon, and iii) determine if species and/or haplotypes are associated to specific climates. Based on the phenotypic variability of *V. planifolia* recorded throughout the region, we hypothesized that standing phenotypic variation exists as a result of phenotypic plasticity in response to contrasting climates and therefore represents a single species with limited genetic variation. It is predicted, however, that local adaptation (here inferred using haplotypes) within the species is present and attributable to long histories of local domestication and/or procurement from wild populations shaped by natural selection. Cultivated plants derived from distinct domestication events may be better adapted to local environmental conditions and could prove to be important for future crop improvement. Through an improved understanding of the genetic variation within regionally cultivated vanilla, we stress the importance of native haplotypes as genetic resources and advocate for their inclusion into conservation strategies.

Here, we aimed to augment previous research by assessing species delimitation and genetic variability of Mexican vanilla along an environmental gradient across the states of Veracruz, Puebla, and Oaxaca using DNA sequence data (plastid gene *rbcL* and nuclear ribosomal ITS region). A dual approach was implemented to assess monophyly by considering both evolutionary relatedness, based on the phylogenetic species concept (Wheeler, 1999), and genetic similarities, based on Kimura's 2-parameter (K2P; Kimura, 1980) genetic distances, which has been widely used to delimitate species using DNA barcoding data (Chen et al., 2010; Yao et al., 2010). To summarize findings and plot intra-specific genetic variation across the landscape, haplotypes were identified within each species unit previously inferred by the phylogenetic species concept. DNA barcoding, which allows for a glimpse into small portions of genomes to resourcefully solve evolutionary relationships, is limited in its depth of resolution. Therefore, species delimitation and barcoding results were further investigated by analyzing full plastome and ribosomal RNA (rRNA) sequences within a subset of samples focusing on *V. planifolia* haplotypes.

Materials and Methods

Sampling

Samples were collected from 58 *Vanilla* plants within the origin and major vanilla cultivation regions in Mexico in October of 2019, mostly from plantations (Table 3.1). Samples were collected from three general latitudinal regions: the northernmost Region 1 around the city of Papantla, Veracruz (including coastal regions to mountainous regions around Cuetzalan, Puebla); the central Region 2, around Cordoba and Xalapa, Veracruz; and the southernmost Region 3, around Chinantla, Oaxaca (Figure 3.1). Due to

abundance, the majority of samples (43) were collected from Region 1; three were collected from Region 2; and 12 were collected from Region 3. When possible, cultivators were asked a series of questions to ascertain the source of collected plants, i.e. from wild or cultivated populations, and from which region (1, 2, or 3). Flowers were absent during the time of collection; therefore, taxonomic identity was inferred using vegetative morphological characters as described in Soto Arenas and Cribb (2010), and based on the descriptions obtained from cultivators. Sample collections aimed to incorporate well-known phenotypes recorded within Mexican vanilla, such as: 'Mansa', the most common; 'Variegata', which exhibits yellow-green stripes on leaves and stems; and 'Oreja del Burro', which is vegetatively similar to 'Mansa', but possesses sulcate fruit (Soto Arenas & Dressler, 2010). Vegetative cuttings were taken from each individual and one gram of leaf material was dried in silica gel for genetic analyses. Voucher specimens for these individuals were deposited at Herbario Eizi Matuda (HEM) in Mexico and are also represented in the living collections at Berriozábal, Chiapas, Mexico. Exact geo-locations are not disclosed here to protect this endangered species and the crop security of local farmers.

Table 3.1 Field-collected *Vanilla* samples with collection regions, population and source type, and region where samples were purportedly sourced. See Appendix D for full list of samples.

Sample ID	Species	Haplotype	Collection Region	Population Type	Source Type	Source Region
MEX20	V. planifolia	HPl1	Papantla	cultivated	n/a	n/a
MEX5	V. planifolia	HPl3	Papantla	cultivated	n/a	n/a
MEX6	V. planifolia	HPl3	Papantla	cultivated	cultivated	Papantla
MEX7	V. planifolia	HPl3	Papantla	cultivated	cultivated	Papantla
MEX9	V. planifolia	HPl3	Papantla	cultivated	cultivated	Papantla
MEX10	V. planifolia	HPl3	Papantla	cultivated	cultivated	Papantla

MEX21	V. planifolia	HPl3	Papantla	cultivated	n/a	n/a
MEX22	V. planifolia	HPl3	Papantla	cultivated	cultivated	Papantla
MEX23	V. planifolia	HPl3	Papantla	cultivated	cultivated	Papantla
MEX24	V. planifolia	HPl3	Papantla	cultivated	cultivated	Papantla
MEX25	V. planifolia	HPl3	Papantla	wild	n/a	n/a
MEX27	V. planifolia	HPl3	Papantla	cultivated	cultivated	Papantla
MEX28	V. planifolia	HPl3	Papantla	wild	n/a	n/a
MEX29	V. planifolia	HPl3	Papantla	wild	n/a	n/a
MEX30	V. planifolia	HPl3	Papantla	wild	n/a	n/a
MEX31	V. planifolia	HPl3	Papantla	wild	n/a	n/a
MEX38	V. planifolia	HPl3	Papantla	cultivated	n/a	n/a
MEX39	V. planifolia	HPl3	Papantla	cultivated	wild	Papantla
MEX40	V. planifolia	HPl3	Papantla	cultivated	n/a	n/a
MEX41	V. planifolia	HPl3	Papantla	cultivated	n/a	n/a
MEX46	V. planifolia	HPl3	Papantla	cultivated	n/a	n/a
MEX47	V. planifolia	HPl3	Papantla	cultivated	n/a	n/a
MEX48	V. planifolia	HPl3	Papantla	cultivated	cultivated	Papantla
MEX52	V. planifolia	HPl3	Papantla	cultivated	n/a	n/a
MEX54	V. planifolia	HPl3	Papantla	cultivated	wild	Cuetzalan
MEX55	V. planifolia	HPl3	Papantla	cultivated	wild	Cuetzalan
MEX56	V. planifolia	HPl3	Papantla	cultivated	wild	Chinantla
MEX59	V. planifolia	HPl3	Xalapa/Cordoba	cultivated	cultivated	Oaxaca
MEX65	V. planifolia	HPl3	Xalapa/Cordoba	cultivated	cultivated	Oaxaca
MEX66	V. planifolia	HPl3	Chinantla	cultivated	wild	Chinantla
MEX69	V. planifolia	HPl3	Chinantla	cultivated	cultivated	Tuxtepec
MEX70	V. planifolia	HPl3	Chinantla	cultivated	cultivated	Tuxtepec
MEX72	V. planifolia	HPl3	Chinantla	cultivated	wild	Chinantla
MEX76	V. planifolia	HPl3	Chinantla	cultivated	wild	Chinantla
MEX75	V. planifolia	HPl3	Chinantla	cultivated	wild	Chinantla
MEX77	V. planifolia	HPl3	Chinantla	cultivated	wild	Chinantla
MEX79	V. planifolia	HPl3	Chinantla	cultivated	wild	Chinantla
MEX13	V. planifolia	HPl4	Papantla	cultivated	n/a	n/a
MEX12	V. planifolia	HPl5	Papantla	cultivated	n/a	Papantla
MEX26	V. planifolia	HPl7	Papantla	cultivated	n/a	n/a
MEX14	V. planifolia	HPl8	Papantla	cultivated	n/a	n/a
MEX18	V. planifolia	HPl14	Papantla	cultivated	n/a	n/a
MEX19	V. planifolia	HPl14	Papantla	cultivated	cultivated	Papantla
MEX67	V. planifolia	HPl16	Chinantla	cultivated	wild	Chinantla
MEX80	V. planifolia	HPl16	Chinantla	cultivated	wild	Chinantla

MEX36	V. planifolia	HPl21	Papantla	cultivated	n/a	n/a
MEX51	V. planifolia	HPl22	Papantla	cultivated	cultivated	Papantla
MEX50	V. insignis	HIn5	Papantla	cultivated	cultivated	Papantla
MEX53	V. insignis	HIn5	Papantla	cultivated	wild	Chinantla
MEX61	V. insignis	HIn5	Xalapa/Cordoba	wild	n/a	n/a
MEX73	V. insignis	HIn5	Chinantla	cultivated	wild	Chinantla
MEX8	V. pompona	HPo3	Papantla	cultivated	cultivated	Papantla
MEX11	V. pompona	HPo3	Papantla	cultivated	cultivated	Papantla
MEX42	V. pompona	HPo3	Papantla	wild	n/a	n/a
MEX44	V. pompona	HPo3	Papantla	wild	n/a	n/a
MEX45	V. pompona	HPo3	Papantla	wild	n/a	n/a
MEX74	V. pompona	HPo3	Chinantla	cultivated	wild	Chinantla
MEX49	V. pompona	HPo7	Papantla	cultivated	cultivated	Papantla

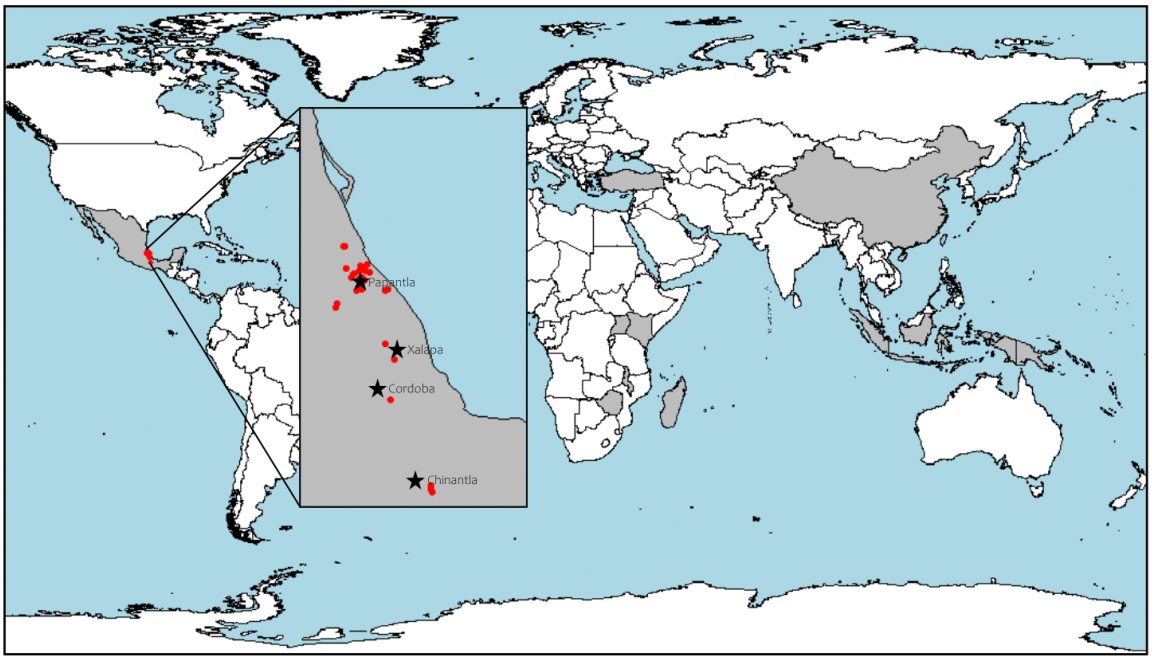


Figure 3.1 World distribution of vanilla production. Countries producing vanilla are shaded in grey. The insert shows the sampling region in Mexico, with red points indicating sampling locations.

Ex-situ samples were included in the analysis to further support taxonomic identification of field-collected samples and to provide a hypothesis on haplotypes widely cultivated outside of Mexico. *Ex-situ* samples consisted of living collections maintained

at Boise State University and Missouri Botanical Garden, GenBank accessions of *Vanilla* spp., and GenBank accessions of *Erythrorchis cassythoides,* which was used as the most external outgroup taxon in subsequent analyses (Appendix D). An additional phenotype, 'Albomarginata', which exhibits leaves with white margins and is not found natively in Mexico (Soto Arenas & Dressler, 2010), was also included within the *ex-situ* accessions. Voucher specimens for specimens obtained from living collections were deposited at SRP and MO. To broaden sampling and species assignment, *ex-situ* samples included representatives from all three sections (*Vanilla, Tethya, Xanata*) within subgenus *Xanata* (Soto Arenas & Cribb, 2010).

DNA Extraction, PCR, and sanger sequencing

DNA barcoding of the plastid gene *rbcL* and nuclear ribosomal ITS region were used to i) infer species delimitation, ii) assess genetic variation, and iii) identify haplotypes within vanilla species. The conserved *rbcL* gene is commonly sequenced in plants and used as a DNA barcode (Hollingsworth et al., 2009). This plastid gene codes for the large subunit ribulose-1,5-bisphosphate carboxylase/oxygenase (RUBISCO), an enzyme essential for carbon fixation, thus genetic variation within this gene could provide insights into local adaptation. The ITS region is also commonly sequenced in plant phylogenetic studies and has been shown to be effective for species discrimination (Feliner & Rosselló, 2007), even within genus *Vanilla* (Besse et al., 2021).

Genomic DNA was extracted from leaf samples dried in silica gel using the Qiagen DNeasy Plant Mini Kit following the manufacturer's protocol. DNA yield was quantified using a Qubit fluorometer (Thermo Fisher Scientific, Inchinnan, UK). PCR amplifications of *rbcL* and ITS regions were carried out using the primers rbclF, rbcLR

(Fazekas et al., 2008), ITSp4 and ITSp5 (Cheng et al., 2015). PCR amplifications were performed in a reaction volume of 25 µL containing 1 µL of genomic DNA template, 12.5 µL of AmpliTaq Gold 360 Master Mix (Thermo Fisher Scientific), 2 µL of each 10 µM primer, and 8.5 µL of sterile distilled water. Thermal cycling programs for the amplification of *rbcL* and ITS were performed on a Biometra® T3 thermocycler following Buerki et al. (2009): initial denaturation at 95 °C for 2 minutes; 35 cycles of denaturation at 95 °C for 45 seconds, annealing at 50 °C for 45 seconds, and extension at 72 °C for 1 minute; final extension at 72 °C for 10 min. Amplification was detected by 2% agarose gel electrophoresis in 1X TAE buffer and visualized under UV light. Samples with verifiable PCR amplification were sent to Genewiz (South Plainfield, New Jersey, USA) for PCR purification and Sanger sequencing using the rbcLF and ITSp4F primers.

Sequence chromatograms from field collected samples were visualized and edited with FinchTV version 1.4.0 (Geospiza Inc.) and online BLAST analyses (Altschul et al., 1990) were performed to validate sequencing of targeted DNA regions and taxonomy. A search for GenBank accessions was conducted using the R package *rentrez* package (D. J. Winter, 2017) for all available *Vanilla* species as well as *E. cassythoides*, the sister genus to *Vanilla*, which was used as outgroup taxon for all phylogenetic analyses (Pansarin, 2016). Sequences resulting from the GenBank query for *rbcL* and ITS were combined to respective sample sequence datasets and aligned using MAFFT V7 (Katoh & Standley, 2013). Alignments were manually edited in MEGA version X (Kumar et al., 2018) to exclude GenBank sequences with minimal nucleotide position overlapping with DNA sequences from samples collected within this study.

Species delimitation and haplotype identification

Phylogenetic analyses of the *rbcL* gene and ITS region of all samples were performed on the CIPRES web server (Miller et al., 2010). Best-fitting substitution models were chosen using the Bayesian Information Criterion output from MEGA. Maximum likelihood analyses and rapid bootstrapping were conducted with RAxML-HPC2 on both regions separately (Stamatakis, 2014). Nodes with bootstrap support values (BS) below 50% were not considered supported, 50%-75% were considered weakly supported, and 75%-100% were considered strongly supported. Bayesian analyses were conducted with MrBayes V3.2.7a (Huelsenbeck & Ronquist, 2001), performing two runs of three chains for 50 million generations and sampling one tree every 1,000 generations. Convergence among chains was verified using Tracer v1.7.1 and 25% of trees were discarded as burn-in (Rambaut et al., 2018). Nodes with Bayesian posterior probabilities (BPP) below 0.70 were considered not supported, 0.70-0.95 were moderately supported, and 0.95-1 were strongly supported. Complementing preliminary morphological identification, species were inferred based on phylogenetic relatedness using *ex-situ* accessions whose identities had been confirmed by previous molecular characterization or identification by a botanical specialist.

Genetic variation between Mexican and *ex-situ* samples was assessed to propose species boundaries and assess inter- and intra-specific variation using the established average threshold of inter-specific genetic variation described by Qin et al. (2017), which showed the average genetic distance of sister species in angiosperms to be 3.98% for ITS2. Pairwise genetic distances among *Vanilla* DNA sequences at each DNA barcoding locus were calculated using the K2P model as implemented in the default *dist.dna*

function within the R package ‘ape’ (Paradis & Schliep, 2019; R Core Team, 2017). Ridgeline plots of *rbcL* and ITS K2P distances were created using ‘ggridges’ (Wilke, 2021) to illustrate inter- and intra-specific genetic variability and estimate/visualize species boundaries among all accessions. Due to the greater variability of the ITS region compared to *rbcL,* species boundaries were only based on ITS data (Cheng et al., 2015). To summarize intra-specific genetic variation, ITS nucleotide variation of samples was assessed in terms of haplotypes, identified using the R package ‘haplotypes’ applying the simple indel coding method (Aktas, 2015). Among *rbcL* sequences, nucleotide variation was also assessed, and the presence of non-synonymous amino acid changes were identified by translating codons to amino acids in AliView (Larsson, 2014).

Genome skimming and phylogenetic reconstruction of haplotypes

Reliability of inferences based on chosen barcodes was further verified through genomic sequencing of a subset of *V. planifolia* samples incorporating one sample representing each haplotype previously identified (Table 3.2). A genome skimming approach was used to reconstruct plastomes and rRNA sequences for phylogenetic analyses. Genomic DNA samples were sent to Genewiz for library preparation and sequencing of 150bp paired-end reads using an Illumina HiSeq platform. Raw sequences were checked for quality using FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed using default parameters in Trimmomatic (Bolger et al., 2014). Plastomes and rRNA were reconstructed *de novo* using the GetOrganelle toolkit to assemble reads following the recommended parameters for Embryophyte plant plastomes and nuclear rRNA (Jin et al., 2020). For sample sequences resulting in incomplete graphs, the number of rounds (-R)

was increased to 75 and the word count (-w) was reduced to 0.6. For the outgroup taxon, plastome and rRNA sequences were downloaded from GenBank for *Vanilla madagascariensis*, which belongs to the sister section, *Tethya*, from the subgenus *Xanata* in which *V. planifolia* is assigned (Soto Arenas & Cribb, 2010). Resulting DNA sequences were aligned, and maximum likelihood and Bayesian phylogenetic analyses were performed using the same approaches as described above for *rbcL* and ITS.

Table 3.2 All accessions from the three clades inferred from ITS phylogenetic analysis (see Fig. 2) containing Mexican samples, along with their respective haplotype assignment. See Appendix D for full list of accessions used in phylogenetic analyses.

ID	Species Assignment	Haplotype	Reference
BSU81	*V. planifolia*	HPl3	SRP/BSU living collection
BSU82	*V. planifolia*	HPl3	SRP/BSU living collection
BSU83	*V. planifolia*	HPl3	SRP/BSU living collection
BSU84	*V. planifolia*	HPl3	SRP/BSU living collection
BSU85	*V. planifolia*	HPl25	SRP/BSU living collection
GB1517415473	*V. planifolia*	HPl13	GenBank (Hu et al., 2019)
GB1517415474	*V. planifolia*	HPl23	GenBank (Hu et al., 2019)
GB1708599397	*V. planifolia*	HPl3	GenBank (Kim et al., 2020)
GB1789804134	*V. planifolia*	HPl6	GenBank (Salazar,G.A.)
GB1789804135	*V. planifolia*	HPl3	GenBank (Salazar,G.A.)
GB1789804136	*V. planifolia*	HPl10	GenBank (Salazar,G.A.)
GB1789804137	*V. planifolia*	HPl15	GenBank (Salazar,G.A.)
GB1789804138	*V. planifolia*	HPl17	GenBank (Salazar,G.A.)
GB1789804139	*V. planifolia*	HPl9	GenBank (Salazar,G.A.)
GB1789804140	*V. planifolia*	HPl18	GenBank (Salazar,G.A.)
GB1789804142	*V. planifolia*	HPl2	GenBank (Salazar,G.A.)
GB24397240	*V. planifolia*	HPl11	GenBank (Salazar,G.A.)
GB260401091	*V. planifolia*	HPl3	GenBank (Belanger & Havkin-Frenkel, 2010)
GB260401092	*V. planifolia*	HPl19	GenBank (Belanger & Havkin-Frenkel, 2010)
GB260401093	*V. planifolia*	HPl16	GenBank (Belanger & Havkin-Frenkel, 2010)
GB260401094	*V. planifolia*	HPl24	GenBank (Belanger & Havkin-Frenkel, 2010)
GB260401095	*V. planifolia*	HPl19	GenBank (Belanger & Havkin-

			Frenkel, 2010)
GB260401096	*V. planifolia*	HPl12	GenBank (Belanger & Havkin-Frenkel, 2010)
GB260401097	*V. planifolia*	HPl20	GenBank (Belanger & Havkin-Frenkel, 2010)
GB260401098	*V. planifolia*	HPl3	GenBank (Belanger & Havkin-Frenkel, 2010)
MEX10	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX12	*V. planifolia*	HPl5	HEM/UNICACH Living Collection
MEX13	*V. planifolia*	HPl4	HEM/UNICACH Living Collection
MEX14	*V. planifolia*	HPl8	HEM/UNICACH Living Collection
MEX18	*V. planifolia*	HPl14	HEM/UNICACH Living Collection
MEX19	*V. planifolia*	HPl14	HEM/UNICACH Living Collection
MEX20	*V. planifolia*	HPl1	HEM/UNICACH Living Collection
MEX21	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX22	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX23	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX24	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX25	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX26	*V. planifolia*	HPl7	HEM/UNICACH Living Collection
MEX27	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX28	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX29	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX30	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX31	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX36	*V. planifolia*	HPl21	HEM/UNICACH Living Collection
MEX38	*V. planifolia*	HPl3	HEM/UNICACH Living

			Collection
MEX39	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX40	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX41	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX46	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX47	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX48	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX5	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX51	*V. planifolia*	HPl22	HEM/UNICACH Living Collection
MEX52	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX54	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX55	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX56	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX59	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX6	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX65	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX66	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX67	*V. planifolia*	HPl16	HEM/UNICACH Living Collection
MEX69	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX7	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX70	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX72	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX75	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX76	*V. planifolia*	HPl3	HEM/UNICACH Living

			Collection
MEX77	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX79	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MEX80	*V. planifolia*	HPl16	HEM/UNICACH Living Collection
MEX9	*V. planifolia*	HPl3	HEM/UNICACH Living Collection
MO87	*V. planifolia*	HPl3	MO/MBG Living Collection
MO90	*V. planifolia*	HPl3	MO/MBG Living Collection
MO92	*V. planifolia*	HPl3	MO/MBG Living Collection
GB1789804118	*V. insignis*	HIn2	GenBank (Salazar,G.A.)
GB1789804119	*V. insignis*	HIn3	GenBank (Salazar,G.A.)
GB1789804120	*V. insignis*	HIn1	GenBank (Salazar,G.A.)
GB1789804122	*V. insignis*	HIn4	GenBank (Salazar,G.A.)
MEX50	*V. insignis*	HIn5	HEM/UNICACH Living Collection
MEX53	*V. insignis*	HIn5	HEM/UNICACH Living Collection
MEX61	*V. insignis*	HIn5	HEM/UNICACH Living Collection
MEX73	*V. insignis*	HIn5	HEM/UNICACH Living Collection
GB1517415475	*V. pompona*	HPo12	GenBank (Hu et al., 2019)
GB1546056	*V. pompona*	HPo3	GenBank (Mai & Coleman, 1997)
GB170284138	*V. pompona*	HPo14	GenBank (Pansarin, Salatino, & Salatino, 2008)
GB1789804143	*V. pompona*	HPo13	GenBank (Salazar,G.A.)
GB1789804144	*V. pompona*	HPo11	GenBank (Salazar,G.A.)
GB1789804145	*V. pompona*	HPo10	GenBank (Salazar,G.A.)
GB1789804146	*V. pompona*	HPo5	GenBank (Salazar,G.A.)
GB1789804147	*V. pompona*	HPo9	GenBank (Salazar,G.A.)
GB1789804148	*V. pompona*	HPo9	GenBank (Salazar,G.A.)
GB1789804149	*V. pompona*	HPo4	GenBank (Salazar,G.A.)
GB1789804150	*V. pompona*	HPo4	GenBank (Salazar,G.A.)
GB1789804152	*V. pompona*	HPo3	GenBank (Salazar,G.A.)
GB1789804153	*V. pompona*	HPo2	GenBank (Salazar,G.A.)
GB1789804154	*V. pompona*	HPo7	GenBank (Salazar,G.A.)
GB1789804155	*V. pompona*	HPo8	GenBank (Salazar,G.A.)
GB1789804156	*V. pompona*	HPo8	GenBank (Salazar,G.A.)
GB1789804157	*V. pompona*	HPo1	GenBank (Salazar,G.A.)
GB1789804158	*V. pompona*	HPo8	GenBank (Salazar,G.A.)

GB260401086	*V. pompona*	HPo7	GenBank (Belanger & Havkin-Frenkel, 2010)
GB260401087	*V. pompona*	HPo7	GenBank (Belanger & Havkin-Frenkel, 2010)
GB260401088	*V. pompona*	HPo7	GenBank (Belanger & Havkin-Frenkel, 2010)
GB260401089	*V. pompona*	HPo7	GenBank (Belanger & Havkin-Frenkel, 2010)
GB260401090	*V. pompona*	HPo7	GenBank (Belanger & Havkin-Frenkel, 2010)
GB6652551	*V. pompona*	HPo6	GenBank (Belanger & Havkin-Frenkel, 2010)
MEX11	*V. pompona*	HPo3	HEM/UNICACH Living Collection
MEX42	*V. pompona*	HPo3	HEM/UNICACH Living Collection
MEX44	*V. pompona*	HPo3	HEM/UNICACH Living Collection
MEX45	*V. pompona*	HPo3	HEM/UNICACH Living Collection
MEX49	*V. pompona*	HPo7	HEM/UNICACH Living Collection
MEX74	*V. pompona*	HPo3	HEM/UNICACH Living Collection
MEX8	*V. pompona*	HPo3	HEM/UNICACH Living Collection
MN221420	*V. madagascariensis*	NA	GenBank (Kim et al., 2020)

Assessing the climatic niche of haplotypes

Variation in climatic niche of Mexican vanilla was inferred from corresponding climatic variables following the procedure in Bone et al. (2015). The 19 bioclimatic variables reflecting temperature and precipitation regimes were retrieved for each occurrence from WorldClim at a resolution of 2.5 minutes (Fick & Hijmans, 2017). Because the BioClim variables were highly correlated with each other, a principal component analysis (PCA) was performed to summarize them as highly explanatory

eigenvalues using the R package ‘vegan’ (Oksanen et al., 2019). BioClim variables were standardized, and eigenvalues were plotted.

Sample coordinates were used to build a haplotype richness map to visualize diversity throughout the sampling landscape. The R package ‘raster’ (Hijmans, 2019a) was used to create a blank raster at a resolution of 22km to record the number of haplotypes per geographic cell following the methods of Buerki et al. (2015). Coordinates were then used to assess pairwise distances among samples with the R package ‘geodist’ (Padgham & Sumner, 2020). Pairwise geographic and genetic distances were tested for correlation among Mexican vanilla samples falling within the same clade by calculating Pearson’s correlation coefficient using the *cor* base R function.

Results

Sampling and DNA sequences

Overall, 191 DNA sequences were used for ITS analyses: 58 from Mexican field-collected vanilla samples, 10 from *ex-situ* living collections, and 123 from GenBank accessions. For *rbcL* analyses, 108 DNA sequences were used: 54 from Mexican field-collected vanilla samples, 14 from *ex-situ* living collections, and 40 from GenBank accessions (Appendix D). Information on source identity, type, and origin was obtained from 50 Mexican field-collected vanilla samples (Table 3.1). Of these, 40% were sourced from wild populations, mainly from Region 3, and 60% were sourced from cultivated populations, predominantly from Region 1 (Table 3.1). The identity of most samples was determined using descriptions from *Vanilla* cultivators and vegetative morphological traits (Soto Arenas and Cribb, 2010) due to the absence of flowers when these were collected. Most samples were generally identified as “Vanilla” or “Planifolia” by

cultivators, although some were specified as *V. insignis* or *V. pompona*. Vegetative morphological traits allowed for the verification of *V. insignis* samples due to discernable grooved stems and *V. pompona* due to measurably thick stems and leaves (Soto Arenas & Cribb, 2010), although these were not always congruent with cultivator assignments. In many cases, differentiation between *V. planifolia* and *V. pompona* using vegetative morphological characters was uncertain due to the phenotypic variability exhibited by these plants at different developmental stages.

Species delimitation and haplotype identification

Phylogenetic inferences based on *rbcL* showed little genetic variation among all samples and lacked interspecific resolution (Appendix E and F). On the other hand, ITS phylogenetic inferences provided more polymorphism at inter and intra-specific levels supporting species delimitation and haplotype identification. With strong support from both maximum likelihood and Bayesian phylogenetic analyses, ITS inferences placed all *ex-situ* samples into their corresponding *Vanilla* sections: *Vanilla, Tethya,* and *Xanata* (Appendix G and H). Analyses also revealed that Mexican vanilla samples clustered with strong statistical support (BS=100%, BPP >0.96) into three main clades (Figure 3.2). Within each clade, *ex-situ* samples exhibited mostly congruent taxonomy: the red clade consisting of *V. planifolia* as well as two *V. planifolia x pompona* hybrids, the green clade consisting of all *V. insignis,* and the blue clade consisting mostly of *V. pompona,* but also one *V. planifolia x pompona* hybrid and two samples previously identified as *V. planifolia.*

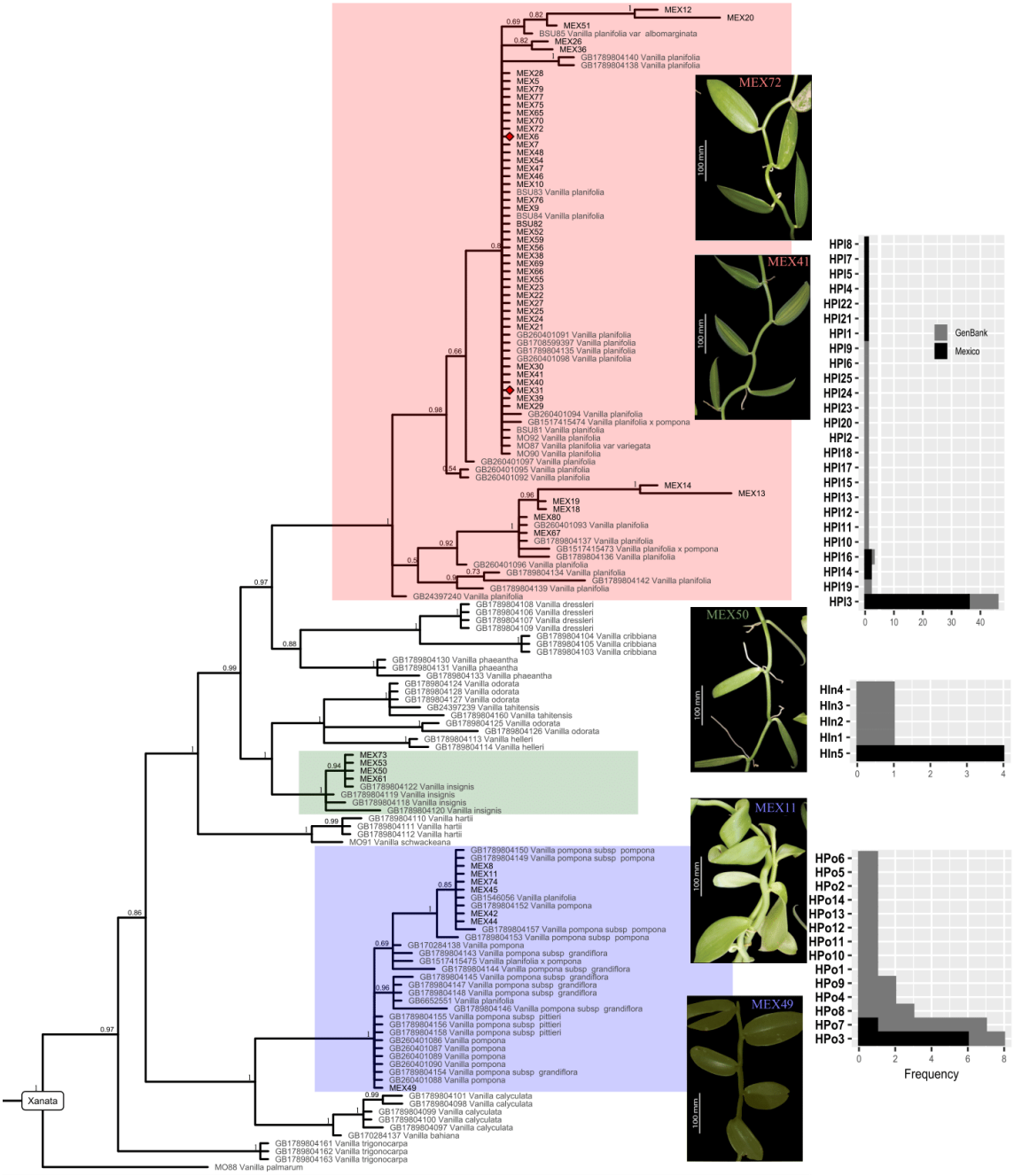


Figure 3.2 A) Phylogenetic tree of *Vanilla* section *Xanata* resulting from the Bayesian analysis of ITS sequences from Mexican samples and available GenBank accessions. The three main groups containing Mexican samples are highlighted: *V. planifolia* (red), *V. insignis* (green), and *V. pompona* (blue). Within each clade, morphological variation is shown and haplotype frequencies are plotted. Posterior probabilities greater than 0.75 are displayed on nodes. Samples exhibiting the non-synonymous substitution for the *rbcL* gene are indicated with red diamonds on the tree.

K2P genetic distances between the only clade recovered from phylogenetic analyses of *rbcL,* which contained all Mexican *Vanilla* samples, and *ex-situ* accessions showed low genetic variation (around ten times smaller than ITS K2P distances; see below) and overlapping distributions between multiple species (Appendix I). ITS K2P distances, however, predominantly supported species delimitations between each of the three clades recovered from phylogenetic analyses of ITS and showed genetic variation within two clades (Figure 3.3). From the red clade, Mexican vanilla exhibited a range of K2P genetic distances between *ex-situ V. planifolia* accessions, averaging at 0.014, with the majority (87%) falling below 0.026 and the most distant at 0.095 between the accession GB1546056. Two *ex-situ* accessions of *V.* x *tahitensis* and six of *V. odorata* resulted in distances below 0.026. Within the green clade, K2P genetic differences between Mexican vanilla samples and *ex-situ* accessions resulted in narrow distributions among all species, with those between *V. insignis* accessions averaging at 0.006. Within the blue clade, the distribution of K2P genetic distances showed two or more peaks among all *ex-situ* species, suggesting a dichotomy within the clade. Distances between Mexican vanilla samples within this green clade and *V. pompona ex-situ* accessions averaged at 0.009. Additionally, low K2P genetic distances (<0.014) resulted from 15 *V. planifolia ex-situ* accessions.

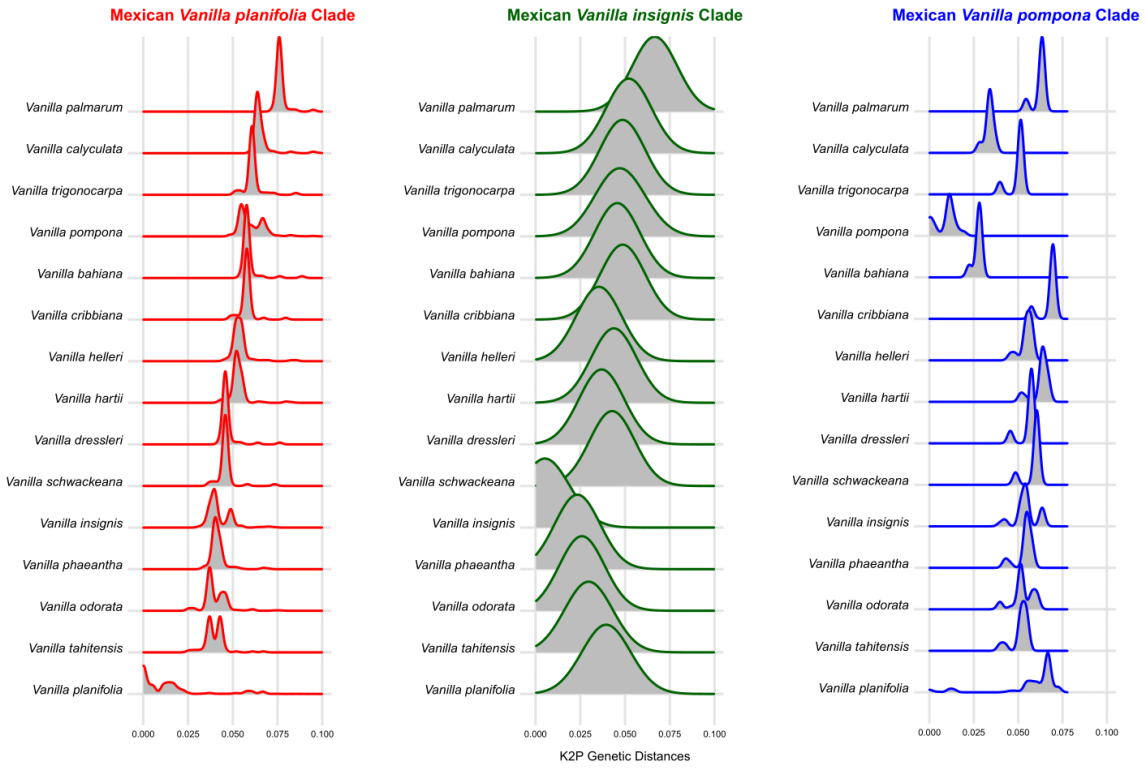


Figure 3.3 Density plots of K2P genetic distances between *ex-situ Vanilla* accessions and Mexican samples from the *V. planifolia, V. insignis,* and *V. pompona* clades.

Genetic variation of clades was summarized by identifying haplotypes within the three clades (Table 3.2; Figure 3.2). Within the largest red clade, 25 haplotypes were identified overall, but only ten comprised Mexican vanilla samples (Figure 3.2). Of these, most samples (61%) and most Mexican vanilla samples (78%) are assigned to haplotype HPl3. Samples within this haplotype exhibited notable phenotypic variation including the most common phenotype, 'Mansa', along with the varieties 'Albomarginata' and 'Variegata' (Figure 3.2). Within the green clade, four haplotypes were identified, and all Mexican vanilla samples belonged to the haplotype HIns4. Within the blue clade, 11 haplotypes were identified with six Mexican vanilla samples belonging to HPo3 and one to HPo6, exhibiting distinct phenotypes, both with thicker leaves and stems, but one

(HPo6) with more elliptic leaves (Figure 3.2). From analyses of the *rbcL* gene, only three haplotypes were identified from all Mexican vanilla samples where two haplotypes exhibited only one nucleotide substitution. When nucleotide sequences of these haplotypes were translated into amino acids, a synonymous substitution was detected for one haplotype and a non-synonymous substitution, from coding for aspartic acid to tyrosine, for the other. Samples exhibiting the non-synonymous substitution for the *rbcL* gene were shown to belong to the red clade and the most common HPl3 haplotype. They are represented within the ITS phylogenetic tree with the red diamonds (Figure 3.2).

Genome skimming and phylogenetic reconstruction of haplotypes

Full plastome and partial rRNA sequences were reconstructed *de novo* from genome skimming data, resulting in aligned sequence length of 161,656 bp and 4,949 bp, respectively (https://doi.org/10.5061/dryad.m905qfv3q). Within resulting rRNA phylogenetic trees inferred using both Bayesian and maximum likelihood methods, three clades were well-supported (BS ≥ 80 and BPP >0.99). Additionally, within the resulting plastome phylogenetic trees, three clades nested within each other had high statistical support (BS =100 and BPP=1). Placement of Mexican vanilla samples within the clades of the plastome and rRNA phylogenetic trees were not congruent (Figure 3.4). One notable sample, MEX67, was isolated within the plastome tree, but grouped with six other samples within the rRNA phylogenetic tree. Conversely, the sample Mex51 was isolated within the rRNA tree, but grouped with two other samples within the plastome tree (Figure 3.4).

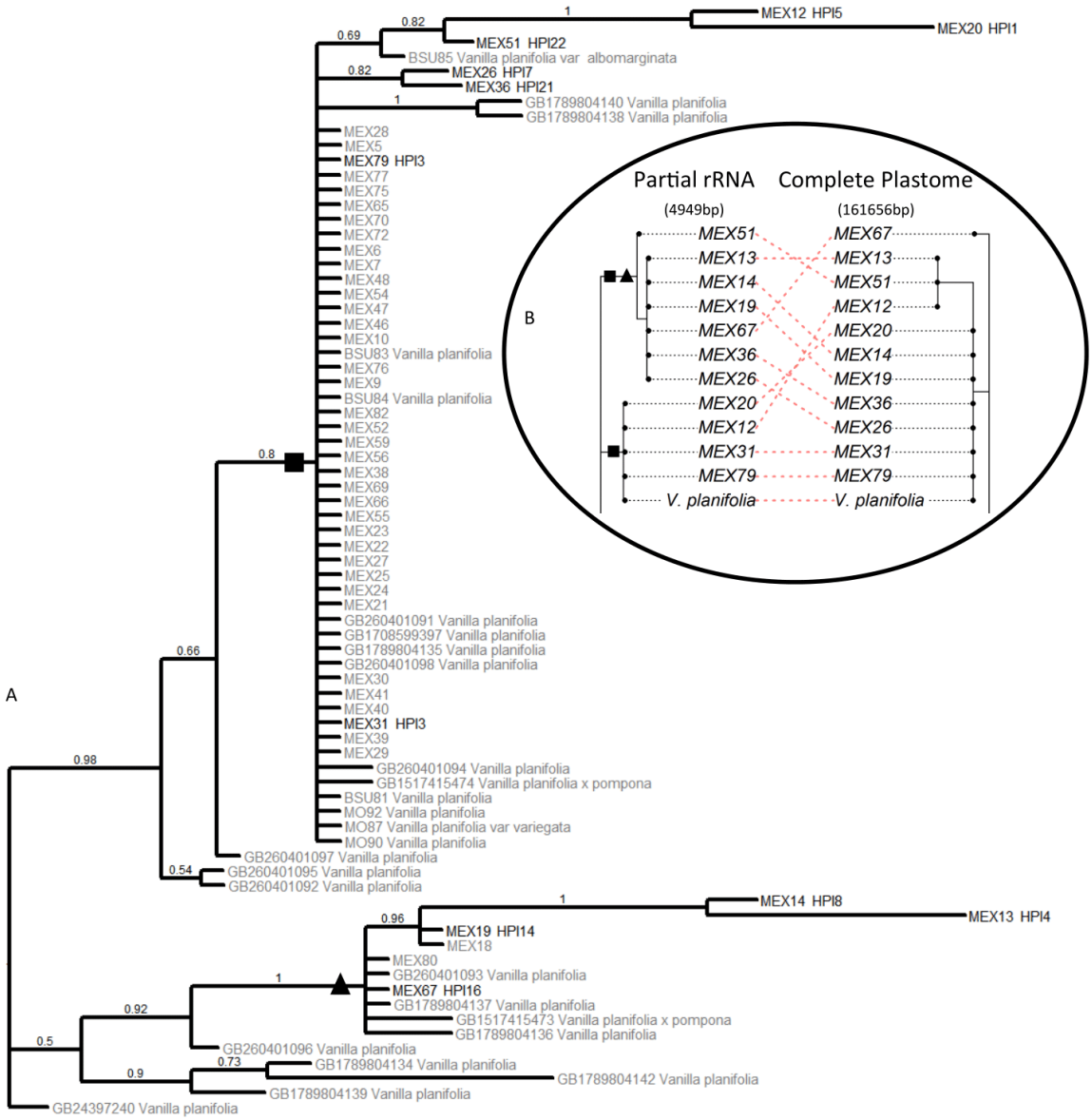


Figure 3.4 A) Bayesian maximum credibility tree based on ITS of the *Vanilla planifolia* clade. Labels in black show samples, along with their haplotype, that were used in B) subsetted phylogenetic analyses of complete plastomes and partial rRNA. Samples within well-supported (> 0.99) clades within the rRNA tree did not completely correspond with clades from the plastome tree, and vice-versa. All Mexican samples from one clade (denoted by the triangle) within the ITS phylogenetic tree, occurred within the same clade of the rRNA phylogenetic tree, however, samples from the other clade (denoted by the square) were spread throughout both clades of the rRNA tree.

Assessing the climatic niche of haplotypes

The climatic niche of all field collected samples was inferred using bioclimatic variables associated with each geo-location. Results from the PCA showed that climatic niche varied among samples, partitioning into three general niches (Figure 3.5). The first two axes of the PCA explained 79.6% of the variance in the data. The most important variables contributing to PC1 were: BIO4 (temperature seasonality), BIO12 (annual precipitation), BIO13 (precipitation of wettest month), BIO16 (precipitation of wettest quarter). The main variables contributing to PC2 are: BIO1 (annual mean temperature), BIO5 (maximum temperature of the warmest month), BIO6 (minimum temperature of the coldest month), and BIO10 (mean temperature of warmest quarter). The contribution of each variable to the first two principal components is available in Appendix J. The climatic niche of samples collected from the northernmost Region 1 grouped together and were positively associated with the temperature-related variables that predominantly influenced the PC2 axis and negatively associated with the precipitation-related variables corresponding to the PC1 axis. Comparatively, the climatic niche of samples collected from the southernmost Region 3 were positively associated with the temperature and precipitation related variables corresponding to both axes. The climatic niche of samples collected from Region 2 varied more considerably than those located within other regions but were mostly negatively associated with the temperature and precipitation related variables corresponding to both axes (Figure 3.5).

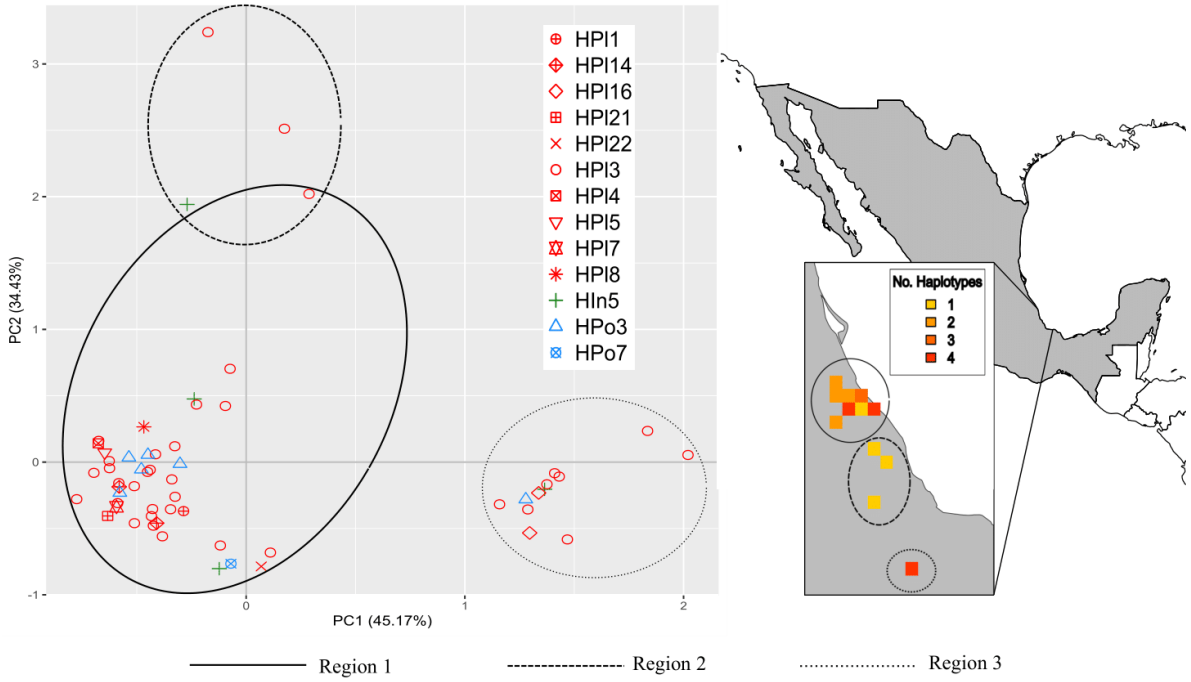


Figure 3.5 A) Principal components analysis of 19 bioclimatic associated with the collection locations of each Mexican vanilla sample. Each haplotype is indicated with a different symbol, with three colors identifying the three clades: *V. planifolia* (red), *V. insignis* (green), and *V. pompona* (blue). B) Haplotype frequency within the three sampling regions.

The predominant HPl3 haplotype occurred in all identified climatic niches and were sources from both wild and cultivated populations throughout all three geographic regions. While most other haplotypes within the red clade (HPl1, HPl4, HPl5, HPl7, HPL8, and HPl14) were constrained to only one partition of the climatic niche, which corresponded to the geographic Region 1. One haplotype consisting of five samples, however, was constrained only to the climatic niche partition corresponding to Region 3. Within the green clade, the Hin4 haplotype also occurred within each of the three climatic niches and geographic regions. Haplotypes within the blue clade, HPo3 and HPo6, corresponded to the Regions 1 and 3 and their respective climatic niches. Haplotype richness was found to be highest in Regions 1 and 3. Both samples with synonymous mutations of the *rbcL* gene were sourced from wild populations and

exhibited the climatic niche from Region 3, while the two samples with non-synonymous mutations were sourced from cultivated populations and exhibited the climatic niche from Region 1. No correlation was revealed between genetic distance and geographic distance.

Discussion

Three cultivated Mexican *Vanilla* species

Results from this study indicate that the phenotypic variation in vanilla cultivated within Mexico is not only attributed to phenotypic plasticity, but also to species diversity and intra-specific haplotype diversity. Phylogenetic analyses of the ITS region recovered three clades of Mexican vanilla, indicating the presence of at least three species in cultivation: *V. planifolia, V. pompona,* and *V. insignis*. Inter- and intra- specific K2P genetic distances of taxa within this study additionally supported the delimitation of *Vanilla* species based on findings from Qin et al. (2017). Within all three clades containing Mexican vanilla samples, the majority of interspecific K2P genetic distances between field-collected samples and GenBank accessions fell below the average genetic distance of sister species threshold of 3.98% advocated for angiosperms by Qin et al. (2017; Figure 3.3). This serves as additional evidence to support species delimitation within *Vanilla*. One field collected sample belonging to haplotype HPl4 in the Papantla region and four accessions from GenBank, one of which was labeled as a *V. planifolia x pompona* hybrid, exceeded this threshold and may therefore be the result of hybridization and introgression between *V. planifolia* and its crop wild relative (Figure 3.3). This possible hybrid origin is supported by similar reproductive biology, phenology (Pansarin, 2016), and chromosome numbers (median n = 16; Rice et al., 2015), as well as variable

intra-specific ploidy levels of *V. planifolia, V. pompona,* and other members within the subgenus *Xanata* (Bory, Catrice, et al., 2008; Jones et al., 1998). Given that only *V. planifolia* and *V. tahitensis* are recognized for commercial vanilla production, the reconsideration of species' requirements to include congeneric species, such as the ones identified within this study, may offer novel alternative sources for vanilla production, and catalyze more inclusive conservation strategies for *Vanilla.*

High haplotype diversity within *V. planifolia*

Phenotypic variability and haplotype diversity was revealed within Mexican accessions of *V. planifolia* and *V. pompona,* but none within *V. insignis* (although this could be due to the low sample size for this species; Figure 3.2). Of the three identified species, the highest haplotype variation was found within Mexican *V. planifolia,* which consisted of ten haplotypes, while *V. pompona* consisted of two haplotypes, and *V. insignis* of only one (Figure 3.2). The two haplotypes found within *V. pompona* exhibited distinct morphology, with leaves more elliptic in HPo3 and more oval in the one sample (MEX49) belonging to HPo7, and grouped with high statistical support into two clades (Figure 3.2). E*x-situ* samples identified as the two accepted subspecies, *V. pompona* subsp. *pittieri* (Schltr.) Dressler and *V. pompona* subsp. *grandiflora* (Lindl.) Soto Arenas (Soto Arenas & Dressler, 2010) were paraphyletic within the *V. pompona* clade, thus haplotypes could not be assigned with confidence to a subspecies (Table 3.2, Figure 3.2). These results call for further taxonomic work within the *V. pompona* complex. The high haplotype diversity uncovered within *V. planifolia* in this study supplements previous findings of endemic haplotypes stemming from wild populations in regions within the species' extended native distribution (Azofeifa-Bolaños et al., 2017; Schlüter et al., 2007;

Villanueva-Viramontes et al., 2017). These haplotypes may offer important genetic resources to support crop improvement in the face of climate change.

Phylogenetic analyses using genomic data increased the resolution of clades within *V. planifolia,* however, results from analyses using rRNA were not completely congruent with those using only ITS (Figure 3.4). Within the ITS phylogenetic tree, all samples belonging to one clade (denoted by the black triangle symbol in Figure 3.4) were also found to be within the same clade within the rRNA phylogenetic tree. The other clade within the ITS phylogenetic tree (denoted by the black square in Figure 3.4), however, consisted of samples from both clades within the rRNA tree. Phylogenetic comparisons of complete plastome and partial rRNA haplotype sequences suggest past events of hybridization and introgression. Phylogenetic reconstruction of plastomes revealed three well-supported clades, which did not align with the clades resulting from the phylogenetic reconstruction of rRNA (Figure 3.4). Chloroplast capture from a sister species may explain this observed plastome sequence variability as is evidenced more prominently in *Piper* (Simmonds et al., 2021).

Characterizing the predominant haplotype using ITS

Encompassing multiple *ex-situ V. planifolia* samples, the haplotype HPl3 probably comprises most vanilla cultivated outside of its native range (Figure 3.2) and exemplifies the degree to which genetic diversity is limited within globally cultivated vanilla, as shown by previous genetic studies (Bory, Grisoni, et al., 2008; Y. Hu et al., 2019; Lubinsky, Bory, et al., 2008; Villanueva-Viramontes et al., 2017). Phenotypic variability is also observed within this haplotype, as evidenced by the distinctive leaf morphology observed among field-collected samples, some of which his shown in Figure

3.2, as well as the inclusion of *ex-situ* 'Albomarginata' phenotypes. In addition, results from the *rbcL* analysis identified two samples within this predominant ITS haplotype, which contained a non-synonymous mutation within the coding region. Additional analyses will be required, e.g. sequencing and assembling the full protein to identify potential changes in catalytic rate (Sen et al., 2011), to ascertain the effect of this mutation on the plants' fitness and potentially uncover the signature of natural selection or local adaptation. Additional genomic research will be needed to further assess variability within this predominant haplotype and identify the mechanisms underpinning its phenotypic variability and ability to adapt to a changing climate.

Phylogenetic, geographic, and climatic characteristics suggest local and widespread landraces

The inclusion of external samples and widespread cultivation throughout the extent of the sampling region and climatic niches suggests that the *V. planifolia* haplotype HPl3 is the predominant cultivated vanilla in Mexico and has been traded by humans within Mexico and globally (Figure 3.5). The main indigenous groups of each region (the Totonaco, Nahuatl, and Chinanteco in Regions 1, 2, and 3, respectively; Geographic, n.d.) are likely to have played a major role in the domestication of regional haplotypes and the trade of this predominant *V. planifolia* haplotype. Our results, however, do not indicate directional dispersal of this haplotype because individuals are reported to be sourced both from wild populations in Regions 1 and 2 (Table 3.1). Additionally, it is necessary to emphasize that the long histories of vanilla cultivation in Mexico, which blur the lines between wild, cultivated, and feral populations, complicate inferences on directional dispersal and domestication.

Most observed haplotype variation occurs within the northernmost Region 1 around Papantla (Figure 3.5), which is credited to be the origin of cultivated vanilla and boasts a long pre-Hispanic history of cultivation by the Totonaco people (Rain, 2004). Haplotypes within this region are genetically related to the predominant HPl3 haplotype, yet still distinct, suggesting that multiple domestication events may have occurred in this area stemming from locally available wild populations that may no longer exist. In addition, their co-occurrence with *V. insignis* and *V. pompona* haplotypes suggests the possibility of hybridization between congeners. Haplotype richness was also high in Region 3 where the majority of samples were sourced from local wild populations and where, similarly, all three species were observed (Figure 3.5; Table 3.1). Two samples exhibiting synonymous mutations of the *rbcL* gene were sourced from locally available wild populations in this region, which may suggest it as the natural origin of his haplotype. While other haplotypes are also found to grow wild in this region, they are not restricted to this area but occur throughout the sampling range.

Due to ecogeographic constraints and long histories of regional cultivation, the haplotypes identified within this study could be considered landraces and should be prioritized in conservation planning. Further research incorporating archeological and genetic data is needed to test the hypothesis of multiple origins resulting in genetically distinct landraces. Domestication syndrome traits of crops that have been predominantly propagated are less apparent than those of sexually propagated crops and are often further obscured by phenotypic plasticity and/or spontaneous sexual reproduction with sympatric wild or cultivated populations (Denham et al., 2020). Furthermore, cultivated vanilla may not necessarily be morphologically distinct from its wild relatives and therefore is only

considered ‘semi-domesticated’, as described by Meyer et al. (2012). Nonetheless, regional haplotypes identified here may fit within a more inclusive definition of landraces, including the dynamic domestication of regional cultivated varieties influenced by local human culture (Casañas et al., 2017). Diverse landraces within a crop’s native distribution provide an important source of genetic diversity to potentially increase its capacity to cope with environmental change (Bellon et al., 2015). Maintaining agricultural diversity and conserving landraces within vanilla’s center of domestication benefits the livelihoods of farmers as well as global vanilla production.

Perspectives on regional sustainability of vanilla

Vanilla has played a prominent role in the culture and economy of Mexico (especially in Region 1) for thousands of years, and although vanilla is now cultivated globally, Papantla remains to be known as “the city that perfumes the world” (Rain, 2004). Climate change proves a serious and imminent threat to vanilla cultivation in this culturally and genetically important region. Over the past ten to twenty years, the duration and intensity of droughts has been increasing (Varela et al., 2021). Annually intensifying conditions has led to drought stress in traditionally cultivated systems and has increased susceptibility to deadly epidemics of the fungus *Fusarium* spp. and other pests (Pinaria et al., 2010). Due to six months of extreme drought in 2019, over 90% of the vanilla harvest was lost in the state of Veracruz (Agropecuaria, 2019). On top of increasingly unsuitable environmental factors, farmers continue to face the challenges of a volatile market, thievery of the expensive fruits, and little to no governmental support. In many communities, the tradition of small-scale vanilla farming has been left to the elders of families and the vocation is slowly fading in favor of less risky agricultural

pursuits. Within the survey conducted for this study, most farmers recounted stories of wild vanilla in their region, however, few believe that these populations still exist. A continued trend of deforestation and agricultural land-use changes (Bonilla-Moheno & Aide, 2020; Dalrymple, 2006) threaten the remaining wild populations and the regional landraces within their narrow distribution, adding urgency to improve conservation. Without increased support for local farmers and sustainable farming practices, the region so well-known for vanilla may be losing a key part to its culture and economy, and the genetic resources of this globally important crop may be further diminished (Donatti et al., 2019).

Conclusion

Haplotypes identified within this study provide a starting point for future studies to further assess congeneric species delimitation, intraspecific genetic variation, and phenotypic plasticity within regional vanilla. In addition, a deeper understanding of the genetic resources and landraces of this important crop will benefit local farmers as well as the global vanilla production. The haplotype composition and frequency within vanilla's center of origin represents the importance of *V. planifolia* relatives, especially *V. pompona* and *V. insignis*, within regionally cultivated systems, as biodiversity and hybridization may be key factors to maintaining genetic diversity within this crop. While the DNA sequence data used here offered a glimpse into its genetic variability, further genomic research on Mexico's haplotypes, landraces, and wild populations is needed to uncover vanilla's genetic basis for phenotypic variability and capacity to cope with environmental changes and to identify populations better adapted to drought. Conservation of the variation of vanilla's genetic resources within this region, both wild

and cultivated, is crucial for advancements in crop improvement and ensuring the future sustainability of this beloved spice.

CHAPTER FOUR: GENOMIC INSIGHTS INTO CULTIVATED MEXICAN *VANILLA PLANIFOLIA* REVEAL HIGH LEVELS OF HETEROZYGOSITY STEMMING FROM HYBRIDIZATION

The final version of this article has undergone full peer review and has been published. Please see:
Ellestad, P., Pérez-Farrera, M. A., & Buerki, S. (2022). Genomic Insights into Cultivated Mexican Vanilla planifolia Reveal High Levels of Heterozygosity Stemming from Hybridization. *Plants 2022, Vol. 11, Page 2090*, *11*(16), 2090.
https://doi.org/10.3390/PLANTS11162090

Abstract

Although vanilla is one of the most valuable spices, there is a lack of understanding of the genomic variability of the main vanilla producing species, *Vanilla planifolia*, within its cultivated origin, Mexico. High genomic heterozygosity levels within the globally cultivated "Daphna" genome have raised questions on the possibility of a hybrid origin and analogous genomic signatures of vanilla cultivated within its origin. This study investigated these questions by assessing whether the genomic structure of Mexican *V. planifolia* reflected domestication events. Whole genome re-sequencing was used to compare genome complexity between 15 cultivated accessions from different regions and gene pools. Results showed high levels of heterozygosity, ranging from 2.48% to 2.85%, in all but one accession, which exhibited a low level (0.403%). Chromosome-level comparative analyses revealed genomic variability among samples, but no signals of chromosome rearrangements. These findings support the hypotheses that cultivated vanilla resulted from hybridization and that multiple

domestication events have shaped cultivated vanilla leading to the formation of landraces. High cultural diversity within this region further supports the occurrence of multiple domestication processes. These results may help to improve breeding and conservation efforts aiming to preserve the genetic diversity of this beloved spice threatened by climate change

Introduction

Vanilla planifolia Andrews is a tropical vine of the family Orchidaceae, which produces vanilla, one of the most widely known and valuable spices worldwide (Bruman, 1948). With a cultivated origin in Mexico, it has been introduced across the globe to be cultivated for use in the culinary, cosmetic, and medicinal industries (Bruman, 1948; Lubinsky, Bory, et al., 2008). By country, vanilla production is currently led by Madagascar, followed by Indonesia then Mexico (Food and Agriculture Organization of the United Nations, 2020). *Vanilla planifolia* is self-compatible, but incapable of self-fertilization without natively co-occurring pollinators (Bory, Grisoni, et al., 2008). Outside of its native range, a labor-intensive technique is required to manually pollinate the flower (Cameron, 2011). Inhibiting natural genetic recombination, manual self-pollination and clonal vegetative propagation practices have resulted in low genetic diversity within the cultivated species, overall hindering its ability to cope with changing environmental conditions (Bory, Grisoni, et al., 2008; Chambers et al., 2021; Y. Hu et al., 2019; Lubinsky, Bory, et al., 2008; Soto Arenas & Dressler, 2010). On top of increasing drought conditions (Varela et al., 2021) and fungal outbreaks associated with climate change (Pinaria et al., 2010), the rapid loss of wild populations due to land-use change, habitat fragmentation, and illegal harvesting poses an immediate and irreversible threat to

the preservation of genetic variation within this crop (Hinsley et al., 2018; Soto Arenas & Dressler, 2010). Genetic resources within *V. planifolia'*s cultivated center of origin may offer a novel gene pool to increase the genetic diversity within the species and ensure crop sustainability under future climate scenarios. Analyzing the genomic structure of regionally cultivated vanilla in Mexico will offer important insight into this crop's genetic resources and a better understanding of the processes leading to its domestication.

Ancient and contemporary cultural groups have shaped vanilla in its center of cultivation for centuries. Historical records indicate that vanilla was used as a flavoring and medicinal beverage by multiple cultures in Mesoamerica, including the Totonacs, the Mayans and the Aztecs (Bruman, 1948; Rain & Lubinsky, 2011). After the Spanish conquest of the Aztecs in 1520 AD, it was transported to Europe, but was not cultivated outside of its native range until 1832, when Edmond Albius, from Reunion Island, developed a technique for manually pollinating the flowers (Bruman, 1948; Rain, 2004). This human-mediated dispersal has led many researchers to believe that globally cultivated vanilla (i.e., cultivated outside of the species' native range (Ellestad et al., 2021) comes from a single origin in Mexico, specifically in the Papantla region, and this hypothesis has been supported by genetic data (Bory, Grisoni, et al., 2008; Lubinsky, Bory, et al., 2008; Minoo et al., 2008). Within Mexico, however, high levels of genetic variability have led to the hypothesis of multiple origins shaping regionally cultivated vanilla (Ellestad, Pérez-Farrera, Forest, et al., 2022; Lubinsky, Bory, et al., 2008; Ramos-Castellá et al., 2017; Schlüter et al., 2007; Villanueva-Viramontes et al., 2017), although these limited results have not been able to fully disentangle the native crop's evolutionary history. Challenges, rooting from unclear species boundaries, intra-specific phenotypic

variability, and congeneric hybridization, have hindered an accurate understanding of the processes that have shaped the genetic resources of vanilla in its origin. Additionally, the cultivation of multiple *Vanilla* species in Mexico (Ellestad, Pérez-Farrera, Forest, et al., 2022), which exhibit similar vegetative morphological characteristics, muddles inferences on the genetic resources of the main vanilla producing species, *V. planifolia.*

Recent advances in genomic sequencing technology and the publication of a reference genome (Hasing et al., 2020; Y. Hu et al., 2019) have helped to elucidate vanilla's genetic resources and uncover greater levels of genetic variation than previously expected (Chambers et al., 2021; Favre et al., 2022; Hasing et al., 2020; Y. Hu et al., 2019), providing more insights into the domestication processes that have affected vanilla cultivated both in its native and its global range. Various methods to infer genetic variation have exposed high levels of variability within *V. planifolia.* Single nucleotide polymorphism (SNP) analyses have revealed variation, clustering vanilla accessions into three main groups (types 1-3), with accessions cultivated in Mexico clustering into only two (Chambers et al., 2021). Furthermore, within cultivated *V. planifolia* in Mexico, haplotype variation, inferred using ITS, was revealed, uncovering ten different haplotypes (Ellestad, Pérez-Farrera, Forest, et al., 2022). At the population level, a clear demarcation in observed heterozygosity (Ho) was found between cultivated and wild *V. planifolia*, where cultivated vanilla exhibited substantially higher levels (Favre et al., 2022).

For examining an organism's genetic diversity and evolutionary history, genome-wide patterns of heterozygosity offer a valuable metric. Using GenomeScope, a recently developed software designed to assess the relative abundance of homozygous and

heterozygous sequences within k-mer frequency distributions (Ranallo-Benavidez et al., 2020; Vurture et al., 2017), recent studies have reported genome-wide heterozygosity levels of globally cultivated *V. planifolia* to be 2.32% (Y. Hu et al., 2019) and 2.47% (Hasing et al., 2020), therefore suggesting this species to be highly outbred. These high levels found within *V. planifolia* cultivated outside of its native range raise the questions of what evolutionary processes contributed to this genomic structure and whether vanilla cultivated in its origin exhibits the same genomic signals. It has been hypothesized that these high levels of heterozygosity within cultivated vanilla were attributed to the accumulation of somatic point mutations brought about by clonal propagation (Favre et al., 2022), as shown in *Manihot esculenta* Crantz (McKey et al., 2010). The extent of these levels, however, points to the contribution of additional, more effecting, evolutionary processes, such as hybridization and/or polyploidization.

Hybridization has previously been suspected as a contributing agent to phylogenetic incongruences between nuclear and plastid signals (Ellestad, Pérez-Farrera, Forest, et al., 2022) and chromosomal abnormalities (Nair & Ravindran, 1994; Ravindran, 1979) within cultivated vanilla, and may additionally offer an explanation for these high levels of heterozygosity. Hybridization has been shown to occur between *V. planifolia* and six species: *V. pompona* Schiede (Delassus, 1963), *V. aphylla* Blume (Divakaran et al., 2006), *V. odorata* C. Presl, *V. x tahitensis* J.W. Moore (Lubinsky, Cameron, et al., 2008), *V. phaeantha* Rchb.f. (Y. Hu et al., 2019), and *V. palmarum* Salzm. ex Lindl (J. Li et al., 2020). Owing to the ancient and contemporary cultivation histories in Mexico, synthetic hybridization events between local congeners is a likely possibility. On the other hand, natural polyploidization has also been shown to occur

within cultivated *V. planifolia* (Bory, Catrice, et al., 2008) and could explain unexpectedly high levels of genome-wide heterozygosity within some individuals, although it is unlikely that these phenomena would occur in widespread cultivation.

Within this study, we aimed to explore the evolutionary mechanisms underpinning the high levels of genome-wide heterozygosity in vanilla and shed light onto the evolutionary processes that have affected this crop in its cultivated center of origin, Mexico. Due to the phenotypic and genetic variation observed in Mexico and the long histories of regional cultivation by different ethnic groups, we hypothesized that the gene pool of cultivated vanilla in Mexico has been influenced by multiple domestication events. On top of that, due to the extent of genome-wide heterozygosity levels found within globally cultivated vanilla, we hypothesized that cultivated vanilla stems from a hybrid origin. To assess if the genome structure of regionally cultivated haplotypes reflects domestication processes, we compared the genome structure of regionally cultivated vanilla against the available reference 'Daphna' genome (Hasing et al., 2020), evaluating genome-wide heterozygosity, ploidy, synteny, and SNP relatedness. To obtain a reference scale of genome-wide heterozygosity levels in plants, we first conducted a literature review to extract all genome-wide heterozygosity values inferred using the software GenomeScope and GenomeScope 2.0 (Ranallo-Benavidez et al., 2020; Vurture et al., 2017). Our sampling consisted of 15 plants cultivated around the main vanilla producing regions of Veracruz and Oaxaca, Mexico and encompassed the breadth of haplotypic and phenotypic diversity as inferred by Ellestad, Pérez-Farrera, Forest, et al. 2022). Genomic insights into cultivated *V. planifolia* in its origin will help shed new light

on the domestication processes and genetic resources of this beloved spice threatened by climate change.

Results

The data and reproducible workflow (the code, including citations and versions of all packages) associated with this study are available on GitHub (GitHub, n.d.), and a companion GitHub Pages website (*Vanilla Genomic Project*, n.d.) was developed to fully explain our analyses.

Review of plant levels of genomic heterozygosity inferred using GenomeScope

The query for studies that have used GenomeScope to infer genomic heterozygosity resulted in 455 publications deposited on PubMed, of these 142 pertained to plants (Table S1). For all plants assessed, the average level of genomic heterozygosity was found to be 1.59% (min 0.04%, max 12.02%; Figure 4.1A). For just diploid plants, the average was found to be 1.10% (min 0.07%, max 4.48%). Over half of the plants assessed in these studies were cultivated for human use (Figure 4.1B) and Orchidaceae was only represented by three other species (Figure 4.1A), therefore it should be noted that these values may offer a skewed scale of heterozygosity levels since genomic research on cultivated plant species often employs inbred and/or solely diploid accessions for genomic sequencing to effectively perform genomic tasks, such as read mapping and alignment. Nonetheless, the previously reported genomic heterozygosity levels (2.32% and 2.47%) for diploid *V. planifolia,* a predominantly vegetatively propagated crop, were comparatively high (Hasing et al., 2020; Y. Hu et al., 2019) (Figure 4.1A).

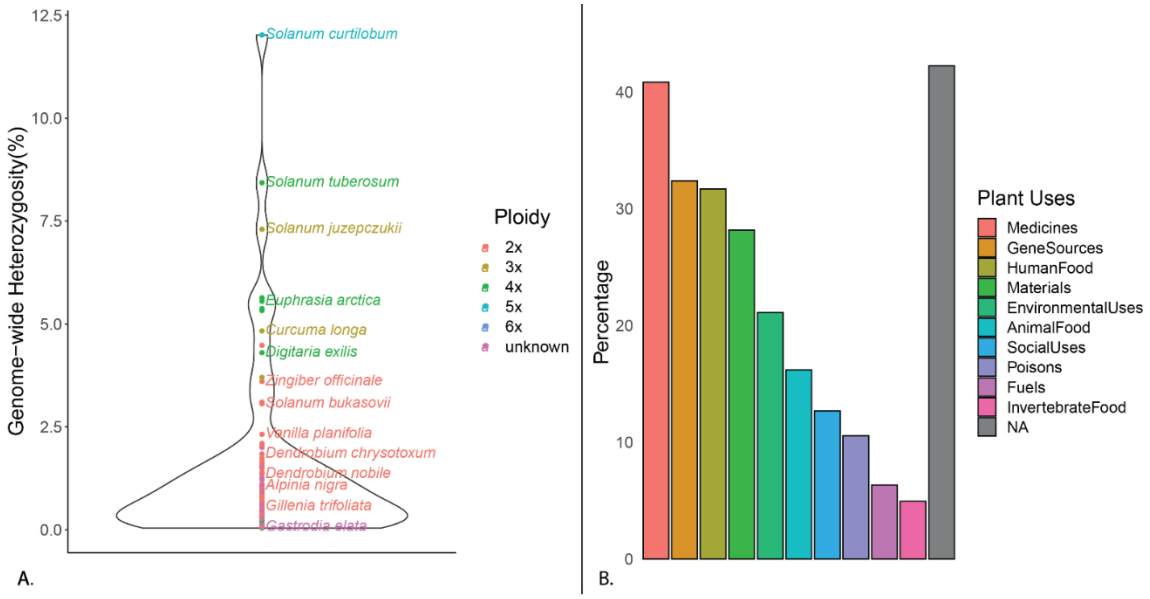


Figure 4.1 Summarized data of plant attributes extracted through a literature review of studies using GenomeScope and Smudgeplot A) Violin plot of genome-wide heterozygosity levels. B) Bar plot of human uses for each species studied.

Sampling, DNA extraction, and whole genome re-sequencing

Fifteen samples were collected from eight municipalities within the main cultivation regions in Mexico (Figure 4.2). Samples included 10 haplotypes inferred from ITS haplotype analyses (Ellestad, Pérez-Farrera, Forest, et al., 2022). Thirteen samples exhibited the most common 'Mansa' phenotype and two samples exhibited 'Variegata' phenotype, with yellow and green striped leaves, as described by Soto Arenas and Dressler (Soto Arenas & Dressler, 2010). Whole genome re-sequencing of quality extracted genomic DNA (see Methods for DNA concentration threshold) resulted in an average of 195 million reads (paired-end), yielding an average of 5.861 Gb per sample. Re-sequenced samples were found to have an average sequencing coverage of 80x (Table 4.1) to the reference genome *V. planifolia* 'Daphna', which had a genome length of 736,752,966 bp (Hasing et al., 2020). After trimming, samples had an average size of 13.4 Gb per paired-end read (Table 4.1).

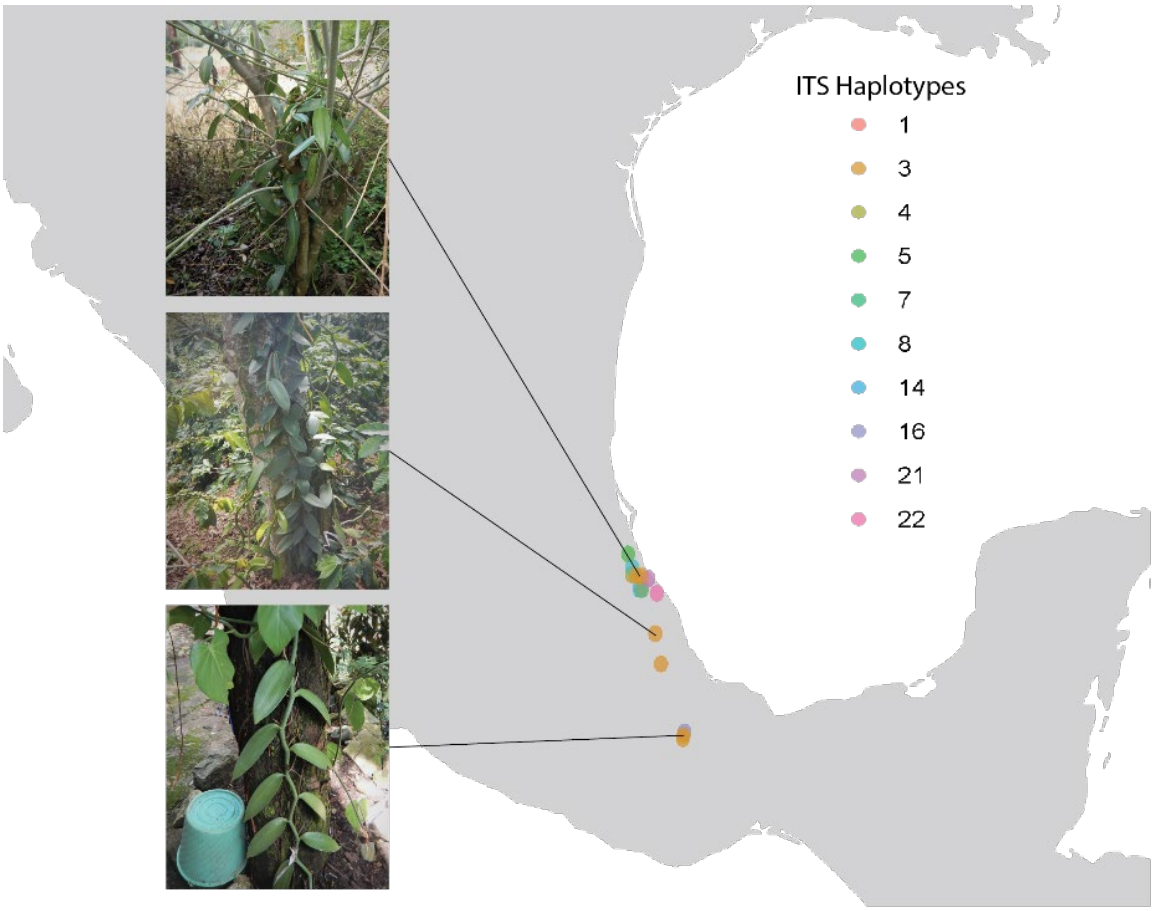


Figure 4.2 Map of sampling locations in Mexico. Point colors show haplotype IDs of samples (Ellestad, Pérez-Farrera, Forest, et al., 2022)**. From top to bottom, photos show samples MEX41, MEX59, and MEX67.**

Table 4.1 Attributes and identifiers of generated genomic data used in this study. Raw genome coverage was calculated based on the 1-C reference genome size (736,752,966 bp) from the BioProject PRJNA633886 published in Hasing et al. (2020a)**.**

Sample ID	BioSample	SRA	# Reads	Yield (Mbases)	Sequencing coverage (x)
MEX12	SAMN28632720	SRR19374418	187296525	5618.896	76.27
MEX13	SAMN28632721	SRR19374411	191490389	5744.712	77.97
MEX14	SAMN28632722	SRR19374410	186065732	5581.972	75.76
MEX19	SAMN28632723	SRR19374409	169276636	5078.299	68.93
MEX20	SAMN28632724	SRR19374408	185396535	5561.896	75.49
MEX26	SAMN28632725	SRR19374407	181142904	5434.287	73.76
MEX31	SAMN28632726	SRR19374406	187560091	5626.803	76.37
MEX36	SAMN28632727	SRR19374405	169478017	5084.341	69.01
MEX41	SAMN28632728	SRR19374404	220584559	6617.537	89.82
MEX51	SAMN28632729	SRR19374417	215541915	6466.257	87.77

MEX59	SAMN28632730	SRR19374416	206393194	6191.796	84.04
MEX65	SAMN28632731	SRR19374415	214461750	6433.852	87.33
MEX67	SAMN28632732	SRR19374414	194615015	5838.450	79.25
MEX69	SAMN28632733	SRR19374413	207909587	6237.288	84.66
MEX79	SAMN28632734	SRR19374412	213402841	6402.085	86.90

Genomic heterozygosity, ploidy, and complexity

Output from GenomeScope 2.0 analyses conducted on cleaned trimmed reads revealed similar genome-wide heterozygosity levels between the reference 'Daphna' genome and most Mexican *V. planifolia* samples, but a strong divergence was revealed with one Mexican sample, MEX67 (Figure 4.3, S1). Within 14 Mexican *V. planifolia* samples, heterozygosity levels were high, ranging from 2.48% to 2.85%. Contrastingly, MEX67 exhibited a much lower heterozygosity level of 0.403% (Table 4.2). Haploid genome size estimations all ranged between 513Mbp and 613Mbp. Heterozygous k-mer pair coverage distributions from Smudgeplot revealed signals of diploidy in all samples with an average of 97.3% of k-mer pairs in an AB ratio (Figure 4.4). Other ratios AABB, AAB, and AAABB were also found, but only in small percentages (<5%; Figures 4.4 and S2).

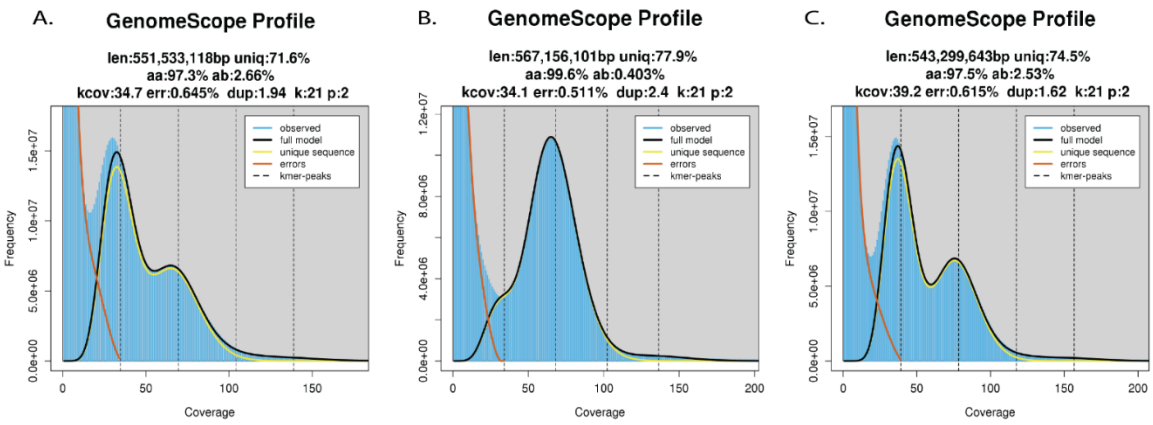


Figure 4.3 GenomeScope2.0 output showing variation in genome-wide heterozygosity levels among Mexican vanilla accessions: A) MEX41, B) MEX67 and C) MEX79.

Table 4.2 Genomic structure attributes of reference 'Daphna' genome and Mexican *V. planifolia* samples, inferred through GenomeScope2.0 and Smudgeplot (ploidy).

Sample ID	Ploidy	Genome-wide Heterozygosity (ab%)	K-mer Coverage (kcov)
'Daphna'	2x	2.48	99
MEX12	2x	2.74	32.3
MEX13	2x	2.67	32.6
MEX14	2x	2.66	34.7
MEX19	2x	2.52	32.4
MEX20	2x	2.56	33.7
MEX26	2x	2.57	31.7
MEX31	2x	2.61	32.5
MEX36	2x	2.52	28.2
MEX41	2x	2.77	40.9
MEX51	2x	2.62	39.3
MEX59	2x	2.79	38.5
MEX65	2x	2.57	16
MEX67	2x	0.403	34.1
MEX69	2x	2.85	38.9

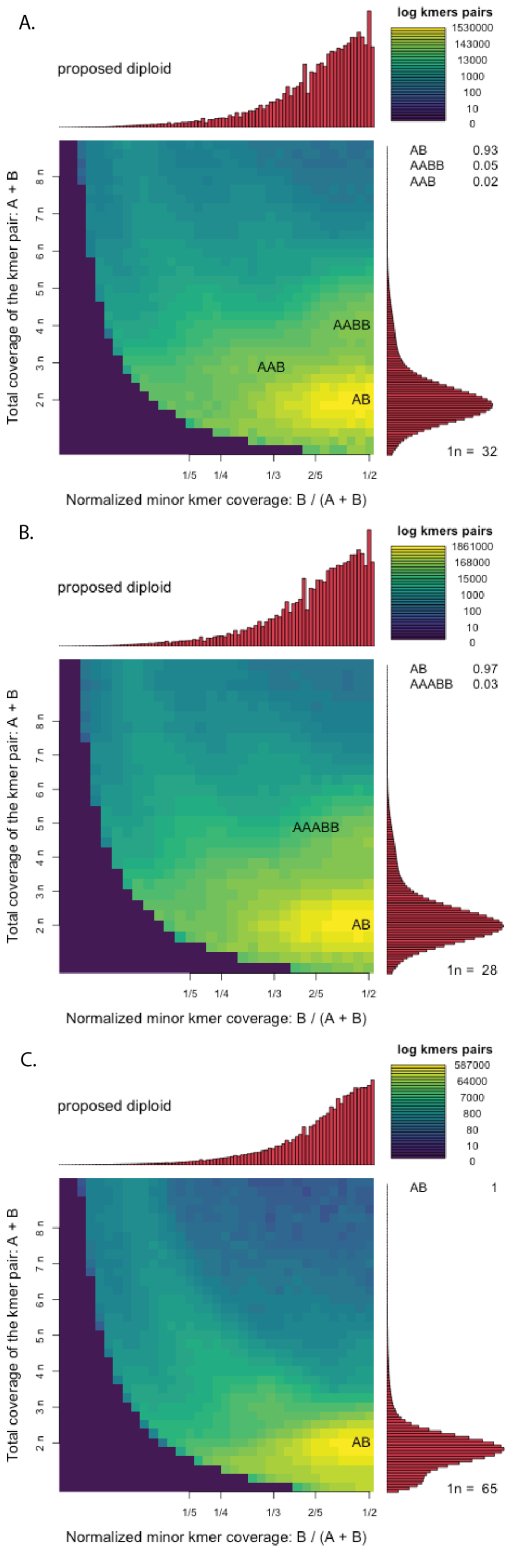


Figure 4.4 Smudgeplot output for A) MEX13, B) MEX26, and C) MEX67. The color intensity of the smudge indicates how frequently the haplotype structure is represented within each genome and the bar plots represent sequencing coverage.

Genome reconstructions to infer structural variation and synteny

Genomic alignments from MiniMap2 revealed that all reconstructed genomes exhibited full coverage on the 'Daphna' reference genome (Figure 4.5) and, in congruence with results from the dotplot analyses (Figure S3), suggested no chromosomal rearrangements among the accessions. Overall, genomic comparisons to the reference genome revealed a variation in structural similarities among Mexican samples with MEX67 exhibiting the most similarities (Figures 4.6 and S3). Genomic synteny to the reference genome, visualized using 'dotPlotly' (Figure S3), showed mean percentages of identity between 99.0 and 99.6% on all chromosomes of MEX67, while the rest of the samples showed much lower percentages of identities (98.4-99.0%). Among all samples, chromosome two (CM028151.1) matched the least to the reference genome (Figures 4.6, S3). Variation among samples was best visualized by the heat map of relative percentage identities by chromosome and further exemplified the extent of differences between most samples and the reference, especially on chromosome two (CM028151.1; Figure 4.6). Samples clustered into three main groups: the first consisting of MEX67; the second consisting of MEX65, MEX51, MEX79, and MEX41; and the third consisting of the remaining samples. Samples did not cluster by geography. Relative to other samples, MEX67 showed remarkable similarities to the reference genome.

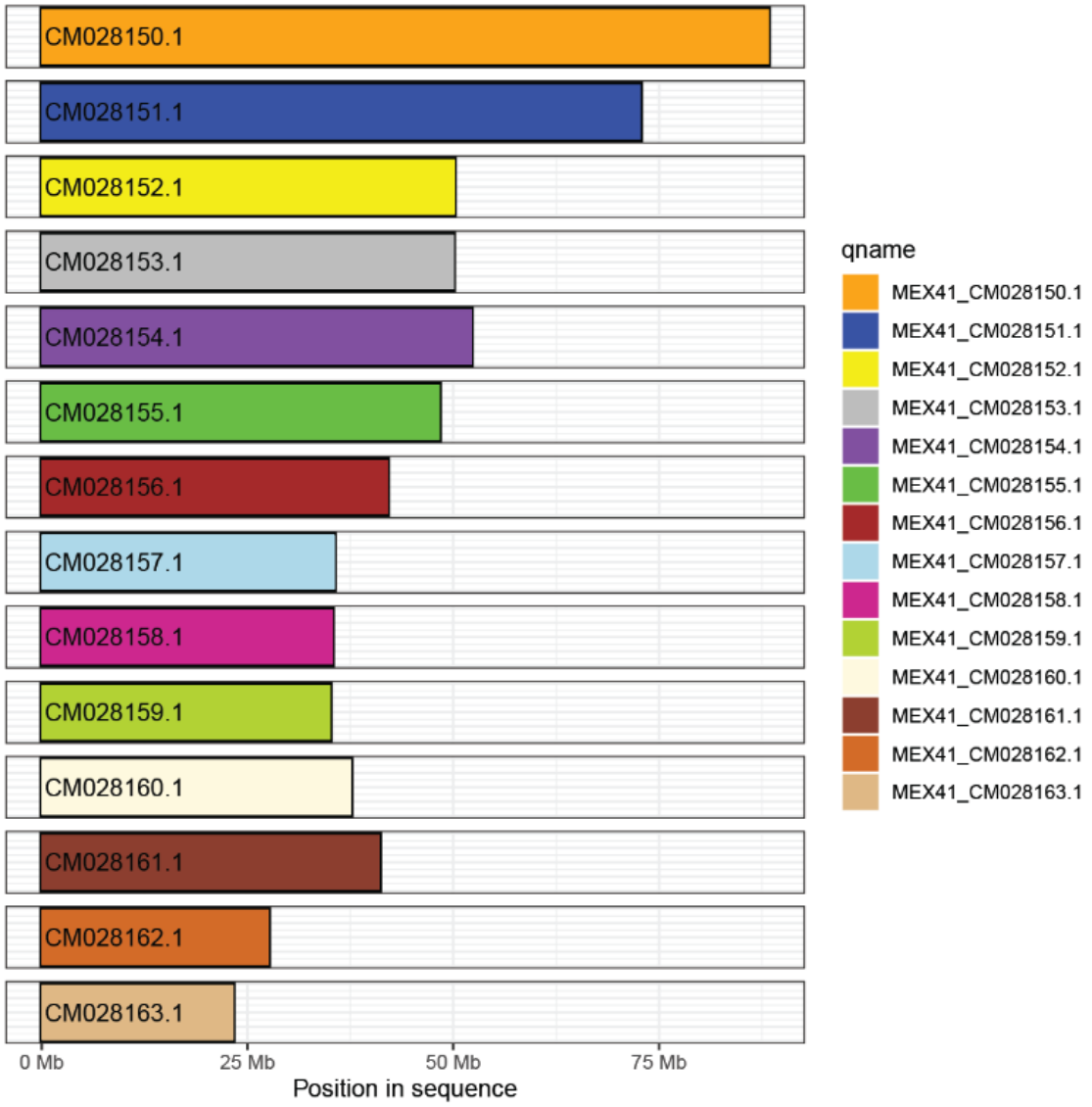


Figure 4.5 **Genome coverage results of MEX41 on the reference genome. Fully colored chromosomes show complete coverage.**

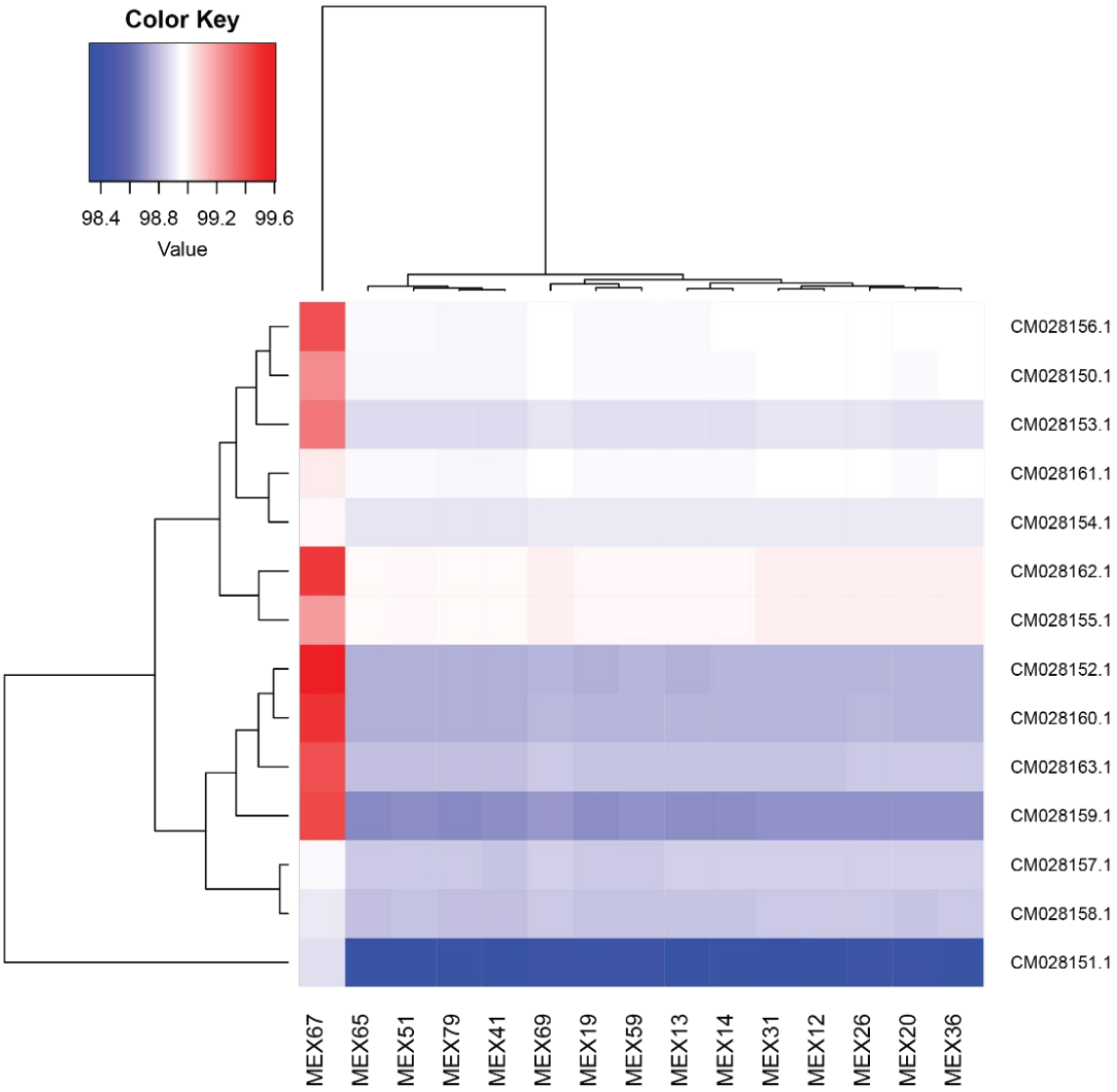


Figure 4.6 Heat map of chromosomal similarities between Mexican *V. planifolia* samples and reference 'Daphna' genome.

SNP calling and clustering analyses

A total of 7,468,839 high-quality biallelic single nucleotide polymorphisms (SNPs) were detected among all samples. Pruning for linkage disequilibrium (LD) at thresholds 0.2 and 0.8 reduced the number of filtered SNPs to 9,297 and 419,885, respectively. Independently of LD thresholds, results from the principal components analysis (PCA) remained similar (Figures 4.7A, S4A). With a LD threshold set to 0.2, the

top two eigen vectors explained 10.39 and 8.16% of variance (Figure 4.7A). Within this more conservative PCA (Figure 4.7A), most samples clustered together exhibiting slightly more variability in eigen vector 2 values than in eigen vector 1 values. Two samples, MEX31 and MEX67, did not cluster with the rest, nor each other. MEX31, exhibited distinctively low values along eigen vector 2 and MEX67 exhibited distinctively high values along eigen vector 1. Within the PCA set with a LD threshold of 0.8, MEX31 clustered with the other samples along eigen vector 2, but MEX67 remained distantly separated. At both thresholds, SNPs were scattered across all chromosomes, but were most numerous on chromosome 2 (Table 4.3, Figure S4B). When SNP density was mapped onto the 14 chromosomes (using a 500 kb sliding window), non-randomly distributed SNP hotspots were revealed (Figure 4.7B). The most prominent hotspot occurred along a terminal region of chromosome 2; other notable hotspots occurred on terminal regions of chromosomes 9 and 14 (Figure 4.7B).

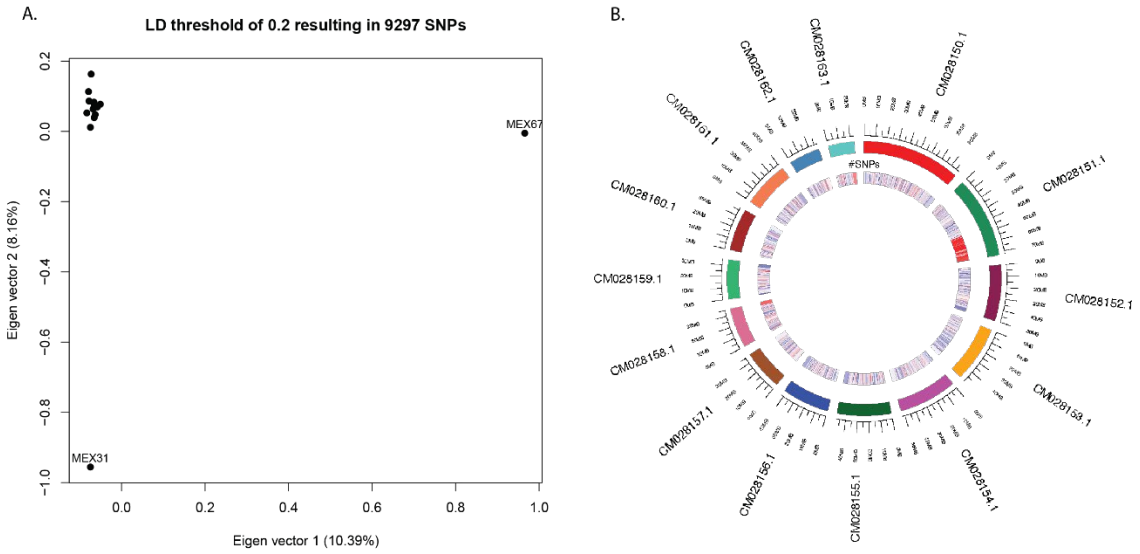


Figure 4.7 SNP analyses using linkage disequilibrium thresholds set to 0.2 resulting in 9,297 SNP markers: A) Principal component analysis (PCA) and B) SNP density along 500 Kb sliding window on 14 chromosomes (colored). Regions of SNP density are illustrated on a color gradient from blue (low) to red (high).

Table 4.3 SNP quantity by chromosome based on filtering with linkage disequilibrium (LD) thresholds set to 0.2 and 0.8

Chromosome Number	Chromosome ID	SNP Quantity (LD threshold = 0.2)	SNP Quantity (LD threshold = 0.8)
1	CM028150.1	1248	45148
2	CM028151.1	1315	129408
3	CM028152.1	710	27900
4	CM028153.1	680	23651
5	CM028154.1	723	24225
6	CM028155.1	678	23166
7	CM028156.1	570	19088
8	CM028157.1	525	19218
9	CM028158.1	524	17213
10	CM028159.1	487	23185
11	CM028160.1	508	21826
12	CM028161.1	564	17157
13	CM028162.1	418	14437
14	CM028163.1	347	14263
	Total	9297	419885

Discussion

High genome-wide heterozygosity stemming from hybridization

Compared to other plants (Figure 4.1A), the reference 'Daphna' genome and all Mexican *V. planifolia* samples exhibited high genome-wide heterozygosity levels, except one, MEX67, which exhibited a notably low level (0.403%), (Figure 4.3). Similar levels have been reported in diploid *Artemisia tridentata* Nutt. (Asteraceae) and, based on the abundance of AAB (26%) and AABB (14%) k-mer ratios along with the dominant AB (49%) k-mer pairs (Melton et al., 2022), were attributed to past polyploidization followed by diploidization events within the evolution of the species. For *V. planifolia*, however, results from Smudgeplot confirmed the diploid status of all accessions by uncovering almost exclusive AB k-mer pairs (average 97.3%; Figure 4.4), indicating the prominence of two sub-genomes, and therefore ruling out these latter evolutionary events as sources

of the observed high heterozygosity. Together, these findings support the hypothesis of a past hybridization event within cultivated *V. planifolia*. Concordantly, a hybrid origin has been previously proposed based on the evidence of chromosomal structure (Nair & Ravindran, 1994)and contrasting phylogenetic signals between chloroplast and nuclear DNA sequences (Ellestad et al., 2021). Previous hypotheses have attributed these levels to the accumulation of somatic point mutations (Favre et al., 2022), however, the extent of this heterozygosity as compared to other plants (Figure 4.1A) suggests the occurrence of a more extreme evolutionary event, like hybridization between *V. planifolia* and a congener or two genetically differentiated *V. planifolia*. Similar levels of heterozygosity (2.27%) were observed in *Litchi chinensis* Sonn. and were attributed to the hybridization of two distinct haplotypes(G. Hu et al., 2022). In addition to distinct haplotypes of *V. planifolia,* candidate parental species may include other less cultivated species like *V. pompona* or *V. odorata*. One sample, MEX67, did not exhibit the genomic signal of hybridization as the others did (Figure 4.3B), and therefore may represent the most in-bred form of *V. planifolia.* Further research is needed to understand the implications that these genomic structures have on fitness or other desirable traits.

Comparative chromosomal analyses suggest multiple domestication events in Mexico

Varying chromosomal structure (Figure 4.6), as compared to the reference 'Daphna' genome, and clustering patterns based on SNP relatedness (Figure 4.7A) suggest that multiple evolutionary pathways have shaped the genomes of cultivated Mexican *V. planifolia* leading to their similar, but variable, levels of heterozygosity. These results indicate that multiple haplotypes exist within the AB sub-genomes identified through Smudgeplot analyses (Figure 4.4). The accessions within this study are

most likely not clones, but the result of several domestication events in Mexico. One sample, MEX67, exhibited notable differences to all other samples as shown by substantially lower genomic heterozygosity levels (Figure 4.3), high degree of similarity to the reference chromosome (Figures 4.5 and 4.6), and distant positioning in the PCA (Figure 4.7A). Largely congruent chromosomal structuring indicates similarity to the reference genome, but the conflicting heterozygosity levels contradict this similarity. Therefore, it is probable that MEX67 matches to the one haplotype that is referenced in the *V. planifolia* 'Daphna' genome, but not the other haplotype, which is not referenced. This sample, which was cultivated from a wild source in the Chinantla region of Oaxaca, may represent the true *V. planifolia*, from a natural non-hybrid origin.

Not considering MEX67, the two groups of samples on the chromosomal similarity heat map (Figure 4.6) and the clustering of most samples, except MEX31, in the PCA (Figure 4.7A) support the hypothesis that multiple evolutionary or domestication processes have affected vanilla cultivated in this region. Grouping of samples within the heat map did not reflect geography. The distribution of these groups throughout the entire sampling region and the additional, but less extreme, chromosomal variation within groups shows that these groups have been dispersed by humans throughout the entire sampling region and that additional domestication processes like introgression and/or the accumulation of somatic point mutations may have contributed to their genomic makeup. Although grouping within the heat map is not completely mirrored by the PCA, these results show that at least two main domestication events of hybridization have occurred within vanilla cultivated in Mexico. Considering the long histories of cultivation by groups such as the Aztecs, Mayans, and Totonacs, the findings prove reasonable in that

ancient cultures might have separately influenced the genomic make-up of regionally cultivated species through the passing down of cultivation knowledge and plant material.

Conservation of Mexican vanilla landraces and implications for production

Diverse landraces within a crop's native distribution provide an important source of genetic diversity to potentially increase its capacity to cope with environmental change (Bellon et al., 2015). The genomic signals of multiple origins of cultivated vanilla within Mexico support the hypothesis of landrace cultivation, which was previously suggested based on ITS haplotype analyses (Ellestad, Pérez-Farrera, Forest, et al., 2022). Additionally, results from this study suggest that most cultivated vanilla comes from a hybrid origin between either two genetically differentiated *V. planifolia,* or between *V. planifolia* and another species. Other species found in cultivation such as, *V. x tahitensis* (Lubinsky, Cameron, et al., 2008)*, V. pompona,* and *V. insignis* (Ellestad, Pérez-Farrera, Forest, et al., 2022)*,* may offer parental candidates for cultivated vanilla. Given that only *V. planifolia* and *V. x tahitensis* are recognized for commercial vanilla production (CBI, 2018; U.S. FDA, 2022), the reconsideration of species' requirements to include congeneric species may offer novel alternative sources for vanilla production and catalyze more inclusive conservation strategies for *Vanilla*. The prioritization of agricultural diversity and the conservation of landraces within this biologically, culturally, and economically important region, will not only benefit global vanilla production and sustainability, but will also benefit the livelihoods of farmers and may help to encourage the protection of cultural diversity in Mexico.

Materials and Methods

A more comprehensive, reproducible workflow (including code, citations and package version) of methods within this study are available on GitHub (GitHub, n.d.). Additionally, a companion GitHub Pages website (*Vanilla Genomic Project*, n.d.) was developed to fully explain our analyses.

Review of plant levels of genomic heterozygosity inferred using GenomeScope

To obtain a reference of plant genomic heterozygosity levels inferred using GenomeScope (Ranallo-Benavidez et al., 2020; Vurture et al., 2017), a literature review was conducted using the R package 'easyPubMed' (Fantini, 2019) and 'rentrez' (D. J. Winter, 2017) querying all studies that have used this software (using the two PubMed accession numbers associated to publications related to GenomeScope) since March 29th, 2022 and are deposited on PubMed. From each study, the following attributes were manually recorded by inspecting publications: species, ploidy levels, genomic heterozygosity, and estimated genome size. Additionally, a list of possible human uses for each species was obtained using categories provided in the World Checklist of Useful Plant Species, compiled by the Royal Botanic Gardens, Kew (UK)(Diazgranados et al., 2020).

Sampling, DNA extraction, and whole genome resequencing

Samples were collected from 15 *Vanilla planifolia* plants within the origin of vanilla cultivation in Mexico in October of 2019 from the northernmost region around Papantla, Veracruz to the southernmost region around Chinantla, Oaxaca (Figure 2). Samples included the breadth of genetic, phenotypic, and climatic variation as inferred from ITS haplotype analyses in Ellestad, Pérez-Farrera, Forest, et al. (2022). Voucher

specimens for these individuals were deposited at Herbario Eizi Matuda (HEM) and are represented as a living collection maintained in Berriozábal, Chiapas, Mexico. From each individual, vegetative cuttings were taken, and one gram of leaf material was dried in silica gel for genomic analyses. Additionally, the publicly available phased *V. planifolia* 'Daphna' genome (BioProject ID: PRJNA633886), downloaded from the National Center for Biotechnology (NCBI) website, was used as a reference in this study.

Genomic DNA was extracted from all lyophilized leaf samples using the Qiagen DNeasy Plant Mini Kit (Hilden, Germany) following the manufacturer protocol. DNA yield was quantified using a Qubit fluorometer (Thermo Fisher Scientific, Inchinnan, UK). Extracted genomic DNA with concentrations greater than 20 ng/µL was sent to GENEWIZ, inc.. (South Plainfield, New Jersey, USA) for library preparation and sequencing of 150bp paired-end reads using an Illumina HiSeq platform aiming for a sequencing depth of 50x to allow for sufficient coverage on the reference 'Daphna' genome (736,752,966 bp) (Hasing et al., 2020). Raw sequences were checked for quality using FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and all reads were cleaned and trimmed using Trimmomatic (Bolger et al., 2014) with minimum length (MINLEN) reads set to 100bp and a Phred score of 33.

Genomic heterozygosity, ploidy, and complexity

Genomic sequences were characterized (size, heterozygosity, repetitiveness) by k-mer frequency analyses (k=21) using Jellyfish (Marçais & Kingsford, 2011) and GenomeScope 2.0 (Ranallo-Benavidez et al., 2020). Using k-mer (k=21) histograms obtained by KMC3 (Kokot et al., 2017), heterozygous k-mer pairs were analyzed through Smudgeplot (Ranallo-Benavidez et al., 2020) to estimate ploidy levels and infer genomic

complexity. Lower (L) and upper (U) end cut-off values, below and above which all k-mers were discarded as errors, were set using k-mer coverage output from Genomescope 2.0, as recommended in the Smudgeplot documentation (https://github.com/KamilSJaron/smudgeplot) using k-mer coverage (kcov) values inferred from GenomeScope, where L = (kcov/2)-5.

Genome reconstructions to infer structural variation and synteny

Chromosome-scale genomes were reconstructed by mapping cleaned, trimmed reads to the reference *V. planifolia* 'Daphna' genome using Bowtie 2 (Langmead & Salzberg, 2012). Variants were called, filtered, and normalized and consensus genome sequences were created using SAMtools and BCFtools (Danecek et al., 2021). Chromosome-level genome sequences were compared against the reference genome using Minimap2 (H. Li, 2018) to assess similarity. In R (R Core Team, 2017), chromosomal coverage was evaluated using the 'pafr' package (D. Winter, 2020) and chromosomal rearrangements and synteny were assessed using both 'pafr' and 'dotPlotly' packages (https://github.com/tpoorten/dotPlotly). For visualization of genomic variability among samples, a heat map was produced in R using 'gplots' (Warnes et al., 2022) to show the percentage of identities between each sample and the reference genome at a chromosome level.

SNP calling and clustering analysis

Reconstructed genomes were analyzed using BCFtools (Danecek et al., 2021)to call and filter variants with Phred quality scores greater than 20. Using the R package 'SNPRelate V1.6.4' (Zheng et al., 2012), indexed calls were further filtered to include only biallelic SNPs in linkage equilibrium with each other. Since the population

processes affecting this species as a cultivated plant is unclear, linkage disequilibrium thresholds were set to the wide-ranging values 0.2 and 0.8. Also using ‘SNPRelate’ (Zheng et al., 2012), principal components analyses were conducted with both linkage disequilibrium thresholds to observe and minimize the effect of SNP clusters. Results were plotted using the top two eigenvectors explaining the largest percent of variance among the data. Additionally using both linkage disequilibrium thresholds, SNP density along a 500 Kb sliding window was mapped onto the 14 chromosomes. To observe the chromosome level distribution of SNP hotspots, results were plotted using the R packages ‘seqinr’ (Charif & Lobry, 2007) and ‘RCircos’ (Zhang et al., 2013).

CHAPTER FIVE: ON THE GENOMIC ORIGIN OF CULTIVATED VANILLA

Abstract

Although the vanilla spice is so well-known, there is an unexpected lack of knowledge on the domestication processes that have shaped its genetic resources. Within the main vanilla producing plant species, *Vanilla planifolia,* which has a long history of cultivation primarily through vegetative cuttings, unexpected high levels of genome-wide heterozygosity have been found in the global "Daphna" cultivar and in accessions cultivated in Mexico, its center of origin. These findings have pointed to a hybrid origin within the crop; however, the source and quantity of hybridization events remained unclear. This study aimed to disentangle these events by identifying parental candidate species and investigating the domestication processes underlying its hybrid origin. The less commonly cultivated species, *V. pompona,* was identified as a likely parental candidate and hypothesized to have hybridized with *V. planifolia* causing the high levels of genome-wide heterozygosity observed within cultivated individuals. Chromosomal structure and SNP distributions were compared between a *V. planifolia* "Daphna" cultivar*,* 15 *V. planifolia* accessions from Mexico, and one *V. pompona* accession to answer the questions: 1) do both haplotypes of the "Daphna" genome show signatures of hybridization with *V. pompona*?, 2) do the highly heterozygous Mexican *V. planifolia* accessions exhibit the same hybridization signatures as the "Daphna" genome?, and 3) do the latter accessions show signatures of one or multiple domestication events? Results provided evidence for multiple genomic origins within cultivated *V. planifolia* and

revealed more variation than previously recognized. At least two distinct hybridization events were shown to have occurred: one showing signatures of introgressive hybridization between *V. planifolia* and *V. pompona* in the “Daphna” cultivar, and at least one more in Mexican cultivars. This parental origin, however, has yet to be identified. Additional research incorporating genomic data from other crop-wild relatives is necessary to further disentangle these events observed in Mexico. Findings from this study offer a clearer illustration of vanilla’s evolutionary history and genetic resources, highlight their importance for crop sustainability, and provide foundations for future research into the origin of this important spice.

Introduction

Understanding the genetic identities that make up the gene pool of crop species is essential for ensuring their sustainability in the face of climate and landscape changes (Ebert & Engels, 2020). The genetic resources within a crop's center of origin are especially important due to the fact that they often encompass the breadth of genetic diversity resulting from natural evolutionary and/or domestication processes (Engels et al., 2006). This fundamental knowledge, however, is surprisingly limited for many crop species (Gepts, 2006). For example, the genetic resources are even unclear for *Vanilla planifolia* Andrews (Orchidaceae), the plant that produces one of the most well-known and valuable spices, vanilla. For a crop that is already at risk due to limited genetic diversity from centuries of clonal propagation through vegetative cuttings (Bory, Grisoni, et al., 2008), this gap in knowledge hinders the effective preservation of the resources needed for future crop improvements. Multiple studies have shown that the cultivation of the main vanilla producing species, *Vanilla planifolia*, originated in Mexico (Lubinsky, Bory, et al., 2008) with the crop experiencing multiple domestication events and hybridization (Ellestad, Pérez-Farrera, & Buerki, 2022), however the factors contributing to its genomic make up remain mostly unclear. Co-occurring crop-wild relatives exhibiting traits that may increase crop resilience, e.g. *Vanilla pompona* Schiede, are likely to have contributed to this hybrid origin. In this study, we aimed to augment knowledge on the history of *V. planifolia* cultivation by assessing the genomic origin of vegetatively propagated cultivars through the comparison of genomic structure and variation between the predominantly cultivated *V. planifolia* and a relative with which it is hypothesized to have hybridized, *V. pompona,*

Within cultivated vanilla, varying degrees of genomic variability (Chambers et al., 2021; Favre et al., 2022; Hasing et al., 2020; Y. Hu et al., 2019) and high levels of genome-wide heterozygosity (GWH) have provided potential evidence for multiple domestication events and past hybridization (Ellestad, Pérez-Farrera, & Buerki, 2022). Genomic data from 15 cultivated accessions from Mexico, encompassing the breadth of phenotypic and haplotypic diversity – identified by (Ellestad, Pérez-Farrera, Forest, et al., 2022) using ITS sequences – showed two distinct groups based on levels of GWH. Relative to the average level of GWH (1.10%) found within diploid plants (Ellestad, Pérez-Farrera, Forest, et al., 2022), accessions were split into a low heterozygosity group and a high heterozygosity group. The first group contained only one sample, MEX67, which exhibited a low heterozygosity level of 0.403%, and the second group contained all other 14 samples, which exhibited levels ranging from 2.52% to 2.85%. All five samples of the most commonly and widely distributed haplotype, Hpl3 (Table 5.1), belonged to this highly heterozygous group along with nine other distinct haplotypes (Ellestad, Pérez-Farrera, Forest, et al., 2022). Additionally, a similarly high level (2.48%) was also observed within the publicly available genome of the vanilla cultivar “Daphna” (Hasing et al., 2020). Although we have limited data on the occurrence of this cultivar, it was assumed that this was a common global cultivar because it was chosen as a reference for genome assembly. The low heterozygosity group, MEX67, which did not show signatures of hybridization, was hypothesized to represent an ancestral *V. planifolia* genome unaffected by the hybridization event(s) that produced the highly heterozygous accessions.

To test if the high levels of heterozygosity resulted from a shared domestication event, Ellestad, Pérez-Farrera, & Buerki (2022) mapped genomic sequences from the Mexican cultivated *V. planifolia* accessions onto one haplotype of the reference "Daphna" genome to compare differences. Non-random regions of high single nucleotide polymorphism (SNP) densities along chromosomes revealed that these genomes may have originated independently (Ellestad, Pérez-Farrera, & Buerki, 2022). The resulting density map of consensus SNPs among all accessions, though, lacked the ability to test whether each individual shared the same pattern and, due to the availability of only one haplotype, the genomic processes underlying this pattern were unable to be fully disentangled (Ellestad, Pérez-Farrera, & Buerki, 2022). More recently, however, the fully phased genome has become available (Hasing et al., 2020). Using both haplotypes of this phased diploid "Daphna" genome as a reference, we aimed to identify the parental origin of these highly heterozygous vanilla accessions and clarify the distinct events and mechanisms underpinning vanilla's genomic structure.

To identify a parental origin which hybridized with *V. planifolia,* a crop-wild relative was selected for genomic comparison. One likely parental candidate was *V. pompona,* a sympatric species with morphological characteristics and climatic requirements similar to those of *V. planifolia* (Flores Jiménez et al., 2017; Soto Arenas & Cribb, 2010). *Vanilla pompona* was hypothesized to be the most likely parental candidate because it has been shown to hybridize with *V. planifolia* (Delassus, 1963; Y. Hu et al., 2019) and has been found to be in cultivation in Mexico (Ellestad, Pérez-Farrera, Forest, et al., 2022). On top of that, *V. pompona* has been shown to exhibit a high resistance to stem and root rot, therefore making it desirable in cultivation (Koyyappurath et al., 2015).

Therefore, we hypothesized that a hybridization event between *V. planifolia* and *V. pompona* was responsible for the observed high levels of genomic heterozygosity observed within Mexican accessions and/or the reference "Daphna" genome.

Using genomic data from the 15 cultivated Mexican *V. planifolia* accessions from Ellestad, Pérez-Farrera, & Buerki (2022), *V. pompona,* and both haplotypes of the phased "Daphna" genome, we aimed to answer the following questions: 1) do both haplotypes of the Daphna genome show signatures of hybridization with *V. pompona*?, 2) do the highly heterozygous Mexican *V. planifolia* accessions exhibit the same hybridization signatures as the "Daphna" genome?, and 3) do the latter accessions show signatures of one or multiple domestication events? For the first questions, we hypothesized that the observed signature of hybridization was the result of one of three processes distinctly impacting one or both homologous chromosomes in the phased diploid reference genome: 1) an F1 hybridization event in which the maternal *V. planifolia* contributed to one haplotype and the paternal *V. pompona* contributed to the other haplotype, 2) a hybridization event followed by backcrossing with *V. pompona*, where signatures of *V. pompona* have been conserved on each parental chromosome, and 3) a hybridization event between *V. planifolia* and another unidentified genome. For the second question, we hypothesized that two or more of these hybridization processes (i.e., domestication events) have affected these highly heterozygous accessions causing the observed genomic dissimilarities between the "Daphna" genome, the highly heterozygous Mexican accessions, and the minimally heterozygous accession. For the third question, we hypothesized that multiple domestication events within the Mexican accessions contributed to their genomic variability.

Due to the predominance of vegetative cuttings in the cultivation of vanilla, different domestication events can be inferred through genome comparison. If a hybridization event were to have occurred between *V. planifolia* and *V. pompona* to produce the "Daphna" cultivar (Q1), we would expect the minimally heterozygous *V. planifolia* (MEX67) and *V. pompona* genomes to exhibit contrasting patterns of genomic similarity along chromosomes of one or both of the haplotypes of the reference genome. If the highly heterozygous Mexican accessions were to have resulted from a different hybridization process than the reference (Q2), we would expect them to exhibit corresponding genomic regions of high similarity as well as regions with low similarity. Finally, if multiple domestication events were to have occurred among these highly heterozygous Mexican accessions (Q3), we would expect to see contrasting patterns of genomic similarity.

To test these hypotheses, we first assessed genomic compatibility of *V. pompona* through estimations of ploidy and GWH levels, then examined chromosome-level genomic similarity between one *V. pompona* and 15 Mexican *V. planifolia* accessions by mapping cleaned genomic reads to each reference haplotype (A and B) of the phased "Daphna" genome and calculating genetic distances. We then identified SNPs for each accession on each haplotype of the reference genome and compared their distribution and density along parental homologous chromosomes. Additionally, we analyzed the possibility of multiple domestication events by assessing relationships between accessions based on SNPs in a principal component analysis (PCA).

Materials and Methods

Genomic data

Publicly available genomic data (SRR19374404 – SRR19374417) were used in this study for 15 cultivated *V. planifolia* plants collected in Mexico from the northernmost region around Papantla, Veracruz to the southernmost region around Chinantla, Oaxaca (Ellestad, Pérez-Farrera, & Buerki, 2022). Samples included the breadth of genetic, phenotypic, and climatic variation as inferred from ITS haplotype analyses in Ellestad, Pérez-Farrera, Forest, et al. (2022). Additionally, genomic data for *V. pompona* 'King' (BioProject ID: PRJNA633886, BioSample ID: SAMN16056350, SRA: SRR12628844) and both haplotypes of the phased *V. planifolia* 'Daphna' genome (BioProject IDs: PRJNA633886 and PRJNA668740, BioSample: SAMN14973820, Assemblies: GCA_016413885.1 and GCA_016413895.1) were downloaded from the National Center for Biotechnology (NCBI) website. Both "Daphna" haplotypes, referred to as A and B, were used as a reference in this study. All genomic reads were cleaned and trimmed using the procedures described in Ellestad, Pérez-Farrera, & Buerki (2022).

Ploidy and GWH of *V. pompona*

The parental candidate, *V. pompona,* was assessed for genomic compatibility with *V. planifolia.* Following the same methods as Ellestad, Pérez-Farrera, & Buerki (2022), genomic sequences were analyzed for ploidy level and genomic complexity using KMC3 (Kokot et al., 2017) and Smudgeplot (Ranallo-Benavidez et al., 2020) then characterized by heterozygosity using Jellyfish (Marçais & Kingsford, 2011) and GenomeScope 2.0 (Ranallo-Benavidez et al., 2020).

Reconstructing genome sequences for chromosome-level comparison by phased haplotype

For all *V. planifolia* and *V. pompona* accessions, chromosome-level genomic similarity was assessed, based on genetic distances and SNP distribution following the methods of Ellestad, Pérez-Farrera, & Buerki (2022), and compared along each haplotype, A and B, of the reference *V. planifolia* 'Daphna' genome. Genomes were reconstructed by mapping cleaned, trimmed reads to each haplotype using Bowtie 2 (Langmead & Salzberg, 2012). Variants were then called, filtered, and normalized, and consensus genome sequences were created using SAMtools and BCFtools (Danecek et al., 2021). Chromosome-level genome sequences were compared against each haplotype genome using Minimap2 (H. Li, 2018) to assess similarity. In R (R Core Team, 2017), chromosomal coverage was evaluated using the 'pafr' package (D. Winter, 2020) to identify potential chromosomal rearrangements. Genomic variability was then visualized through a heat map, produced using 'gplots' (Warnes et al., 2022), to show the percentage of identities between each sample and each haplotype genome by chromosome.

SNP distribution and comparative clustering

Mapped reads from each accession were individually analyzed using BCFtools (Danecek et al., 2021) to call and filter variants following the methods of (Ellestad, Pérez-Farrera, & Buerki, 2022). Using the R package 'SNPRelate V1.6.4' (Zheng et al., 2012), individually indexed calls were further filtered to include only biallelic SNPs, then SNP density was mapped along a 500 Kb sliding window onto each chromosome of each reference haplotype. To observe the chromosome level distribution of SNP densities and

compare haplotype SNP distributions, results were plotted using the R packages 'seqinr' (Charif & Lobry, 2007) and 'RCircos' (Zhang et al., 2013). Also using 'SNPRelate', SNPs between all highly heterozygous *V. planifolia* accessions were called with a linkage disequilibrium threshold of 0.2 and principal components analyses were conducted on each haplotype to observe the effect of individual clustering among genetically similar samples. Results were plotted using the top two eigenvectors explaining the largest percent of variance among the data.

Results

Ploidy and GWH of *V. pompona*

Results from the Smudgeplot analysis revealed that although *V. pompona* is diploid and is mostly made up of an AB genome (67%), it also exhibits remnants of an AABB genome (33%; Figure 1B). Additionally, *V. pompona* was found to have a relatively low level of GWH (0.734%; Table 5.1; Figure 5.1A).

Table 5.1 Identification and attributes of genomic data.

Sample ID	SRA	Ploidy	ITS Haplotype	GWH (ab%)
MEX12	SRR19374411	2x	Hpl5	2.74
MEX13	SRR19374410	2x	Hpl4	2.67
MEX14	SRR19374409	2x	Hpl8	2.66
MEX19	SRR19374408	2x	Hpl14	2.52
MEX20	SRR19374407	2x	Hpl1	2.56
MEX26	SRR19374406	2x	Hpl7	2.57
MEX31	SRR19374405	2x	Hpl3	2.61
MEX36	SRR19374404	2x	Hpl21	2.52
MEX41	SRR19374417	2x	Hpl3	2.77
MEX51	SRR19374416	2x	Hpl22	2.62
MEX59	SRR19374415	2x	Hpl3	2.79
MEX65	SRR19374414	2x	Hpl3	2.57
MEX67	SRR19374413	2x	Hpl16	0.403
MEX69	SRR19374412	2x	Hpl3	2.85
MEX79	SRR19374418	2x	Hpl3	2.53
Daphna	SRR12628848	2x	-	2.48

Pompona	SRR12628844	2x	-	0.734

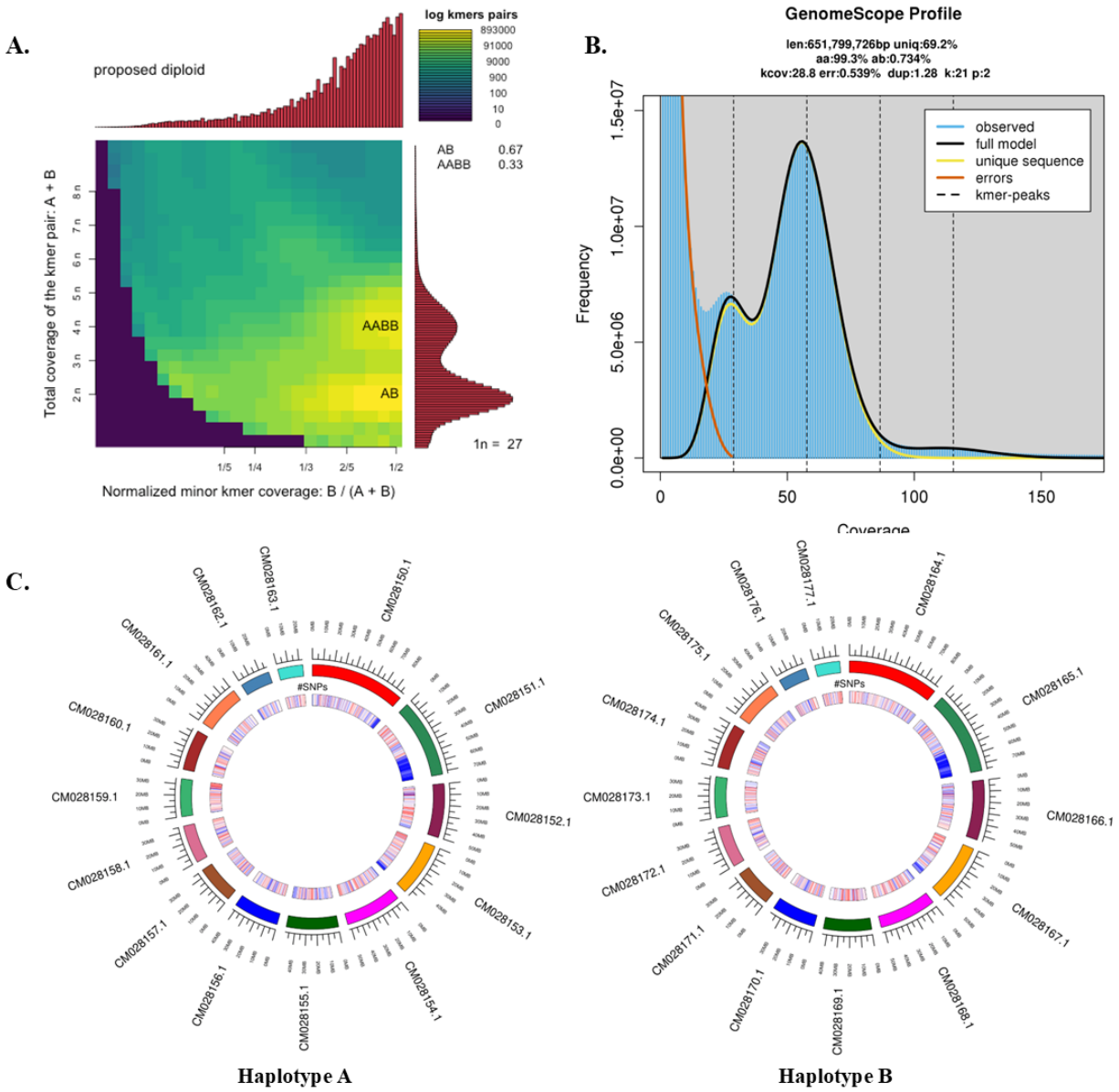


Figure 5.1 Genome structure analyses of *V. pompona:* A) k-mer frequency distribution (k=21) assessing the relative abundance of homozygous and heterozygous sequences, B) Smudgeplot output indicating ploidy and the frequency of each haplotype structure within the genome, and C) SNP density from *V. pompona* along 500 Kb sliding windows on 14 homologous chromosomes (colored) of both reference haplotypes. Regions of SNP density are illustrated on a color gradient from blue (low) to red (high).

Reconstructing genome sequences for chromosome-level comparison by phased haplotype

Genomic alignments from MiniMap2 revealed that all reconstructed genomes exhibited full coverage on both haplotypes of the 'Daphna' reference genome, therefore suggesting no chromosomal rearrangements had occurred. For all accessions, results showed similar but slightly varying patterns of genomic similarity along homologous chromosomes of both haplotypes of the reference genome (Figure 5.2A). This variation was particularly evident along the sixth (CM028155.1 and CM028169.1) and tenth (CM028159.1 and CM028173.1) homologous chromosomes. Along both haplotypes, the *V. pompona* accession exhibited low levels of genomic similarity on all chromosomes, but higher on the second chromosomes (CM028151.1 and CM028165.1). In contrast, the minimally (MEX67) and the highly heterozygous *V. planifolia* accessions exhibited high levels of genomic similarity on all homologous chromosomes except the second chromosomes.

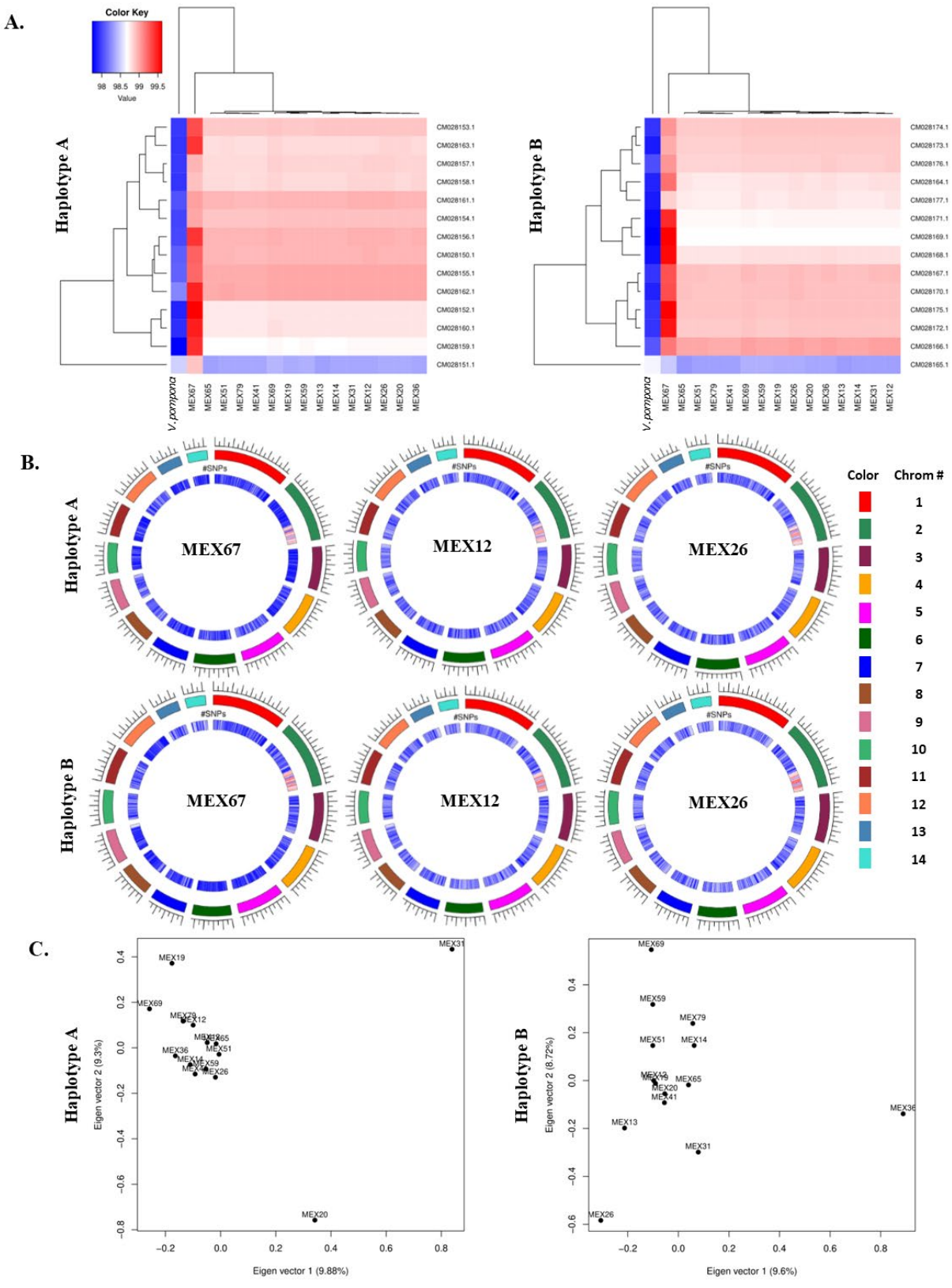

A.
Color Key
98 98.5 99 99.5
Value
Haplotype A
Haplotype B
V. pompona
MEX67
MEX65
MEX51
MEX79
MEX41
MEX69
MEX19
MEX59
MEX13
MEX14
MEX31
MEX12
MEX26
MEX20
MEX36
B.
Haplotype A
Haplotype B
#SNPs
MEX67
MEX12
MEX26
Color
Chrom #
1
2
3
4
5
6
7
8
9
10
11
12
13
14
C.
Haplotype A
Eigen vector 2 (9.3%)
Eigen vector 1 (9.88%)
MEX19
MEX69
MEX12
MEX65
MEX36
MEX51
MEX14
MEX59
MEX26
MEX31
MEX20
Haplotype B
Eigen vector 2 (8.72%)
Eigen vector 1 (9.6%)
MEX69
MEX59
MEX79
MEX51
MEX14
MEX65
MEX20
MEX41
MEX13
MEX31
MEX36
MEX26

Figure 5.2 A) Heat map of chromosomal similarities between Mexican *V. planifolia* accessions, *V. pompona* accession and both haplotypes of the reference 'Daphna' genome. B) SNP density from Mexican *V. planifolia* accessions along 500 Kb sliding windows on 14 homologous chromosomes (colored) of both reference haplotypes. Regions of SNP density are illustrated on a color gradient from blue (low) to red (high). C) Principal components analysis of SNPs, using a linkage disequilibrium threshold of 0.2, from highly heterozygous Mexican *V. planifolia* accessions along each reference haplotype.

SNP distribution and comparative clustering

For *V. pompona,* SNP distributions along chromosomes revealed large regions of high density throughout most of the genome (represented by red in Figure 5.1C), however, were punctuated with non-random regions of very low density (represented by blue in Figure 5.1C). Regions of low SNP density were most notable on the terminal regions of homologous chromosomes 2 and 4 (Figure 5.1C). Along other chromosomes, most notably the ends of first homologous chromosome set (CMO28150.1 and CM028164.1), a higher SNP density was found on haplotype B than haplotype A (Figure 5.1C) for *V. pompona*. For the Mexican accessions, SNP distributions revealed similar patterns among each other, but contrasted sharply with those of *V. pompona* (Figures 1C and 2B). SNP distributions revealed large regions of low SNP density throughout most of the genome (represented by blue in Figure 5.2B) but were punctuated by non-random regions of high SNP density (represented by red in Figure 5.2B), notable along the second homologous chromosomes. Along both reference haplotypes, SNP density patterns were analogous for the Mexican accessions (Figure 5.2B).

Among the 14 highly heterozygous Mexican accessions, SNP calling resulted in 33,357 and 32,273 SNPs on the reference haplotypes A and B, respectively. After filtering for biallelic SNPs with a linkage disequilibrium threshold set to 0.2; 1,171 and

1,180 SNPs remained for each respective haplotype. Further exploration of relationships within these highly heterozygous Mexican accessions through a PCA of filtered SNPs on the reference haplotype A, where eigenvectors one and two were found to explain 9.88% and 9.3% of the variance showed that most samples grouped together but that MEX31 and MEX20 were distantly positioned (Figure 5.2C). On the reference haplotype B, where eigenvectors one and two were found to explain 9.6% and 8.72% of the variance, samples varied uniformly along the eigen vector 2, however, MEX36 was positioned distantly from the rest along eigen vector 1 (Figure 5.2C).

Discussion

Hybridization between *V. planifolia* and *V. pompona* within the "Daphna" genome

The reference "Daphna" genome reveals unique signatures of hybridization between *V. planifolia* and *V. pompona* on each haplotype (Figures 5.1 and 5.2). This signature is most clear on the end of the second chromosome where the *V. pompona* accession exhibits unusually low SNP densities (Figure 5.1C) and the *V. planifolia* accessions exhibit unusually high SNP densities (Figure 5.2B). Incongruent values of genomic similarity (Figure 5.2A) and SNP density (most evident on the first chromosome; Figure 5.2B) along homologous chromosomes, indicates that distinct hybridization events occurred within each haplotype, therefore suggesting introgression within this "Daphna" cultivar. These results reflect our second hypothesized process where a region on the end of the second chromosome was integrated into the *V. planifolia* genome from *V. pompona* through mechanisms such as crossing-over and genetic recombination, then repeated introgression. Furthermore, results indicated a more

complicated evolutionary past within *V. pompona* (Figure 5.1A and B) than has been observed within *V. planifolia* (Ellestad, Pérez-Farrera, & Buerki, 2022). The large proportion of an AABB subgenome within *V. pompona* (Figure 5.1B) was most likely caused by more ancient events, such as polyploidization followed by diploidization. To a lesser degree, nine of the highly heterozygous Mexican accessions also exhibited remnants of this AABB genome, ranging from 3%-5%, and the "Daphna" reference genome exhibited 7% (Ellestad, Pérez-Farrera, & Buerki, 2022). Polyploidization has been previously shown to naturally occur within *Vanilla* (Bory, Catrice, et al., 2008) and similar signatures of polyploidization followed by diploidization have been shown within *Artemesia tridentata* Nutt. (Melton et al., 2022).

Distinct hybridization events between "Daphna" and highly heterozygous Mexican accessions

For the highly heterozygous Mexican accessions, SNP distributions along homologous chromosomes of each haplotype mirrored those of MEX67, therefore indicating that *V. pompona* was not involved in this particular hybridization event (Figure 5.2B). These results support our third hypothesized process by which these high levels of GWH originated from a hybridization event between *V. planifolia* and another haplotype or closely related species, as has been evidenced with *Litchi chinensis* Sonn. (G. Hu et al., 2022). Furthermore, it is possible that these accessions resulted from multiple domestication processes and therefore might reveal multiple unique parental origins. Using the "Daphna" cultivar as a genomic reference could inhibit the identification of parental species because dissimilar parental genomic regions might not map to the "Daphna" genome; therefore, signatures will not be detected. A genome assembly of a

less introgressed *V. planifolia* cultivar, such as MEX67 (which did not exhibit a signature of hybridization), might offer a better reference for the species and further help to disentangle the domestication processes that have affected this crop. Additional research is required to identify alternative parental candidates and determine the genomic origin of these Mexican *V. planifolia* accessions. Of the 106 species within the *Vanilla* genus, parental candidates should include the 13 species within the *V. planifolia* taxonomical group (Soto Arenas & Cribb, 2010) and the ten species co-occurring within *V. planifolia's* current extended distribution (Ellestad et al., 2021). A comprehensive phylogenomic framework of crop-wild relatives would greatly advance this objective.

Multiple domestication events within Mexico

Results additionally reveal signatures of multiple domestication events within the highly heterozygous Mexican *V. planifolia* accessions. Since this crop is typically propagated clonally, the three distinct clusters within the PCA on haplotype A (Figure 5.2C) suggest multiple origins. Within the main cluster containing all accessions except MEX31 and MEX20, the slight variation between individuals may reflect the accumulation of somatic point mutations as has been reported among vegetatively propagated individuals (Favre et al., 2022). Variations in the positioning of individuals within the PCA of SNPs along each reference haplotype (Figure 5.2B) may also indicate introgression among these individuals.

Maintaining vanilla's genetic resources

Within this study, multiple genomic origins of cultivated *V. planifolia* involving different species have been identified revealing more variation within this predominantly clonally propagated crop than previously thought. At least two distinct hybridization

events that have occurred within cultivated *V. planifolia*: one in "Daphna" cultivars between *V. planifolia* and *V. pompona,* and at least one more in highly heterozygous cultivated accessions from Mexico. These findings provide a clearer, yet incomplete, illustration of vanilla's evolutionary histories and highlights their importance for maintaining genetic diversity within the crop. Especially within its center of origin, Mexico, the conservation of crop-wild relative diversity and the genetic diversity within cultivated *V. planifolia* originating from different domestication events is needed for global crop improvement and to ensure the sustainability of this iconic spice.

REFERENCES

Ackerman, J. D. (1983). Specificity and mutual dependency of the orchid-euglossine bee interaction. *Biological Journal of the Linnean Society*, *20*(3), 301–314. https://doi.org/10.1111/j.1095-8312.1983.tb01878.x

Agropecuaria, T. F. M. (2019). *Reducirá Veracruz 90 % la cosecha de vainilla por sequía*. https://www.tierrafertil.com.mx/producira-veracruz-90-menos-de-vainilla-por-sequia/

Aktas, C. (2015). *R Package haplotypes: Haplotype Inference and Statistical Analysis of Genetic Variation*. https://CRAN.R-project.org/package=haplotypes

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andrews, H. C., Andrews, H. C., Bensley, T., Haworth, A. H., Jackson, G., & Kennedy, J. (1797). *The botanist's repository, for new and rare plants*. https://www.biodiversitylibrary.org/item/109767

Arditti, J., Rao, A. N., & Nair, H. (2009). History-pollination hand-pollination of Vanilla: How many discoverers? In *Orchid Biology: Reviews and Perspectives, X* (pp. 233–249). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8802-5_7

Argout, X., Salse, J., Aury, J., Guiltinan, M., … G. D.-N., & 2011, undefined. (2011). The genome of Theobroma cacao. *Nature*. https://www.nature.com/articles/ng.736

Armenta-Montero, S., Menchaca-García, R., Pérez-Silva, A., & Velázquez-Rosas, N. (2022). Changes in the Potential Distribution of Vanilla planifolia Andrews under Different Climate Change Projections in Mexico. *Sustainability 2022, Vol. 14, Page 2881*, *14*(5), 2881. https://doi.org/10.3390/SU14052881

Arnone, I. S., Trajano, E., Pulchério-Leite, A., & Passos, F. D. C. (2016). Deslocamento de longa distancia por um grande morcego frugıvoro, artibeus lituratus (Olfers, 1818), no sudeste do brasil (chiroptera, phyllostomidae): Evidencia de migracao de morcegos neotropicais? *Biota Neotropica*, *16*(1), 1–6. https://doi.org/10.1590/1676-0611-BN-2015-0026

Azevedo da Silva, R., Souza, G., Santos Lima Lemos, L., Vanderlei Lopes, U., Geó rgia Ribeiro Braz Patrocínio, N., Moysés Alves, R., Helena Marcellino, L., Clement, D., Micheli, F., & Peres Gramacho, K. (2017). *Genome size, cytogenetic data and transferability of EST-SSRs markers in wild and cultivated species of the genus Theobroma L. (Byttnerioideae, Malvaceae)*. https://doi.org/10.1371/journal.pone.0170799

Azofeifa-Bolaños, J. B., Gigant, L. R., Nicolás-García, M., Pignal, M., Tavares-González, F. B., Hágsater, E., Salazar-Chávez, G. A., Reyes-López, D., Archila-Morales, F. L., García-García, J. A., Da Silva, D., Allibert, A., Solano-Campos, F., Rodríguez-Jimenes, G. D. C., Paniagua-Vásquez, A., Besse, P., Pérez-Silva, A., & Grisoni, M. (2017). A new vanilla species from Costa Rica closely related to V. planifolia (orchidaceae). *European Journal of Taxonomy*, *2017*(284). https://doi.org/10.5852/ejt.2017.284

Bacon, C. D., Mora, A., Wagner, W. L., & Jaramillo, C. A. (2013). Testing geological models of evolution of the Isthmus of Panama in a phylogenetic framework. *Botanical Journal of the Linnean Society*, *171*(1), 287–300. https://doi.org/10.1111/j.1095-8339.2012.01281.x

Baker, A. (2018). *The Bittersweet Cost of Madagascar's Vanilla Boom | Time*. http://time.com/5308143/vanilla-price-climate-change-madagascar/

Bellon, M. R., Gotor, E., & Caracciolo, F. (2015). Conserving landraces and improving livelihoods: how to assess the success of on-farm conservation projects? *International Journal of Agricultural Sustainability*, *13*(2), 167–182. https://doi.org/10.1080/14735903.2014.986363

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., & Wheeler, D. L. (2002). GenBank. *Nucleic Acids Research*, *30*(1), 17–20. https://doi.org/10.1093/nar/30.1.17

Besse, P., Da Silva, D., Bory, S., Grisoni, M., Le Bellec, F., & Duval, M. F. (2004). RAPD genetic diversity in cultivated vanilla: Vanilla planifolia, and relationships with V. tahitensis and V. pompona. *Plant Science*, *167*(2), 379–385. https://doi.org/10.1016/j.plantsci.2004.04.007

Besse, P., Da Silva, D., & Grisoni, M. (2021). Plant DNA Barcoding Principles and Limits: A Case Study in the Genus Vanilla. In *Methods in Molecular Biology* (Vol. 2222, pp. 131–148). Humana Press Inc. https://doi.org/10.1007/978-1-0716-0997-2_8

Bini, L. M., Diniz-Filho, J. A. F., Rangel, T. F. L. V. B., Bastos, R. P., & Pinto, M. P. (2006). Challenging Wallacean and Linnean shortfalls: knowledge gradients and

conservation planning in a biodiversity hotspot. *Diversity and Distributions*, *12*(5), 475–482. https://doi.org/10.1111/j.1366-9516.2006.00286.x

Bivand, R., & Rundel, C. (2019). *Interface to Geometry Engine - Open Source ('GEOS') [R package rgeos version 0.4-2]*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/rgeos/index.html

Black, W. (1970). The nature and inheritance of field resistance to late blight (Phytophthora infestans) in potatoes. *American Potato Journal*, *47*(8), 279–288. https://doi.org/10.1007/BF02863003

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/BIOINFORMATICS/BTU170

Bone, R. E., Smith, J. A. C., Arrigo, N., & Buerki, S. (2015). A macro-ecological perspective on crassulacean acid metabolism (CAM) photosynthesis evolution in Afro-Madagascan drylands: Eulophiinae orchids as a case study. *New Phytologist*, *208*(2), 469–481. https://doi.org/10.1111/nph.13572

Bonilla-Moheno, M., & Aide, T. M. (2020). Beyond deforestation: Land cover transitions in Mexico. *Agricultural Systems*, *178*(November 2019), 102734. https://doi.org/10.1016/j.agsy.2019.102734

Bory, S., Catrice, O., Brown, S., Leitch, I. J., Gigant, R., Chiroleu, F., Grisoni, M., Duval, M. F., & Besse, P. (2008). Natural polyploidy in Vanilla planifolia (Orchidaceae). *Genome*, *51*(10), 816–826. https://doi.org/10.1139/G08-068

Bory, S., Grisoni, M., Duval, M.-F., & Besse, P. (2008). Biodiversity and preservation of vanilla: present state of knowledge. *Genetic Resources and Crop Evolution*, *55*(4), 551–571. https://doi.org/10.1007/s10722-007-9260-3

Bouetard, A., Lefeuvre, P., Gigant, R., Bory, S., Pignal, M., Besse, P., & Grisoni, M. (2010). Evidence of transoceanic dispersion of the genus Vanilla based on plastid DNA phylogenetic analysis. *Molecular Phylogenetics and Evolution*, *55*(2), 621–630. https://doi.org/10.1016/j.ympev.2010.01.021

Brewer, G. E., Clarkson, J. J., Maurin, O., Zuntini, A. R., Barber, V., Bellot, S., Biggs, N., Cowan, R. S., Davies, N. M. J., Dodsworth, S., Edwards, S. L., Eiserhardt, W. L., Epitawalage, N., Frisby, S., Grall, A., Kersey, P. J., Pokorny, L., Leitch, I. J., Forest, F., & Baker, W. J. (2019). Factors Affecting Targeted Sequencing of 353 Nuclear Genes From Herbarium Specimens Spanning the Diversity of Angiosperms. *Frontiers in Plant Science*, *10*, 1102. https://doi.org/10.3389/fpls.2019.01102

Brondizo, E. S., Settele, J., Diaz, S., & Ngo, H. T. (2019). *Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*.

Bruman, H. (1948). The Culture History of Mexican Vanilla. *The Hispanic American Historical Review*, *28*(3), 360. https://doi.org/10.2307/2507753

Brummitt, N., Bachman, S. P., Aletrari, E., Chadburn, H., Griffiths-Lee, J., Lutz, M., Moat, J., Rivers, M. C., Syfert, M. M., & Nic Lughadha, E. M. (2015). The Sampled Red List Index for Plants, phase II: ground-truthing specimen-based conservation assessments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1662), 20140015–20140015. https://doi.org/10.1098/rstb.2014.0015

Buerki, S., & Baker, W. J. (2016). Collections-based research in the genomic era. *Biological Journal of the Linnean Society*, *117*(1), 5–10. https://doi.org/10.1111/bij.12721

Buerki, S., Callmander, M. W., Bachman, S., Moat, J., Labat, J.-N., & Forest, F. (2015). Incorporating evolutionary history into conservation planning in biodiversity hotspots. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1662), 20140014. https://doi.org/10.1098/rstb.2014.0014

Buerki, S., Forest, F., Acevedo-Rodríguez, P., Callmander, M. W., Nylander, J. A. A., Harrington, M., Sanmartín, I., Küpfer, P., & Alvarez, N. (2009). Plastid and nuclear DNA markers reveal intricate relationships at subfamilial and tribal levels in the soapberry family (Sapindaceae). *Molecular Phylogenetics and Evolution*, *51*(2), 238–258. https://doi.org/10.1016/j.ympev.2009.01.012

Cameron, K. Michael. (2011). *Vanilla orchids: natural history and cultivation*. Timber Press.

Casañas, F., Simó, J., Casals, J., & Prohens, J. (2017). Toward an evolved concept of landrace. *Frontiers in Plant Science*, *8*(FEBRUARY), 145. https://doi.org/10.3389/FPLS.2017.00145/BIBTEX

CBI. (2018). *Exporting vanilla to Europe*. https://www.cbi.eu/market-information/spices-herbs/vanilla#product-description

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(30), E6089–E6096. https://doi.org/10.1073/pnas.1704949114

Chambers, A., Cibrián-Jaramillo, A., Karremans, A. P., Moreno Martinez, D., Hernandez-Hernandez, J., Brym, M., Resende, M. F. R., Moloney, R., Sierra, S. N., Hasing, T., Alomia, Y. A., & Hu, Y. (2021). Genotyping-By-Sequencing diversity analysis of international Vanilla collections uncovers hidden diversity and enables plant improvement. *Plant Science*, *311*, 111019. https://doi.org/10.1016/J.PLANTSCI.2021.111019

Charif, D., & Lobry, J. R. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In U. Bastolla, M. Porto, H. E. Roman, & M. Vendruscolo (Eds.), *Structural approaches to sequence evolution: Molecules, networks, populations* (pp. 207–232). Springer Verlag.

Chaves, A. v., Freitas, G. H. S., Vasconcelos, M. F., & Santos, F. R. (2015). Biogeographic patterns, origin and speciation of the endemic birds from eastern Brazilian mountaintops: A review. *Systematics and Biodiversity*, *13*(1), 1–16. https://doi.org/10.1080/14772000.2014.972477

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y., & Leon, C. (2010). Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. *PLoS ONE*, *5*(1), e8613. https://doi.org/10.1371/journal.pone.0008613

Cheng, T., Xu, C., Lei, L., & Li, C. (2015). Barcoding the kingdom Plantae : New PCR primers for ITS regions of plants with improved universality and specificity Barcoding the kingdom Plantae: new PCR primers for ITS regions of plants with

improved universality and specificity. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.12438

Chumacero de Schawe, C., Durka, W., Tscharntke, T., Hensen, I., & Kessler, M. (2013). Gene flow and genetic diversity in cultivated and wild cacao (Theobroma cacao) in Bolivia. *American Journal of Botany*, *100*(11), 2271–2279. https://doi.org/10.3732/ajb.1300025

Claus, G., Vanhove, W., Damme, P. Van, & Smagghe, G. (2018). Challenges in Cocoa Pollination: The Case of Côte d'Ivoire. In *Pollination in Plants*. InTech. https://doi.org/10.5772/intechopen.75361

Cronquist, A. (1978). Once again, what is a species? *Biosystematics in Agriculture*, 3–20.

Cuatrecasas, J. (1964). *Cacao and Its Allies, A Taxonomic Revision of the Genus Theobroma - Google Play*. Contribution from the United States National Herbarium.

Dalrymple, M. (2006). Deforestation and Land-Use Change in Veracruz, Mexico: A Remote Sensing Analysis. *University of New Orleans Theses and Dissertations*. https://scholarworks.uno.edu/td/395

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. https://doi.org/10.1093/GIGASCIENCE/GIAB008

Delassus, M. (1963). La lutte contre la fusariose du vanillier par les methodes genetiques. *L'agronomie Tropicale*, *18*, 245–246.

Denham, T., Barton, H., Castillo, C., Crowther, A., Dotte-Sarout, E., Florin, S. A., Pritchard, J., Barron, A., Zhang, Y., & Fuller, D. Q. (2020). The domestication

syndrome in vegetatively propagated field crops. *Annals of Botany*, *125*(4), 581–597. https://doi.org/10.1093/AOB/MCZ212

Diazgranados, M., Allkin, B., Black, N., Cámara-Leret, R., Canteiro, C., Carretero, J., Eastwood, R., Hargreaves, S., Hudson, A., Milliken, W., Nesbitt, M., Ondo, I., Patmore, K., Pironon, S., Turner, R., & Ulian, T. (2020). *World Checklist of Useful Plant Species*. https://doi.org/10.5063/F1CV4G34

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N. D., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., Hansen, M., Locke, H., Ellis, E. C., Jones, B., Barber, C. V., Hayes, R., Kormos, C., Martin, V., Crist, E., … Saleem, M. (2017). An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm. *BioScience*, *67*(6), 534–545. https://doi.org/10.1093/biosci/bix014

Divakaran, M., Nirmal Babu, K., Ravindran, P. N., & Peter, K. v. (2006). Interspecific hybridization in vanilla and molecular characterization of hybrids and selfed progenies using RAPD and AFLP markers. *Scientia Horticulturae*, *108*, 414–422. https://doi.org/10.1016/j.scienta.2006.02.018

Doebley, J. (1989). Isozymic Evidence and the Evolution of Crop Plants. In *Isozymes in Plant Biology* (pp. 165–191). Springer Netherlands. https://doi.org/10.1007/978-94-009-1840-5_9

Doebley, J., Gaut, B., & Smith, B. (2006). The Molecular Genetics of Crop Domestication. *Cell*, *127*(7), 1309–1321. https://doi.org/10.1016/j.cell.2006.12.006

Donatti, C. I., Harvey, C. A., Martinez-Rodriguez, M. R., Vignola, R., & Rodriguez, C. M. (2019). Vulnerability of smallholder farmers to climate change in Central America and Mexico: current knowledge and research gaps. In *Climate and*

Development (Vol. 11, Issue 3, pp. 264–286). Taylor and Francis Ltd. https://doi.org/10.1080/17565529.2018.1442796

Donini, P., Bory, S., Catrice, O., Brown, S., Leitch, I. J., Gigant, R., Chiroleu, F., Grisoni, M., Duval, M.-F., & Besse, P. (2008). Natural polyploidy in *Vanilla planifolia* (Orchidaceae). *Genome*, *51*(10), 816–826. https://doi.org/10.1139/G08-068

Ebert, A. W., & Engels, J. M. M. (2020). Plant Biodiversity and Genetic Resources Matter! *Plants*, *9*(12). https://doi.org/10.3390/plants9121706

Ellestad, P., Forest, F., Serpe, M., Novak, S. J., & Buerki, S. (2021). Harnessing large-scale biodiversity data to infer the current distribution of *Vanilla planifolia* (Orchidaceae). *Botanical Journal of the Linnean Society*, *XX*, 1–16. https://doi.org/10.1093/botlinnean/boab005

Ellestad, P., Pérez-Farrera, M. A., & Buerki, S. (2022). Genomic Insights into Cultivated Mexican Vanilla planifolia Reveal High Levels of Heterozygosity Stemming from Hybridization. *Plants 2022, Vol. 11, Page 2090*, *11*(16), 2090. https://doi.org/10.3390/PLANTS11162090

Ellestad, P., Pérez-Farrera, M. A., Forest, F., & Buerki, S. (2022). Uncovering haplotype diversity in cultivated Mexican vanilla species. *American Journal of Botany*. https://doi.org/10.1002/ajb2.16024

Ellis, E. C., Antill, E. C., & Kreft, H. (2012). All Is Not Loss: Plant Biodiversity in the Anthropocene. *PLoS ONE*, *7*(1), e30535. https://doi.org/10.1371/journal.pone.0030535

Engels, J. M. M., Ebert, A. W., Thormann, I., & de Vicente, M. C. (2006). Centres of crop diversity and/or origin, genetically modified crops and implications for plant

genetic resources conservation. *Genetic Resources and Crop Evolution*, *53*(8), 1675–1688. https://doi.org/10.1007/S10722-005-1215-Y/METRICS

Eyre-Walker, A., Gaut, R. L., Hilton, H., Feldman, D. L., & Gaut, B. S. (1998). Investigation of the bottleneck leading to the domestication of maize. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(8), 4441–4446. https://doi.org/10.1073/pnas.95.8.4441

Fantini, D. (2019). *easyPubMed*. https://www.data-pulse.com/dev_site/easypubmed/

Favre, F., Jourda, C., Grisoni, M., Piet, Q., Rivallan, R., Dijoux, J.-B., Hascoat, J., Lepers-Andrzejewski, S., Besse, P., & Charron, C. (2022). A genome-wide assessment of the genetic diversity, evolution and relationships with allied species of the clonally propagated crop Vanilla planifolia Jacks. ex Andrews. *Genetic Resources and Crop Evolution*. https://doi.org/10.1007/s10722-022-01362-1

Fazekas, A. J., Burgess, K. S., Kesanakurti, P. R., Graham, S. W., Newmaster, S. G., Husband, B. C., Percy, D. M., Hajibabaei, M., & Barrett, S. C. H. (2008). Multiple Multilocus DNA Barcodes from the Plastid Genome Discriminate Plant Species Equally Well. *PLOS ONE*, *3*(7), e2802. https://doi.org/10.1371/JOURNAL.PONE.0002802

Feliner, G. N., & Rosselló, J. A. (2007). Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. In *Molecular Phylogenetics and Evolution* (Vol. 44, Issue 2, pp. 911–919). Mol Phylogenet Evol. https://doi.org/10.1016/j.ympev.2007.01.013

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Figueira, A., Janick, J., & Goldsbrough, P. (2019). Genome Size and DNA Polymorphism in Theobroma cacao. *Journal of the American Society for Horticultural Science*, *117*(4), 673–677. https://doi.org/10.21273/jashs.117.4.673

Flores Jiménez, Á., Reyes Lopez, D., Jiménez García, D., Romero Arenas, O., Rivera Tapia, J. A., Huerta Lara, M., & Pérez Silva, A. (2017). Diversidad de Vanilla spp. (Orchidaceae) y sus perfiles bioclimáticos en México. *Revista de Biología Tropical*, *65*(3), 975. https://doi.org/10.15517/rbt.v65i3.29438

Fontaine, C., Dajoz, I., Meriguet, J., & Loreau, M. (2005). Functional Diversity of Plant–Pollinator Interaction Webs Enhances the Persistence of Plant Communities. *PLoS Biology*, *4*(1), e1. https://doi.org/10.1371/journal.pbio.0040001

Food and Agriculture Organization of the United Nations (FAO). (2020). *FAOSTAT*. http://www.fao.org/faostat/en/#rankings/countries_by_commodity

Ford-Lloyd, B. V., Schmidt, M., Armstrong, S. J., Barazani, O., Engels, J., Hadas, R., Hammer, K., Kell, S. P., Kang, D., Khoshbakht, K., Li, Y., Long, C., Lu, B.-R., Ma, K., Nguyen, V. T., Qiu, L., Ge, S., Wei, W., Zhang, Z., & Maxted, N. (2011). Crop Wild Relatives—Undervalued, Underutilized and under Threat? *BioScience*, *61*(7), 559–565. https://doi.org/10.1525/bio.2011.61.7.10

Gao, L. Z. (2003). The conservation of Chinese rice biodiversity: Genetic erosion, ethnobotany and prospects. *Genetic Resources and Crop Evolution*, *50*(1), 17–32. https://doi.org/10.1023/A:1022933230689

GBIF.org (16 April 2019) GBIF Occurrence Download https://doi.org/10.15468/dl.sifewb. (n.d.).

Geographic, N. (n.d.). *Indigenous People of Mexico*. Retrieved May 31, 2021, from https://maps.esri.com/legends/Mex_Indians5/index.html

Gepts, P. (2006). Plant Genetic Resources Conservation and Utilization: The Accomplishments and Future of a Societal Insurance Policy. *Crop Science*, *46*(5), 2278–2292. https://doi.org/10.2135/CROPSCI2006.03.0169GAS

GitHub. (n.d.). *https://github.com/svenbuerki/VanillaGenomicsCode*. Retrieved July 4, 2022, from https://github.com/svenbuerki/VanillaGenomicsCode

Gomes, V. H. F., Vieira, I. C. G., Salomao, R. P., & ter Steege, H. (2019). Amazonian tree species threatened by deforestation and climate change. *Nature Climate Change*, *9*, 547–553.

Harlan, J. R., & Wet, J. M. J. (1971). TOWARD A RATIONAL CLASSIFICATION OF CULTIVATED PLANTS. *TAXON*, *20*(4), 509–517. https://doi.org/10.2307/1218252

Hasing, T., Tang, H., Brym, M., Khazi, F., Huang, T., & Chambers, A. H. (2020). A phased Vanilla planifolia genome enables genetic improvement of flavour and production. *Nature Food*, *1*(12), 811–819. https://doi.org/10.1038/s43016-020-00197-2

Hernández-Ruíz, J., Herrera-Cabrera, B. E., Delgado-Alvarado, A., Salazar-Rojas, V. M., Bustamante-Gonzalez, Á., Campos-Contreras, J. E., & Ramírez-Juarez, J. (2016). Distribución potencial y características geográficas de poblaciones silvestres de

Vanilla planifolia (Orchidaceae) en Oaxaca, México. *Revista de Biologia Tropical*, *64*(1), 235–246. https://doi.org/10.15517/rbt.v64i1.17854

Hijmans, R. J. (2019a). *Geographic Data Analysis and Modeling [R package raster version 2.8-19]*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/raster/index.html

Hijmans, R. J. (2019b). *Spherical Trigonometry [R package geosphere version 1.5-7]*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/geosphere/index.html

Hinsley, A., de Boer, H. J., Fay, M. F., Gale, S. W., Gardiner, L. M., Gunasekara, R. S., Kumar, P., Masters, S., Metusala, D., Roberts, D. L., Veldman, S., Wong, S., & Phelps, J. (2018). A review of the trade in orchids and its implications for conservation. *Botanical Journal of the Linnean Society*, *186*(4), 435–455. https://doi.org/10.1093/botlinnean/box083

Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M. W., Cowan, R. S., Erickson, D. L., Fazekas, A. J., Graham, S. W., James, K. E., Kim, K. J., John Kress, W., Schneider, H., van AlphenStahl, J., Barrett, S. C. H., van den Berg, C., Bogarin, D., … Little, D. P. (2009). A DNA barcode for land plants. *PNAS*, *106*(31), 12794–12797. https://doi.org/10.1073/pnas.0905845106

Hu, G., Feng, J., Xiang, X., Wang, J., Salojärvi, J., Liu, C., Wu, Z., Zhang, J., Liang, X., Jiang, Z., Liu, W., Ou, L., Li, J., Fan, G., Mai, Y., Chen, C., Zhang, X., Zheng, J., Zhang, Y., … Li, J. (2022). Two divergent haplotypes from a highly heterozygous lychee genome suggest independent domestication events for early and late-

maturing cultivars. *Nature Genetics*, *54*(1), 73–83. https://doi.org/10.1038/s41588-021-00971-3

Hu, Y., Resende, M. F. R., Bombarely, A., Brym, M., Bassil, E., & Chambers, A. H. (2019). Genomics-based diversity analysis of Vanilla species using a Vanilla planifolia draft genome and Genotyping-By-Sequencing. *Scientific Reports*, *9*(1), 1–16. https://doi.org/10.1038/s41598-019-40144-1

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. In *BIOINFORMATICS APPLICATIONS NOTE* (Vol. 17, Issue 8). Yang and Rannala.

Hyten, D. L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R. L., Costa, J. M., Specht, J. E., Shoemaker, R. C., Cregan, P. B., & Tanksley, S. D. (2006). *Impacts of genetic bottlenecks on soybean genome diversity*.

IUCN. (2019). *The IUCN Red List of Threatened Species*. International Union for Conservation of Nature - IUCN. https://www.iucnredlist.org/

Jägermeyr, J., Müller, C., Ruane, A. C., Elliott, J., Balkovic, J., Castillo, O., Faye, B., Foster, I., Folberth, C., Franke, J. A., Fuchs, K., Guarin, J. R., Heinke, J., Hoogenboom, G., Iizumi, T., Jain, A. K., Kelly, D., Khabarov, N., Lange, S., … Rosenzweig, C. (2021). Climate impacts on global agriculture emerge earlier in new generation of climate and crop models. *Nature Food 2021 2:11*, *2*(11), 873–885. https://doi.org/10.1038/s43016-021-00400-y

Jarvis, A., Lane, A., & Hijmans, R. J. (2008). The effect of climate change on crop wild relatives. *Agriculture, Ecosystems and Environment*, *126*(1–2), 13–23. https://doi.org/10.1016/j.agee.2008.01.013

Jin, J. J., Yu, W. Bin, Yang, J. B., Song, Y., Depamphilis, C. W., Yi, T. S., & Li, D. Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*(1), 1–31. https://doi.org/10.1186/S13059-020-02154-5/FIGURES/6

Johnson, M. A. G., Okorny, L. I. S. A. P., Odsworth, S. T. D., Otigué, L. A. R. B., Owan, R. O. S. C., & Lison, A. (2018). *A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering*. *0*(0), 1–13. https://doi.org/10.1093/sysbio/syy086

Jones, W. E., Kuehnle, A. R., & Arumuganathan, K. (1998). Nuclear DNA Content of 26 Orchid (Orchidaceae) Genera with Emphasis on Dendrobium. In *Annals of Botany* (Vol. 82).

Joppa, L. N., Roberts, D. L., & Pimm, S. L. (2011). How many species of flowering plants are there? *Proceedings of the Royal Society B: Biological Sciences*, *278*(1705), 554–559. https://doi.org/10.1098/rspb.2010.1004

Jump, A. S., Marchant, R., & Peñuelas, J. (2009). Environmental change and the option value of genetic diversity. *Trends in Plant Science*, *14*(1), 51–58. https://doi.org/10.1016/j.tplants.2008.10.002

Karremans, A. P., Chinchilla, I. F., Rojas-Alvarado, G., Cedeño-Fonseca, M., Damián, A., & Léotard, G. (2020). A REAPPRAISAL OF NEOTROPICAL VANILLA. WITH A NOTE ON TAXONOMIC INFLATION AND THE IMPORTANCE OF ALPHA TAXONOMY IN BIOLOGICAL STUDIES. *LANKESTERIANA*, *20*(3), 395–497. https://doi.org/10.15517/lank.v20i3.45203

Katche, E., Quezada-Martinez, D., Katche, E. I., Vasquez-Teuber, P., & Mason, A. S. (2019). *Interspecific Hybridization for Brassica Crop Improvement*. https://doi.org/10.20900/cbgg20190007

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kiang, Y. T., Antonovics, D., & Wu, L. (1979). The extinction of wild rice (Oryza perennis formosana) in Taiwan. *Journal of Asian Ecology*, *1*, 1–9.

Kier, G., Mutke, J., Dinerstein, E., Ricketts, T. H., Küper, W., Kreft, H., & Barthlott, W. (2005). Global patterns of plant diversity and floristic knowledge. *Journal of Biogeography*, *32*(7), 1107–1116. https://doi.org/10.1111/j.1365-2699.2005.01272.x

Kilian, B., Martin, W., & Salamini, F. (2010). Genetic diversity, evolution and domestication of wheat and barley in the Fertile Crescent. In *Evolution in Action: Case studies in Adaptive Radiation, Speciation and the Origin of Biodiversity* (pp. 137–166). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-12425-9_8

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, *16*(2), 111–120. https://doi.org/10.1007/BF01731581

Kokot, M., Dlugosz, M., & Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, *33*(17), 2759–2761. https://doi.org/10.1093/BIOINFORMATICS/BTX304

Koyyappurath, S., Conéjéro, G., Dijoux, J. B., Lapeyre-Montès, F., Jade, K., Chiroleu, F., Gatineau, F., Verdeil, J. L., Besse, P., & Grisoni, M. (2015). Differential responses of vanilla accessions to root rot and colonization by fusarium oxysporum f. sp. radicis-vanillae. *Frontiers in Plant Science*, *6*(DEC). https://doi.org/10.3389/fpls.2015.01125

Koziol, L., Rieseberg, L. H., Kane, N., & Bever, J. D. (2012). Reduced drought tolerance during domestication and the evolution of weediness results from tolerance-growth trade-offs. *Evolution*, *66*(12), 3803–3814. https://doi.org/10.1111/j.1558-5646.2012.01718.x

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Lahive, F., Hadley, P., & Daymond, A. J. (2019). The physiological responses of cacao to the environment and the implications for climate change resilience. A review. In *Agronomy for Sustainable Development* (Vol. 39, Issue 1, pp. 1–22). Springer-Verlag France. https://doi.org/10.1007/s13593-018-0552-0

Lanaud, C., Hamon, P., & Duperray, C. (1992). Estimation of nuclear DNA content of Theobroma cacao L. by flow cytometry. *Undefined*.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods 2012 9:4*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Larridon, I., Villaverde, T., Zuntini, A. R., Pokorny, L., Brewer, G. E., Epitawalage, N., Fairlie, I., Hahn, M., Kim, J., Maguilla, E., Maurin, O., Xanthos, M., Hipp, A. L., Forest, F., & Baker, W. J. (2020). Tackling Rapid Radiations With Targeted

Sequencing. *Frontiers in Plant Science*, *10*, 1655. https://doi.org/10.3389/FPLS.2019.01655/BIBTEX

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *BIOINFORMATICS APPLICATIONS*, *30*(22), 3276–3278. https://doi.org/10.1093/bioinformatics/btu531

Le Breton, T. D., Zimmer, H. C., Gallagher, R. V., Cox, M., Allen, S., & Auld, T. D. (2019). Using IUCN criteria to perform rapid assessments of at-risk taxa. *Biodiversity and Conservation*, *28*(4), 863–883. https://doi.org/10.1007/S10531-019-01697-9/FIGURES/3

Lepers-Andrzejewski, S., Siljak-Yakovlev, S., Brown, S. C., Wong, M., & Dron, M. (2011). Diversity and dynamics of plant genome size: an example of polysomaty from a cytogenetic study of Tahitian vanilla (Vanilla xtahitensis, Orchidaceae). *American Journal of Botany*, *98*(6), 986–997. https://doi.org/10.3732/AJB.1000415

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/BIOINFORMATICS/BTY191

Li, J., Demesyeux, L., Brym, M., & Chambers, A. H. (2020). Development of species-specific molecular markers in Vanilla for seedling selection of hybrids. *Molecular Biology Reports*, *47*(3), 1905–1920. https://doi.org/10.1007/s11033-020-05287-9

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., & Chen, S. (2015). Plant DNA barcoding: from gene to genome. *Biological Reviews of the Cambridge Philosophical Society*, *90*(1), 157–166. https://doi.org/10.1111/brv.12104

Lira, R., Téllez, O., & Dávila, P. (2009). The effects of climate change on the geographic distribution of Mexican wild relatives of domesticated Cucurbitaceae. *Genetic*

Resources and Crop Evolution, *56*(5), 691–703. https://doi.org/10.1007/s10722-008-9394-y

Lomolino, M. (2004). Conservation Biogeography. In M. Lomolino & L. Heaney (Eds.), *Frontiers of Biogeography*. Sinauer Associates, Inc. .

López-Baucells, A., Rocha, R., Bobrowiec, P. E. D., Palmeirim, J. M., & Meyer, C. F. J. (2016). *Field guide to Amazonian bats*. National Institute of Amazonian Research.

Lozano Rodríguez, M. Á., Luna Rodríguez, M., Pech Canché, J. M., Menchaca García, R. A., & Cerdán Cabrera, C. R. (2022). Visit frequency of Euglossine bees (Hymenoptera: Apidae) to mature fruits of Vanilla planifolia (Orchidaceae). *Acta Botanica Mexicana*, *129*. https://doi.org/10.21829/ABM129.2022.2001

Lubinsky, P., Bory, S., Hernández Hernández, J., Kim, S. C., & Gómez-Pompa, A. (2008). Origins and dispersal of cultivated vanilla (Vanilla planifolia Jacks. [Orchidaceae]). *Economic Botany*, *62*(2), 127–138. https://doi.org/10.1007/s12231-008-9014-y

Lubinsky, P., Cameron, K. M., Molina, M. C., Wong, M., Lepers-Andrzejewski, S., Gómez-Pompa, A., & Kim, S. C. (2008). Neotropical roots of a Polynesian spice: The hybrid origin of Tahitian vanilla, vanilla tahitensis (Orchidaceae). *American Journal of Botany*, *95*(8), 1040–1047. https://doi.org/10.3732/ajb.0800067

Majee, M., Maitra, S., Dastidar, K. G., Pattnaik, S., Chatterjee, A., Hait, N. C., Das, K. P., & Majumder, A. L. (2004). A novel salt-tolerant L-myo-Inositol-1-phosphate synthase from Porteresia coarctata (Roxb.) tateoka, a halophytic wild rice. Molecular cloning, bacterial overexpression, characterization, and functional

introgression into tobacco-conferring salt tolerance phenotype. *Journal of Biological Chemistry*, *279*(27), 28539–28552. https://doi.org/10.1074/jbc.M310138200

Mammadov, J., Buyyarapu, R., Guttikonda, S. K., Parliament, K., Abdurakhmonov, I. Y., & Kumpatla, S. P. (2018). Wild relatives of maize, rice, cotton, and soybean: Treasure troves for tolerance to biotic and abiotic stresses. In *Frontiers in Plant Science* (Vol. 9, p. 886). Frontiers Media S.A. https://doi.org/10.3389/fpls.2018.00886

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/BIOINFORMATICS/BTR011

MarketWatch. (2020). *(New Report)Vanilla Market Size In 2020-Trends, Industry Share, Growth Drivers, Business Opportunities and Demand Forecast to 2026*. https://www.marketwatch.com/press-release/new-reportvanilla-market-size-in-2020-trends-industry-share-growth-drivers-business-opportunities-and-demand-forecast-to-2026-2020-04-22

Martinson, V. A. (1966). Hybridization of Cacao and Theobroma grandiflora. *Journal of Heredity*, *57*(4), 134–136. https://doi.org/10.1093/oxfordjournals.jhered.a107491

Maxted, N., Ford-Lloyd, B. V., Jury, S., Kell, S., & Scholten, M. (2006). Towards a definition of a crop wild relative. *Biodiversity and Conservation*, *15*(8), 2673–2685. https://doi.org/10.1007/s10531-005-5409-6

Maxted, N., & Guarino, L. (2004). Genetic erosion and genetic pollution of crop wild relatives. *Genetic Erosion and Pollution Assessment Methodologies*.

McKey, D., Elias, M., Pujol, M. E., & Duputié, A. (2010). The evolutionary ecology of clonally propagated domesticated plants. *New Phytologist*, *186*(2), 318–332. https://doi.org/10.1111/J.1469-8137.2010.03210.X

Melton, A. E., Child, A. W., Beard, R. S., Dave, C., Dumaguit, C., Forbey, J. S., Germino, M., de Graaff, M.-A., Kliskey, A., Leitch, I. J., Martinez, P., Novak, S. J., Pellicer, J., Richardson, B. A., Self, D., Serpe, M., & Buerki, S. (2022). A haploid pseudo-chromosome genome assembly for a keystone sagebrush species of western North American rangelands. *G3 Genes Genomes Genetics*. https://doi.org/10.1093/g3journal/jkac122

Meyer, R. S., Duval, A. E., & Jensen, H. R. (2012). Patterns and processes in crop domestication: An historical review and quantitative analysis of 203 global food crops. *New Phytologist*, *196*(1), 29–48. https://doi.org/10.1111/J.1469-8137.2012.04253.X

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gateway Computing Environments Workshop, GCE 2010*. https://doi.org/10.1109/GCE.2010.5676129

Minoo, D., Jayakumar, V. N., Veena, S. S., Vimala, J., Basha, A., Saji, K. v., Nirmal Babu, K., & Peter, K. v. (2008). Genetic variations and interrelationships in Vanilla planifolia and few related species as expressed by RAPD polymorphism. *Genetic Resources and Crop Evolution*, *55*(3), 459–470. https://doi.org/10.1007/s10722-007-9252-3

Moat, J. (2017). *Package "rCAT."* https://doi.org/10.3897/zookeys.150.2109

Murphy, B., Forest, F., Barraclough, T., Rosindell, J., Bellot, S., Cowan, R., Golos, M., Jebb, M., & Cheek, M. (2020). A phylogenomic analysis of Nepenthes (Nepenthaceae). *Molecular Phylogenetics and Evolution*, *144*. https://doi.org/10.1016/J.YMPEV.2019.106668

Nair, R. R., & Ravindran, P. N. (1994). Somatic association of chromosomes and other mitotic abnormalities in Vanilla planifolia (Andrews). *Caryologia*, *47*(1), 65–73. https://doi.org/10.1080/00087114.1994.10797284

Nelson, B. W., Ferreira, C. A. C., da Silva, M. F., & Kawasaki, M. L. (1990). Endemism centres, refugia and botanical collection density in Brazilian Amazonia. *Nature*, *345*(6277), 714–716. https://doi.org/10.1038/345714a0

O'Dea, A., Lessios, H. A., Coates, A. G., Eytan, R. I., Restrepo-Moreno, S. A., Cione, A. L., Collins, L. S., de Queiroz, A., Farris, D. W., Norris, R. D., Stallard, R. F., Woodburne, M. O., Aguilera, O., Aubry, M. P., Berggren, W. A., Budd, A. F., Cozzuol, M. A., Coppard, S. E., Duque-Caro, H., … Jackson, J. B. C. (2016). Formation of the Isthmus of Panama. *Science Advances*, *2*(8), 1–11. https://doi.org/10.1126/SCIADV.1600883

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., Minchin, P. R., O'hara, R. B., Simpson, G. L., Solymos, P., Henry, M., Stevens, H., Szoecs, E., & Maintainer, H. W. (2019). *Package "vegan" Title Community Ecology Package*. https://cran.r-project.org/web/packages/vegan/vegan.pdf

Ollerton, J., Winfree, R., & Tarrant, S. (2011). How many flowering plants are pollinated by animals? *Oikos*, *120*(3), 321–326. https://doi.org/10.1111/J.1600-0706.2010.18644.X

Padgham, M., & Sumner, M. D. (2020). *geodist: Fast, Dependency-Free Geodesic Distance Calculations*.

Pansarin, E. R. (2016). Recent advances on evolution of pollination systems and reproductive biology of Vanilloideae (Orchidaceae). *Lankesteriana*, *16*(2), 255–267. https://doi.org/10.15517/lank.v16i2.26010

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pebesma, E., & Bivand, R. (2018). *Classes and Methods for Spatial Data [R package sp version 1.3-1]* (1.3-1). Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/sp/index.html

Pimm, S. L., & Raven, P. (2000). Extinction by numbers. *Nature*, *403*(6772), 843–845. https://doi.org/10.1038/35002708

Pinaria, A. G., Liew, E. C. Y., & Burgess, L. W. (2010). Fusarium species associated with vanilla stem rot in Indonesia. *Australasian Plant Pathology*, *39*(2), 176–183. https://doi.org/10.1071/AP09079

Plants of the World Online | Kew Science. (n.d.). Retrieved November 22, 2020, from http://powo.science.kew.org/

Pokorny, T., Loose, D., Dyker, G., Javier, J., Quezada-Euán, G., Eltz, T., & Dispersal, T. E. (2015). Dispersal ability of male orchid bees and direct evidence for long-range flights for long-range flights. *Apidologie*, *46*(2). https://doi.org/10.1007/s13592-014-0317-yï

Population | United Nations. (n.d.). Retrieved October 28, 2020, from https://www.un.org/en/sections/issues-depth/population/

Qin, Y., Li, M., Cao, Y., Gao, Y., & Zhang, W. (2017). Molecular thresholds of ITS2 and their implications for molecular evolution and species identification in seed plants. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-17695-2

R Core Team. (2017). *R: A language and environment for statistical computing*. https://www.r-project.org/

Rain, P. (2004). *Vanilla : the cultural history of the world's most popular flavor and fragrance*. J.P. Tarcher/Penguin.

Rain, P., & Lubinsky, P. (2011). Vanilla Use in Colonial Mexico and Traditional Totonac. In Odoux & Grisoni (Eds.), *Vanilla. Medicinal and Aromatic Plants – Industrial Profiles* (pp. 251–259). CRC Press; Taylor & Francis Group.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904. https://doi.org/10.1093/sysbio/syy032

Ramos-Castellá, A. L., Iglesias-Andreu, L. G., Martínez-Castillo, J., Ortíz-Garciá, M., Andueza-Noh, R. H., Octavio-Aguilar, P., & Luna-Rodríguez, M. (2017). Evaluation of molecular variability in germplasm of vanilla (Vanilla planifolia G. Jackson in Andrews) in Southeast Mexico: Implications for genetic improvement and conservation. *Plant Genetic Resources: Characterisation and Utilisation*, *15*(4), 310–320. https://doi.org/10.1017/S1479262115000660

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1). https://doi.org/10.1038/S41467-020-14998-3

Ravindran, P. N. (1979). Nuclear Behaviour in the Sterile Pollen of Vanilla planifolia (Andrews). *Cytologia*, *44*, 391–396.

Reid, F. A. (2009). *A Field Guide to the Mammals of Central America & Southeast Mexico* (2nd ed.). Oxford University Press.

Rice, A., Glick, L., Abadi, S., Einhorn, M., Kopelman, N. M., Salman-Minkov, A., Mayzel, J., Chay, O., & Mayrose, I. (2015). The Chromosome Counts Database (CCDB) - a community resource of plant chromosome numbers. In *New Phytologist* (Vol. 206, Issue 1, pp. 19–26). Blackwell Publishing Ltd. https://doi.org/10.1111/nph.13191

Richardson, J. E., Whitlock, B. A., Meerow, A. W., & Madriñán, S. (2015). The age of chocolate: a diversification history of Theobroma and Malvaceae. *Frontiers in Ecology and Evolution*, *3*(NOV), 120. https://doi.org/10.3389/fevo.2015.00120

Rodolphe, G., Severine, B., Michel, G., & Pascale, B. (2012). Biodiversity and Evolution in the Vanilla Genus. *The Dynamical Processes of Biodiversity - Case Studies of Evolution and Spatial Distribution*, *figure 1*. https://doi.org/10.5772/24567

Rosenthal, J. P., & Dirzo, R. (1997). Effects of life history, domestication and agronomic selection on plant defence against insects: Evidence from maizes and wild relatives. *Evolutionary Ecology*, *11*(3), 337–355. https://doi.org/10.1023/A:1018420504439

Sackville Hamilton, N. R. (Institute of G. and E. R. A. (United K. G. R. U. (1999). *Genetic erosion issues in temperate grassland*. Vyzkumny Ustav Rostlinne Vyroby.

Sala, O. E., Chapin, F. S., Armesto, J. J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L. F., Jackson, R. B., Kinzig, A., Leemans, R., Lodge, D. M., Mooney, H. A., Oesterheld, M., Poff, N. L., Sykes, M. T., Walker, B. H., Walker, M., & Wall, D. H. (2000). Global biodiversity scenarios for the year 2100. *Science (New York, N.Y.)*, *287*(5459), 1770–1774. http://www.ncbi.nlm.nih.gov/pubmed/10710299

Sbilordo, S. H., Martin, O. Y., & Ribi, G. (2012). Chromosome inheritance and reproductive barriers in backcrosses between two hybridizing Viviparus snail species. *Journal of Molluscan Studies*, *78*(4), 357–363. https://doi.org/10.1093/mollus/eys023

Schlüter, P. M., Arenas, M. A. S., Harris, S. A., Harris, S. A., Schluter, P. M., & Arenas, M. A. S. (2007). Genetic Variation in Vanilla planifolia (Orchidaceae). *Economic Botany*, *61*(4), 328–336.

Sen, L., Fares, M. A., Liang, B., Gao, L., Wang, B., Wang, T., & Su, Y. J. (2011). Molecular evolution of rbcL in three gymnosperm families: Identifying adaptive and coevolutionary patterns. *Biology Direct*, *6*, 29. https://doi.org/10.1186/1745-6150-6-29

Silva, J. (1995). Birds of the cerrado region, South America. *Streenstrupia*.

Simmonds, S. E., Smith, J. F., Davidson, C., & Buerki, S. (2021). Phylogenetics and comparative plastome genomics of two of the largest genera of angiosperms, Piper and Peperomia (Piperaceae). *Molecular Phylogenetics and Evolution*, *163*, 107229. https://doi.org/10.1016/J.YMPEV.2021.107229

Soto Arenas, M. A., & Cribb, P. (2010). A new infrageneric classification and synopsis of the genus Vanilla Plum. ex mill. (Orchidaceae: Vanillinae). *Lankesteriana*, *9*(3), 355–398. https://doi.org/10.15517/lank.v0i0.12071

Soto Arenas, M. A., & Dressler, R. L. (2010). A revision of the Mexican and Central American species of Vanilla plumier ex miller with a characterization of their its region of the nuclear ribosomal DNA. *Lankesteriana International Journal of Orchidology*, *9*(3).

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stork, N. E. (2018). How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? *Annual Review of Entomology*, *63*(1), 31–45. https://doi.org/10.1146/annurev-ento-020117-043348

Tanksley, S. D., & McCouch, S. R. (1997). Seed banks and molecular maps: Unlocking genetic potential from the wild. *Science*, *277*(5329), 1063–1066. https://doi.org/10.1126/science.277.5329.1063

Thomas, E., van Zonneveld, M., Loo, J., Hodgkin, T., Galluzzi, G., & van Etten, J. (2012). Present Spatial Diversity Patterns of Theobroma cacao L. in the Neotropics Reflect Genetic Differentiation in Pleistocene Refugia Followed by Human-Influenced Dispersal. *PLoS ONE*, *7*(10), e47676. https://doi.org/10.1371/journal.pone.0047676

Trávníček, P., Ponert, J., Urfus, T., Jersáková, J., Vrána, J., Hřibová, E., Doležel, J., & Suda, J. (2015). Challenges of flow-cytometric estimation of nuclear genome size in

orchids, a plant group with both whole-genome and progressively partial endoreplication. *Cytometry Part A*, *87*(10), 958–966. https://doi.org/10.1002/CYTO.A.22681

U.S. FDA. (2022). *CFR - Code of Federal Regulations Title 21*. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=169.3

van Hall, C. J. J. (1914). *Cocoa*. Macmillan and Company, Limited.

Vanilla Genomic Project. (n.d.). Retrieved July 4, 2022, from https://svenbuerki.github.io/VanillaGenomicsCode/index.html

Varela, M., Guerrero, H., & de Miguel, T. (2021). *La sequía que abrasa México, una tragedia predecible y devastadora*. EL PAÍS México. https://elpais.com/mexico/2021-04-24/la-sequia-que-abrasa-mexico-una-tragedia-predecible-y-devastadora.html?prm=enviar_email

Vega, M., Hernandez, M., Herrera-Cabrera, B. E., & Wegier, A. (2017). *Vanilla planifolia. The IUCN Red List of Threatened Species*. E.T103090930A103090933.

Villanueva-Viramontes, S., Hernández-Apolinar, M., Carnevali Fernández-Concha, G., Dorantes-Euán, A., Dzib, G. R., & Martínez-Castillo, J. (2017). Wild Vanilla planifolia and its relatives in the Mexican Yucatan Peninsula: Systematic analyses with ISSR and ITS. *Botanical Sciences*, *95*(2), 169–187. https://doi.org/10.17129/botsci.668

Viruel, J., Kantar, M. B., Gargiulo, R., Hesketh-Prichard, P., Leong, N., Cockel, C., Forest, F., Gravendeel, B., Pérez-Barrales, R., Leitch, I. J., & Wilkin, P. (2020). Crop wild phylorelatives (CWPs): phylogenetic distance, cytogenetic compatibility and breeding system data enable estimation of crop wild relative gene pool

classification. *Botanical Journal of the Linnean Society*, 1–33. https://doi.org/10.1093/botlinnean/boaa064/5903667

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202. https://doi.org/10.1093/BIOINFORMATICS/BTX153

Walther, G.-R., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T. J. C., Fromentin, J.-M., Hoegh-Guldberg, O., & Bairlein, F. (2002). Ecological responses to recent climate change. *Nature*, *416*(6879), 389–395. https://doi.org/10.1038/416389a

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., Venables, B., & Galili, T. (2022). *Package "gplots."* https://cran.r-project.org/web/packages/gplots/index.html

WCS, W. C. S.-, & University, C. for I. E. S. I. N.-C.-C. (2005). *Last of the Wild Project, Version 2, 2005 (LWP-2): Global Human Footprint Dataset (Geographic)*. NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4M61H5F

Wheeler, Q. D. (1999). Why the phylogenetic species concept? - Elementary. *Journal of Nematology*, *31*(2), 134–141.

Whittaker, R. J., Araújo, M. B., Jepson, P., Ladle, R. J., Watson, J. E. M., & Willis, K. J. (2005). Conservation Biogeography: assessment and prospect. *Diversity and Distributions*, *11*(1), 3–23. https://doi.org/10.1111/j.1366-9516.2005.00143.x

Wilke, C. O. (2021). *Ridgeline Plots in "ggplot2" [R package ggridges version 0.5.3]*.

Winter, D. (2020). *pafr: reading, manipulating and plotting genome alignments in the PAF format*. https://cran.r-project.org/web/packages/pafr/vignettes/Introduction_to_pafr.html

Winter, D. J. (2017). rentrez: an R package for the NCBI eUtils API. *The R Journal*, *9*(2), 520–526.

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y., Pang, X., Xu, H., Zhu, Y., Xiao, P., & Chen, S. (2010). Use of ITS2 Region as the Universal DNA Barcode for Plants and Animals. *PLoS ONE*, *5*(10), e13102. https://doi.org/10.1371/journal.pone.0013102

Zhang, H., Meltzer, P., & Davis, S. (2013). RCircos: an R package for Circos 2D track plots. *BMC Bioinformatics*, *14*, 244. https://doi.org/10.1186/1471-2105-14-244

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328. https://doi.org/10.1093/BIOINFORMATICS/BTS606

APPENDIX A

List of dispersers used for GBIF occurrence search.

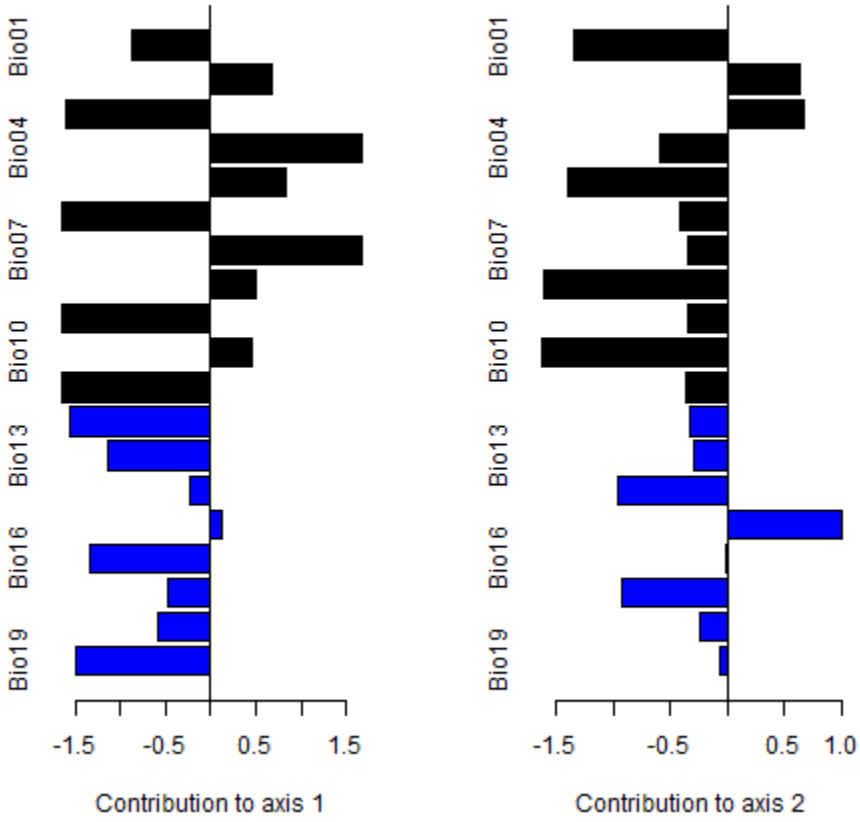
Taxon name:

Micronycteris minuta
Lampronycteris brachyotis
Glyphonycteris sylvestris
Glyphonycteris daviesi
Macrotus waterhousii
Lonchorhina aurita
Lophostoma brasiliense
Lophostoma silvicolum
Phyllostomus discolor
Phyllostomus hastatus
Chropterus auritus
Mimon bennettii
Tonatia saurophila
Tonatia bidens
Phyllostomus elongates
Phyllostomus stenops
Micronycteris megalotis
Micronycteris microtis
Micronycteris brosseti
Glossophaga soricina
Glossophaga leachii
Anoura cultrate
Lichonycteris obscura
Hylonycteris underwoodi
Leptonycteris yerbabuenae
Leptonycteris nivalis
Lonchophylla robusta
Lonchophylla thomasi
Anoura geoffroyi
Scleronycteris ega
Glossophago longiristris
Glossophaga commissarisi
Carollia castanea
Carollia subrufa
Carollia sowelli
Carollia brevicauda
Carollia perspicillata
Rhinophylla pumilio
Rhinophylla fischerae
Sturnira lilium
Sturnira ludovici
Sturnira mordax
Artibeus lituratus
Artibeus jamaicensis
Artibeus toltecas

Artibeus phaeotis
Artibeus watsoni
Enchisthenes hartii
Uroderma bilobatum
Platyrrhinus helleri
Platyrrhinus vittatus
Vampyrodes caraccioli
Chiroderma villosum
Chiroderma salvini
Vampyressa thyone
Ectophylla alba
Ametrida centurio
Centurio senex
Sturnira magna
Sphaeronycteris toxophyllum
Uroderma magnirostrum
Dermanura anderseni
Dermanura glauca
Dermanura gnoma
Artibeus concolor
Artibeus planirostris
Artibeus obscurus
Artibeus amplus
Platyrrhinus infuscus
Platyrrhinus aurarius
Platyrrhinus lineatus
Platyrrhinus fusciventris
Platyrrhinus incarum
Platyrrhinus brachycephalus
Vampyriscus bidens
Vampyriscus brocki

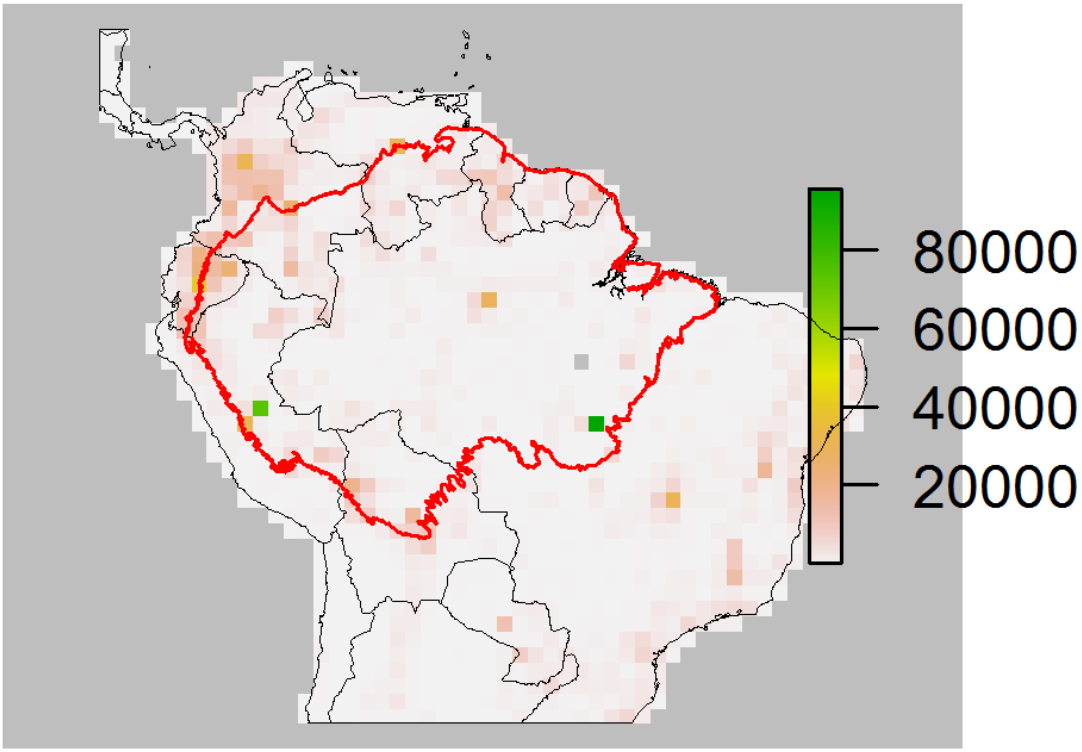
APPENDIX B

Barplots of climatic variable contribution to each axis (PC1 and PC2) withing the principal component analysis.


Bio01
Bio04
Bio07
Bio10
Bio13
Bio16
Bio19
-1.5
-0.5
0.5
1.5
Contribution to axis 1
Bio01
Bio04
Bio07
Bio10
Bio13
Bio16
Bio19
-1.5
-0.5
0.5
1.0
Contribution to axis 2

APPENDIX C

Specimen richness map of plant occurrences from GBIF illustrates a gap of occurrence data in the Amazon Basin, which is outlined in red.


80000
60000
40000
20000

APPENDIX D

Accessions used in phylogenetic and genetic distance analyses.

Voucher id	Voucher source	ex/in-situ	Specimen id	*rbcL* Genbank id	Its Genbank id	BioSample
PE10	HEM/UNICACH Living Collection	in-situ	MEX10	ON531941	ON525213	na
PE11	HEM/UNICACH Living Collection	in-situ	MEX11	ON531940	ON525166	na
PE12	HEM/UNICACH Living Collection	in-situ	MEX12	ON531942	ON525182	SAMN28632720
PE13	HEM/UNICACH Living Collection	in-situ	MEX13	ON531943	ON525174	SAMN28632721
PE14	HEM/UNICACH Living Collection	in-situ	MEX14	ON531939	ON525175	SAMN28632722
PE18	HEM/UNICACH Living Collection	in-situ	MEX18	na	ON525176	na
PE19	HEM/UNICACH Living Collection	in-situ	MEX19	ON531976	ON525177	SAMN28632723
PE20	HEM/UNICACH Living Collection	in-situ	MEX20	ON531955	ON525180	SAMN28632724
PE21	HEM/UNICACH Living Collection	in-situ	MEX21	ON531956	ON525195	na
PE22	HEM/UNICACH Living Collection	in-situ	MEX22	ON531957	ON525199	na
PE23	HEM/UNICACH Living Collection	in-situ	MEX23	ON531975	ON525200	na
PE24	HEM/UNICACH Living Collection	in-situ	MEX24	ON531974	ON525196	na
PE25	HEM/UNICACH	in-situ	MEX25	ON531973	ON525197	na

	Living Collection					
PE26	HEM/UNICACH Living Collection	in-situ	MEX26	ON531958	ON525187	SAMN28632725
PE27	HEM/UNICACH Living Collection	in-situ	MEX27	ON531925	ON525198	na
PE28	HEM/UNICACH Living Collection	in-situ	MEX28	ON531972	ON525228	na
PE29	HEM/UNICACH Living Collection	in-situ	MEX29	na	ON525189	na
PE30	HEM/UNICACH Living Collection	in-situ	MEX30	ON531954	ON525194	na
PE31	HEM/UNICACH Living Collection	in-situ	MEX31	ON531922	ON525191	SAMN28632726
PE36	HEM/UNICACH Living Collection	in-situ	MEX36	ON531982	ON525185	SAMN28632727
PE38	HEM/UNICACH Living Collection	in-situ	MEX38	ON531984	ON525204	na
PE39	HEM/UNICACH Living Collection	in-situ	MEX39	ON531959	ON525190	na
PE40	HEM/UNICACH Living Collection	in-situ	MEX40	ON531952	ON525192	na
PE41	HEM/UNICACH Living Collection	in-situ	MEX41	ON531960	ON525193	SAMN28632728
PE42	HEM/UNICACH Living Collection	in-situ	MEX42	ON531929	ON525163	na
PE44	HEM/UNICACH Living Collection	in-situ	MEX44	ON531932	ON525162	na
PE45	HEM/UNI	in-	MEX45	ON531928	ON525	na

	CACH Living Collection	situ			164	
PE46	HEM/UNI CACH Living Collection	in-situ	MEX46	ON531938	ON525214	na
PE47	HEM/UNI CACH Living Collection	in-situ	MEX47	ON531937	ON525215	na
PE48	HEM/UNI CACH Living Collection	in-situ	MEX48	ON531945	ON525217	na
PE49	HEM/UNI CACH Living Collection	in-situ	MEX49	ON531933	ON525168	na
PE5	HEM/UNI CACH Living Collection	in-situ	MEX5	ON531936	ON525227	na
PE50	HEM/UNI CACH Living Collection	in-situ	MEX50	ON531926	ON525171	na
PE51	HEM/UNI CACH Living Collection	in-situ	MEX51	ON531935	ON525186	SAMN28632729
PE52	HEM/UNI CACH Living Collection	in-situ	MEX52	ON531934	ON525207	na
PE53	HEM/UNI CACH Living Collection	in-situ	MEX53	ON531927	ON525172	na
PE54	HEM/UNI CACH Living Collection	in-situ	MEX54	na	ON525216	na
PE55	HEM/UNI CACH Living Collection	in-situ	MEX55	ON531953	ON525201	na
PE56	HEM/UNI CACH Living Collection	in-situ	MEX56	ON531971	ON525205	na
PE59	HEM/UNI CACH Living Collection	in-situ	MEX59	ON531961	ON525206	SAMN28632730

PE6	HEM/UNICACH Living Collection	in-situ	MEX6	ON531921	ON525219	SAMN28632719
PE61	HEM/UNICACH Living Collection	in-situ	MEX61	na	ON525170	na
PE65	HEM/UNICACH Living Collection	in-situ	MEX65	ON531970	ON525222	SAMN28632731
PE66	HEM/UNICACH Living Collection	in-situ	MEX66	ON531969	ON525202	na
PE67	HEM/UNICACH Living Collection	in-situ	MEX67	ON531919	ON525178	SAMN28632732
PE69	HEM/UNICACH Living Collection	in-situ	MEX69	ON531962	ON525203	SAMN28632733
PE7	HEM/UNICACH Living Collection	in-situ	MEX7	ON531963	ON525218	na
PE70	HEM/UNICACH Living Collection	in-situ	MEX70	ON531968	ON525221	na
PE72	HEM/UNICACH Living Collection	in-situ	MEX72	ON531967	ON525220	na
PE73	HEM/UNICACH Living Collection	in-situ	MEX73	ON531946	ON525173	na
PE74	HEM/UNICACH Living Collection	in-situ	MEX74	ON531930	ON525165	na
PE75	HEM/UNICACH Living Collection	in-situ	MEX75	ON531966	ON525223	na
PE76	HEM/UNICACH Living Collection	in-situ	MEX76	ON531977	ON525211	na
PE77	HEM/UNICACH Living	in-situ	MEX77	ON531947	ON525224	na

	Collection					
PE79	HEM/UNICACH Living Collection	in-situ	MEX79	ON531964	ON525225	SAMN28632734
PE8	HEM/UNICACH Living Collection	in-situ	MEX8	ON531931	ON525167	na
PE80	HEM/UNICACH Living Collection	in-situ	MEX80	ON531918	ON525179	na
PE9	HEM/UNICACH Living Collection	in-situ	MEX9	ON531944	ON525210	na
PE81	SNP/BSU Living Collection	ex-situ	BSU81_Vanilla_planifolia	ON531948	ON525188	na
PE82	SNP/BSU Living Collection	ex-situ	BSU82_Vanilla_planifolia	ON531978	ON525208	na
PE83	SNP/BSU Living Collection	ex-situ	BSU83_Vanilla_planifolia	ON531979	ON525212	na
PE84	SNP/BSU Living Collection	ex-situ	BSU84_Vanilla_planifolia	ON531980	ON525209	na
PE85	SNP/BSU Living Collection	ex-situ	BSU85_Vanilla_planifolia_var__albomarginata	ON531981	ON525226	na
1982-3727	MO/MBG Living Collection	ex-situ	MO86_Vanilla_pompona_x_odorata	ON531923	na	na
2012-0026	MO/MBG Living Collection	ex-situ	MO87_Vanilla_planifolia_var_variegata	ON531949	ON525183	na
1990-2546	MO/MBG Living Collection	ex-situ	MO88_Vanilla_palmarum	ON531986	ON525161	na
2001-1086	MO/MBG Living Collection	ex-situ	MO89_Vanilla_insignis	ON531917	na	na
2015-0108	MO/MBG Living Collection	ex-situ	MO90_Vanilla_planifolia	ON531950	ON525181	na
1992-0683	MO/MBG Living Collection	ex-situ	MO91_Vanilla_schwackeana	ON531985	ON525169	na
2019-0335	MO/MBG Living Collection	ex-situ	MO92_Vanilla_planifolia	ON531951	ON525184	na
1979-1069	MO/MBG Living	ex-situ	MO93_Vanilla_planifolia_var_albo	ON531924	na	na

	Collection					
1989-2217	MO/MBG Living Collection	ex-situ	MO94_Vanilla_tahitensis	ON531920	na	na
na	na	ex-situ	GB1373448127_Vanilla_somae	na	1.37E+09	na
na	na	ex-situ	GB1517415473_Vanilla_planifolia_x_pompona	na	1.52E+09	na
na	na	ex-situ	GB1517415474_Vanilla_planifolia_x_pompona	na	1.52E+09	na
na	na	ex-situ	GB1517415475_Vanilla_planifolia_x_pompona	na	1.52E+09	na
na	na	ex-situ	GB1546056_Vanilla_planifolia	na	1546056	na
na	na	ex-situ	GB170284137_Vanilla_bahiana	na	1.7E+08	na
na	na	ex-situ	GB170284138_Vanilla_pompona	na	1.7E+08	na
na	na	ex-situ	GB170284139_Vanilla_edwallii	na	1.7E+08	na
na	na	ex-situ	GB1708599396_Vanilla_madagascariensis	na	1.71E+09	na
na	na	ex-situ	GB1708599397_Vanilla_planifolia	na	1.71E+09	na
na	na	ex-situ	GB1717501702_Vanilla_roscheri	na	1.72E+09	na
na	na	ex-situ	GB1733398303_Vanilla_borneensis	na	1.73E+09	na
na	na	ex-situ	GB1733398304_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398305_Vanilla_griffithii	na	1.73E+08	na
na	na	ex-situ	GB1733398306_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398308_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398309_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398310_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398311_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398312_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398313_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398314_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398315_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398316_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398317_Vanilla_griffithii	na	1.73E+09	na

na	na	ex-situ	GB1733398318_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398319_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398320_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398321_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398322_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398323_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398324_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398325_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398326_Vanilla_griffithii	na	1.73E+09	na
na	na	ex-situ	GB1733398327_Vanilla_kinabaluensis	na	1.73E+09	na
na	na	ex-situ	GB1733398329_Vanilla_kinabaluensis	na	1.73E+09	na
na	na	ex-situ	GB1733398330_Vanilla_kinabaluensis	na	1.73E+09	na
na	na	ex-situ	GB1733398331_Vanilla_kinabaluensis	na	1.73E+09	na
na	na	ex-situ	GB1733398332_Vanilla_kinabaluensis	na	1.73E+09	na
na	na	ex-situ	GB1789804096_Vanilla_barbellata	na	1.79E+09	na
na	na	ex-situ	GB1789804097_Vanilla_calyculata	na	1.79E+09	na
na	na	ex-situ	GB1789804098_Vanilla_calyculata	na	1.79E+09	na
na	na	ex-situ	GB1789804099_Vanilla_calyculata	na	1.79E+09	na
na	na	ex-situ	GB1789804100_Vanilla_calyculata	na	1.79E+09	na
na	na	ex-situ	GB1789804101_Vanilla_calyculata	na	1.79E+09	na
na	na	ex-situ	GB1789804102_Vanilla_claviculata	na	1.79E+09	na
na	na	ex-situ	GB1789804103_Vanilla_cribbiana	na	1.79E+09	na
na	na	ex-situ	GB1789804104_Vanilla_cribbiana	na	1.79E+09	na
na	na	ex-situ	GB1789804105_Vanilla_cribbiana	na	1.79E+09	na
na	na	ex-situ	GB1789804106_Vanilla_dressleri	na	1.79E+09	na
na	na	ex-situ	GB1789804107_Vanilla_dressleri	na	1.79E+09	na
na	na	ex-situ	GB1789804108_Vanilla_dressleri	na	1.79E+09	na

na	na	ex-situ	GB1789804109_Vanilla_dressleri	na	1.79E+09	na
na	na	ex-situ	GB1789804110_Vanilla_hartii	na	1.79E+09	na
na	na	ex-situ	GB1789804111_Vanilla_hartii	na	1.79E+09	na
na	na	ex-situ	GB1789804112_Vanilla_hartii	na	1.79E+09	na
na	na	ex-situ	GB1789804113_Vanilla_helleri	na	1.79E+09	na
na	na	ex-situ	GB1789804114_Vanilla_helleri	na	1.79E+09	na
na	na	ex-situ	GB1789804116_Vanilla_inodora	na	1.79E+09	na
na	na	ex-situ	GB1789804117_Vanilla_inodora	na	1.79E+09	na
na	na	ex-situ	GB1789804118_Vanilla_insignis	na	1.79E+09	na
na	na	ex-situ	GB1789804119_Vanilla_insignis	na	1.79E+09	na
na	na	ex-situ	GB1789804120_Vanilla_insignis	na	1.79E+09	na
na	na	ex-situ	GB1789804122_Vanilla_insignis	na	1.79E+09	na
na	na	ex-situ	GB1789804124_Vanilla_odorata	na	1.79E+09	na
na	na	ex-situ	GB1789804125_Vanilla_odorata	na	1.79E+09	na
na	na	ex-situ	GB1789804126_Vanilla_odorata	na	1.79E+09	na
na	na	ex-situ	GB1789804127_Vanilla_odorata	na	1.79E+09	na
na	na	ex-situ	GB1789804128_Vanilla_odorata	na	1.79E+09	na
na	na	ex-situ	GB1789804130_Vanilla_phaeantha	na	1.79E+09	na
na	na	ex-situ	GB1789804131_Vanilla_phaeantha	na	1.79E+09	na
na	na	ex-situ	GB1789804133_Vanilla_phaeantha	na	1.79E+09	na
na	na	ex-situ	GB1789804134_Vanilla_planifolia	na	1.79E+09	na
na	na	ex-situ	GB1789804135_Vanilla_planifolia	na	1.79E+09	na
na	na	ex-situ	GB1789804136_Vanilla_planifolia	na	1.79E+09	na
na	na	ex-situ	GB1789804137_Vanilla_planifolia	na	1.79E+09	na
na	na	ex-situ	GB1789804138_Vanilla_planifolia	na	1.79E+09	na
na	na	ex-situ	GB1789804139_Vanilla_planifolia	na	1.79E+09	na
na	na	ex-situ	GB1789804140_Vanilla_planifolia	na	1.79E+09	na

na	na	ex-situ	GB1789804142_Vanilla_planifolia	na	1.79E+09	na
na	na	ex-situ	GB1789804143_Vanilla_pompona_subsp__grandiflora	na	1.79E+09	na
na	na	ex-situ	GB1789804144_Vanilla_pompona_subsp__grandiflora	na	1.79E+09	na
na	na	ex-situ	GB1789804145_Vanilla_pompona_subsp__grandiflora	na	1.79E+09	na
na	na	ex-situ	GB1789804146_Vanilla_pompona_subsp__grandiflora	na	1.79E+09	na
na	na	ex-situ	GB1789804147_Vanilla_pompona_subsp__grandiflora	na	1.79E+09	na
na	na	ex-situ	GB1789804148_Vanilla_pompona_subsp__grandiflora	na	1.79E+09	na
na	na	ex-situ	GB1789804149_Vanilla_pompona_subsp__pompona	na	1.79E+09	na
na	na	ex-situ	GB1789804150_Vanilla_pompona_subsp__pompona	na	1.79E+09	na
na	na	ex-situ	GB1789804152_Vanilla_pompona	na	1.79E+09	na
na	na	ex-situ	GB1789804153_Vanilla_pompona_subsp__pompona	na	1.79E+09	na
na	na	ex-situ	GB1789804154_Vanilla_pompona_subsp__grandiflora	na	1.79E+09	na
na	na	ex-situ	GB1789804155_Vanilla_pompona_subsp__pittieri	na	1.79E+09	na
na	na	ex-situ	GB1789804156_Vanilla_pompona_subsp__pittieri	na	1.79E+09	na
na	na	ex-situ	GB1789804157_Vanilla_pompona_subsp__pompona	na	1.79E+09	na
na	na	ex-situ	GB1789804158_Vanilla_pompona_subsp__pittieri	na	1.79E+09	na
na	na	ex-situ	GB1789804160_Vanilla_tahitensis	na	1.79E+09	na
na	na	ex-situ	GB1789804161_Vanilla_trigonocarpa	na	1.79E+09	na
na	na	ex-situ	GB1789804162_Vanilla_trigonocarpa	na	1.79E+09	na
na	na	ex-situ	GB1789804163_Vanilla_trigonocarpa	na	1.79E+09	na
na	na	ex-situ	GB219964414_Vanilla_imperialis	na	2.2E+08	na
na	na	ex-situ	GB219964418_Vanilla_africana	na	2.2E+08	na
na	na	ex-situ	GB219964419_Vanilla_barbellata	na	2.2E+08	na
na	na	ex-situ	GB219964424_Vanilla_roscheri	na	2.2E+08	na
na	na	ex-situ	GB219964425_Erythrorchis_cassythoides	na	2.2E+08	na
na	na	ex-situ	GB24397239_Vanilla_tahitensis	na	24397239	na
na	na	ex-situ	GB24397240_Vanilla_planifolia	na	24397240	na

na	na	ex-situ	GB260401086_Vanilla_pompona	na	2.6E+08	na
na	na	ex-situ	GB260401087_Vanilla_pompona	na	2.6E+08	na
na	na	ex-situ	GB260401088_Vanilla_pompona	na	2.6E+08	na
na	na	ex-situ	GB260401089_Vanilla_pompona	na	2.6E+08	na
na	na	ex-situ	GB260401090_Vanilla_pompona	na	2.6E+08	na
na	na	ex-situ	GB260401091_Vanilla_planifolia	na	2.6E+08	na
na	na	ex-situ	GB260401092_Vanilla_planifolia	na	2.6E+08	na
na	na	ex-situ	GB260401093_Vanilla_planifolia	na	2.6E+08	na
na	na	ex-situ	GB260401094_Vanilla_planifolia	na	2.6E+08	na
na	na	ex-situ	GB260401095_Vanilla_planifolia	na	2.6E+08	na
na	na	ex-situ	GB260401096_Vanilla_planifolia	na	2.6E+08	na
na	na	ex-situ	GB260401097_Vanilla_planifolia	na	2.6E+08	na
na	na	ex-situ	GB260401098_Vanilla_planifolia	na	2.6E+08	na
na	na	ex-situ	GB347602087_Vanilla_siamensis	na	3.48E+08	na
na	na	ex-situ	GB350999064_Vanilla_somae	na	3.51E+08	na
na	na	ex-situ	GB6652551_Vanilla_planifolia	na	6652551	na
na	na	ex-situ	GB6690511_Vanilla_aphylla	na	6690511	na
na	na	ex-situ	GB1005352026_Vanilla_roscheri	1.01E+09	na	na
na	na	ex-situ	GB1243024771_Vanilla_planifolia	1.24E+09	na	na
na	na	ex-situ	GB170284066_Vanilla_bahiana	1.7E+08	na	na
na	na	ex-situ	GB170284068_Vanilla_edwallii	1.7E+08	na	na
na	na	ex-situ	GB257146093_Vanilla_palmarum	2.57E+08	na	na
na	na	ex-situ	GB257146095_Vanilla_humblotii	2.57E+08	na	na
na	na	ex-situ	GB257146097_Vanilla_albida	2.57E+08	na	na
na	na	ex-situ	GB257146099_Vanilla_planifolia_x_tahitensis	2.57E+08	na	na
na	na	ex-situ	GB257146107_Vanilla_dilloniana	2.57E+08	na	na
na	na	ex-situ	GB257146109_Vanilla_bahiana	2.57E+08	na	na

na	na	ex-situ	GB257146115_Vanilla_odorata	2.57E+08	na	na
na	na	ex-situ	GB257146117_Vanilla_bahiana	2.57E+08	na	na
na	na	ex-situ	GB257146123_Vanilla_africana	2.57E+08	na	na
na	na	ex-situ	GB257146125_Vanilla_imperialis	2.57E+08	na	na
na	na	ex-situ	GB257146127_Vanilla_leprieurii	2.57E+08	na	na
na	na	ex-situ	GB257146131_Vanilla_odorata	2.57E+08	na	na
na	na	ex-situ	GB257146135_Vanilla_madagascariensis	2.57E+08	na	na
na	na	ex-situ	GB257146137_Vanilla_phalaenopsis	2.57E+08	na	na
na	na	ex-situ	GB257146141_Vanilla_tahitensis	2.57E+08	na	na
na	na	ex-situ	GB257146145_Vanilla_pompona	2.57E+08	na	na
na	na	ex-situ	GB257146147_Vanilla_leprieurii	2.57E+08	na	na
na	na	ex-situ	GB257146149_Vanilla_ensifolia	2.57E+08	na	na
na	na	ex-situ	GB257146155_Vanilla_ensifolia	2.57E+08	na	na
na	na	ex-situ	GB257146157_Vanilla_planifolia	2.57E+08	na	na
na	na	ex-situ	GB257146161_Vanilla_chamissonis	2.57E+08	na	na
na	na	ex-situ	GB257146163_Vanilla_lindmaniana	2.57E+08	na	na
na	na	ex-situ	GB257146169_Vanilla_odorata	2.57E+08	na	na
na	na	ex-situ	GB257146181_Vanilla_dilloniana	2.57E+08	na	na
na	na	ex-situ	GB353444828_Vanilla_planifolia	3.53E+08	na	na
na	na	ex-situ	GB3560719_Erythrorchis_cassythoides	3560719	na	na
na	na	ex-situ	GB3560857_Vanilla_aphylla	3560857	na	na
na	na	ex-situ	GB3560859_Vanilla_africana	3560859	na	na
na	na	ex-situ	GB3560863_Vanilla_imperialis	3560863	na	na
na	na	ex-situ	GB3560867_Vanilla_roscheri	3560867	na	na
na	na	ex-situ	GB380749781_Vanilla_planifolia	3.81E+08	na	na
na	na	ex-situ	GB387315781_Vanilla_planifolia	3.87E+08	na	na
na	na	ex-situ	GB39655296_Vanilla_inodora	39655296	na	na

na	na	ex-situ	GB39655298_Vanilla_palmarum	39655298	na	na

APPENDIX E

Complete Bayesian phylogenetic tree inferred using rbcL.

This supplementary file can be found at:

https://bsapubs.onlinelibrary.wiley.com/doi/10.1002/ajb2.16024

APPENDIX F

Complete maximum likelihood phylogenetic tree inferred using rbcL.

This supplementary file can be found at:

https://bsapubs.onlinelibrary.wiley.com/doi/10.1002/ajb2.16024

APPENDIX G

Complete Bayesian phylogenetic tree inferred using ITS.

This supplementary file can be found at:

https://bsapubs.onlinelibrary.wiley.com/doi/10.1002/ajb2.16024

APPENDIX H

Complete maximum likelihood phylogenetic tree inferred using ITS.

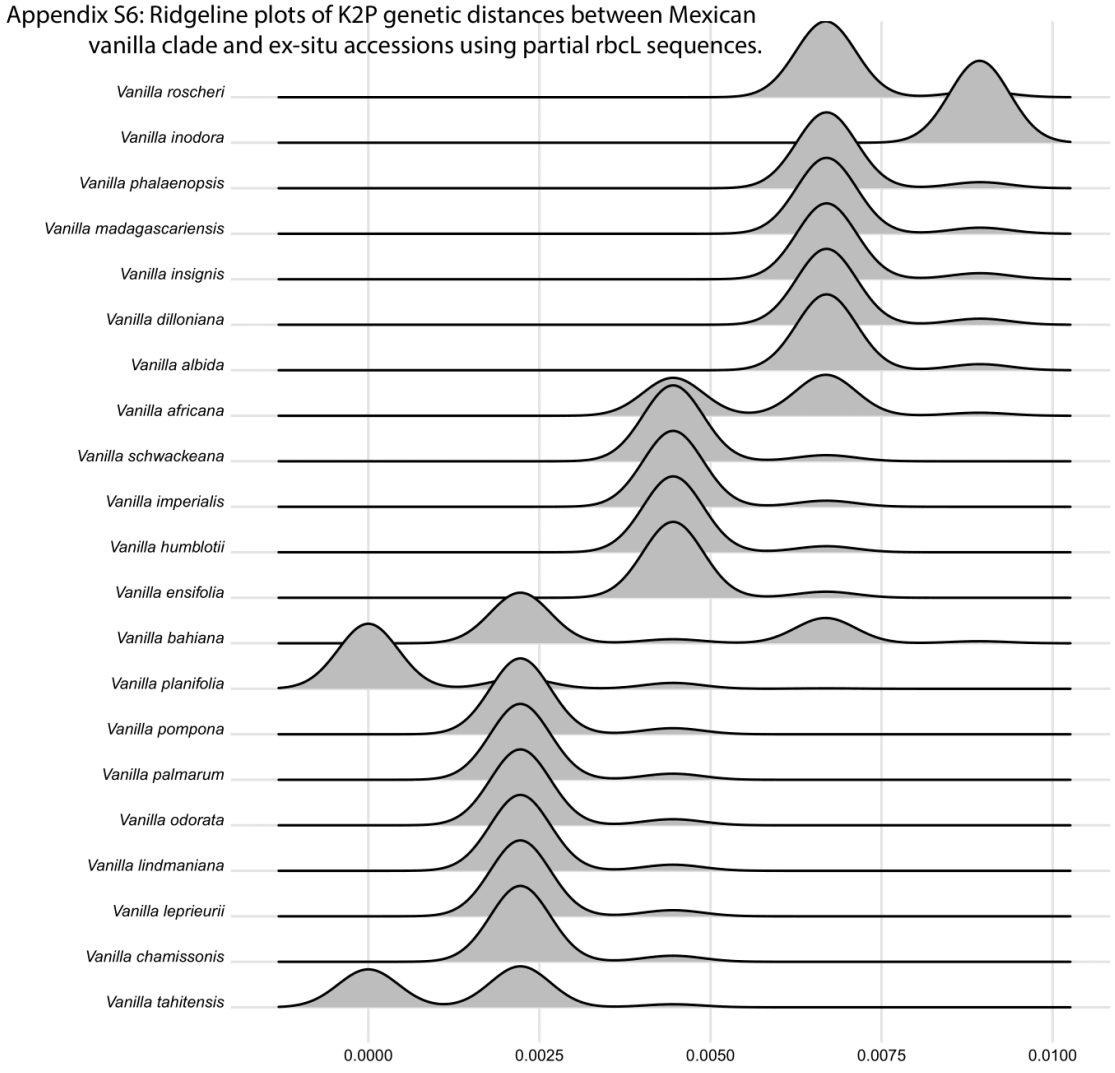
This supplementary file can be found at:

https://bsapubs.onlinelibrary.wiley.com/doi/10.1002/ajb2.16024

APPENDIX I

Ridgeline plots of K2P genetic distances between Mexican vanilla clade and ex-situ accessions using partial rbcL sequences.

Appendix S6: Ridgeline plots of K2P genetic distances between Mexican vanilla clade and ex-situ accessions using partial rbcL sequences.



APPENDIX J

Contributions of each variable to the first two principal components in PCA of 19 bioclimatic variables.

Appendix S7: Contributions of each variable to the first two principal components in PCA of 19 bioclimatic variables.