WILD AVOCADO (*PERSEA AMERICANA)* HERBARIUM GENOMES PROVIDE KEY INSIGHTS INTO ITS WILD POPULATION STRUCTURE AND DOMESTICATION HISTORY

A Thesis

by

KEVIN W. WANN

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

Chair of Committee,	Heather Thakar
Committee Members,	Michael Waters
	Allison Hopkins
	Rodolfo Aramayo
Head of Department,	Darryl De Ruiter

May 2023

Major Subject: Anthropology

ABSTRACT

Avocados (*Persea americana*) are highly nutritious fruits that dominate the global export market and have an extensive genomics research background. They have a complex domestication history with some disagreement on the origins of the three common cultivar varieties (var. *drymifolia*, var *guatemalensis*, and var. *americana*), and most studies have not comprehensively examined the germplasm of wild populations. Our objectives for this study were to better understand how wild avocado populations are structured in the absence of human interference and to assign geographic regions to the origins of domesticated varieties. We sequenced at low coverage the genomes of 25 putatively wild herbarium avocado leaves collected in the last 60 years and spanning their entire native geographic range. We used bioinformatic analyses that examine genotype likelihoods to compare and contrast the population structure of our wild avocados with that of a previously published cultivar dataset. Wild avocados are most likely structured in two distinct populations, one in Central Mexico and one spanning from Chiapas to as far as Peru, and we predict the valley between the Sierra Madre del Sur and the Sierra Madre de Chiapas acts as a reproductive barrier. Overall, wild avocado populations are more genetically differentiated and more diverse compared to cultivars. We attribute the difference to the domestication process which acts to erode genetic variation over time and then reduce differences between varieties through commercial hybridization. In regards to which herbarium specimens have higher genetic affinity to the three common cultivar varieties, our findings support claims that each variety has distinct and separate domestication origin throughout Central America. We also offer a new model fitting our data that includes a single domestication event in Honduras that gives rise to both the var. *guatemalensis* and var. *americana*. We encourage more research including the genomes of ancient specimens to help support or refute this scenario.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Thakar, my committee members, Drs. Kistler, Waters, Hopkins, and Aramayo, and my lab advisor Dr. Hofman for their guidance and support on this project.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of Drs. Heather Thakar, Michael Waters, and Allison Hopkins of the Anthropology Department, Dr. Rodolfo Aramayo of the Biology Department, and Dr. Logan Kistler of the Smithsonian Institution. Dr. Courtney Hofman supervised the lab work portion of this project. Data interpretation was especially aided by Drs. Thakar and Kistler. All lab work and bioinformatic analyses were performed by the author.

Funding Sources

This project was funded by the Charles H. Gregory, class of ’64 Graduate Strategic Fellowship, Department of Liberal Arts, Texas A&M University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

The avocado (*Persea americana* Miller) is a perennial tree species in the Lauraceae family, and it is one of the most popular fruit crops in modern diets. The leading producer and exporter of avocado fruits is Mexico, harvesting nearly 2.5 million tons and exporting over 2.7 billion dollars' worth of fruits in 2021. Globally, avocado fruit production totaled 8.7 million metric tons in 2021, over three times the worldwide harvested volume two decades prior (Shabandeh 2022a; Shabandeh 2022b; Statista Research Department 2022). Their fruits are high in fiber and healthy lipids, contain little sodium and carbohydrates, and have zero cholesterol. They are also beneficial sources of Vitamins B, C, E, and K, magnesium, carotenoids, and potassium, with one avocado containing more two whole bananas (Bhuyan et al. 2019; Harvard School of Public Health).

1.1 Avocado domestication and history

P. americana, while now cultivated worldwide, is native in its wild form to the American tropics, with a range from Central Mexico to The Andes Valley (Dillehay et al. 2017). The species diverged from other *Persea* clades in North America sometime during the Late Miocene to Middle Pliocene, and it then migrated southward to its modern wild range in the early the Pleistocene (Galindo-Tovar et al. 2008, Rendón-Anaya et al. 2019). Prior to the first human migrations into Central and South America, the main dispersal agent of medium- and large-sized fruits were Pleistocene megafauna. The same is true for *P. americana*, dispersed via consumption by the giant ground sloth (Barlow 2002). This places the avocado in an extensive list of New World domesticated plants originally adapted to hybridization via zoochory and left without a proper dispersal agent following megafaunal extinction (Janzen and Martin 1982). The abandonment of their dispersal agent probably restricted the area of wild avocado stands and limited hybridization among them. Had humans not populated the Americas shortly after and established themselves as the new consumer of avocados, this tree species may have undergone a dramatic decrease in fruit size or become extinct altogether (Kistler et al. 2015; Spengler et al. 2021).

Archaeobotanical seed remains suggest that the earliest known wild avocado consumption dates to the Paleoindian Period (11-9ka), and it occurs throughout Central and South America. At the El Gigante Rockshelter, Honduras, archaeologists recovered avocado remains from early Paleoindian strata, and at Huaca Prieta, Peru, avocado seeds were directly dated to 10.5ka (Dillehay et al. 2017; Kennett et al. 2017). Consumption may have also occurred in Mexico by this time as well, although the earliest evidence for avocado use in Mexico dates between 10 and 9ka at Coaxcatlan Cave, Puebla (Smith 1966). Until the beginning of the Archaic Period (9ka), humans only foraged for wild avocados alongside a suite of other more calorically dense plant and animal taxa (Betz 1999). Between 9 and 6ka as temperature and precipitation began to increase, high-energy meat and grains became scarcer, leading to an increase in the exploitation and management of wild avocado trees to sustain diets (Buckler et al. 1998, Mac Neish 1964). After 6ka, Mexican and Central American climates became dryer, threatening wild *P. americana* populations. At this time, further management and exploitation increased as humans begin to cultivate avocado trees in forest and home gardens (Buckler et al. 1998; Weirsum 1997). It is at this point that the archaeological record indicates the avocado's protracted domestication via a gradual increase in seed size, used as a proxy for fruit size (Fuller 2018; Scheffler 2008; Smith 1969) *P. americana* was a primary domesticated fruit during the time of the Maya and later, playing a key role in both diet and religion (Galindo-Tovar 2008).

Mayan mythology interprets trees as symbols of rebirth, including avocados. For example, the sarcophagus of king *Hanab-Pakal* depicts one figure rising from the earth alongside an avocado tree (Schele 1974).

Botanists classify modern domesticated avocados into three distinct varieties or horticultural races based on morphology, preferred environment, probable origin, and genetics. The Mexican race (var. *drymifolia*) originates from the highlands of central Mexico, has the highest cold tolerance, and has the smallest fruits with soft purple skin. The Guatemalan race (var. *guatemalensis*) is thought to originate from the highlands of Guatemala, has moderate cold tolerance, and has moderate fruit size with thick, rough, green skin. The Lowland race (var. *americana*) is the only truly tropical variety, with less agreement as to its origin, and has thick lime green skin and lower oil but higher sugar content (Chanderbali et al. 2013). Archaeologists and geneticists argue that var. *americana* could have arisen from the Atlantic Coast of Yucatán, the Pacific Coast of Guatemala, or between Nicaragua and Panama (Chen et al. 2009; Galindo-Tovar et al. 2010; Storey et al. 1986). The majority of important commercial cultivars are either belong to the Guatemalan group or are a hybrid between the Guatemalan and the Mexican group. The Hass cultivar is particularly popular and accounts for 90% of worldwide consumption (Rendon-Anaya et al. 2019). This cultivar represents a Guatemalan x Mexican hybrid that combines the Guatemalan’s flesh content with the Mexican’s environmental tolerance (Chen et al. 2009; Rendón-Anaya et al. 2019).

Numerous avocado genomic studies implementing a variety of molecular markers have 1) further delineated the three varieties based on genetic differences, 2) identified the varietal/hybrid origin of unknown cultivars for improvement in breeding, 3) uncovered broader patterns of Lauraceae and angiosperm evolution, and 4) supported the protracted history of isolation among the three regions of cultivar origin prior to Spanish Conquest (e.g., Ashworth et al. 2004; Cañas-Gutiérrez et al. 2015; Chanderbali et al. 2008; Chen et al. 2008, 2009; Furnier et al. 1990; Guzmán et al. 2017; Rendón-Anaya et al. 2019; Rubinstein et al. 2019; Sharon et al. 1997; Solares et al. 2022; Song et al. 2016; Talavera et al. 2019). These studies either focused completely on the germplasm of cultivar lineages or examined the population structure of avocados with low sample size or geographic resolution among wild samples. Rendón-Anaya et al. (2019) and Solares et al. (2022) analyzed the genomes of four putative wild trees collected from Chiapas and Costa Rica. They found that these trees were phylogenetically separate from domesticated clades, and those from Chiapas clustered very closely to the Guatemalan variety. Chen et al. (2008) measured the genetic diversity and linkage disequilibrium values of 21 putatively wild avocado trees. However, only 8 of the samples originate outside Mexico, only four countries were sampled, and only four genomic loci were analyzed. In this study, we homogenously sampled the whole geographic range of wild *P. americana* through herbarium leaf genomes to broaden our understanding of its population structure and pinpoint regions of domestication origin for each variety.

1.2 Herbarium specimen genomes

The genomes of historic leaf specimens housed in herbaria offer an alternate method to accessing plant diversity that could otherwise be impossible to capture. As of 2017, there are around 350 million historic plant samples secured in the world's 3400 herbaria, and the rate at which their images and associated metadata are becoming digitized is increasing rapidly (Soltis 2017; Tulig et al 2012). Such metadata typically includes the date of collection, location (either coordinate or relative), elevation, local environment, and any botanical characteristics the collector deems important. The DNA of these leaves can be used as representatives of their species with respect to a given time, location, phenotype, or historical disease (if present on the specimen). Analyzing herbarium genomes can also increase the feasibility of this type of study by decreasing the cost of acquiring specimens. For example, Simon et al. (2022) outlined the phylogeny, biogeography, and admixture among *Manihot* species in South America using mostly herbarium specimens. Through low-coverage shotgun sequencing, they used the known collection locations of each sample to cluster the species based on both ancestry and habitat. Konrade et al. (2019) used microsatellite loci to genotype over 500 herbarium *Prunus serotina* leaves covering the species' entire geographic range of the eastern United States. They found moderate population structure and weak levels of isolation by distance, both attributed to intense and distant gene flow among trees.

While herbaria may promote the preservation of leaf tissue better than tropical archaeological sediments, prolonged exposure to heat treatments and other common preservatives still damage DNA over time (Gutaker and Burbano 2017; Weiß et al. 2016). Therefore, it may be necessary to treat historic herbarium leaves, particularly older ones, with the same delicacy as archaeological specimens (i.e., all extraction steps carried out in a designated clean lab and a protocol optimized for retaining ultrashort DNA fragments). Marinček et al. (2022) compared the efficacy of two extraction protocols on a subset of herbarium specimens in a clean lab setting. They determined that a PTB-DTT-based method better suited for degraded tissue was necessary to produce yields sufficient for whole-genome sequencing for older leaves. The other protocol, a Qiagen kit used for modern leaf tissue with slight modification, produced comparable yields to the stringent method when applied to more recent (<60 years old) samples. They also observed a slight negative correlation between the yield and age of specimens. In Simon et al.'s (2022) study of *Manihot* herbarium genomes, only recent leaves were utilized, and they implemented a similarly modified kit with success. These studies did not experiment with different bioinformatic pipelines used for modern or ancient DNA. Sequence reads are normally mapped to a reference genome using default alignment parameters for modern samples, but archaeogenomes require a more relaxed alignment on the ends of reads to account for nucleotide misincorporations caused by chemical damage. Through this project, we sought to refine the bioinformatic processing of herbarium specimen genomes.

We report here the first attempted DNA extraction from herbarium avocado leaves and the first broad-range sampling of wild avocado populations. By comparing the genomes and population structure of our herbarium genomes to that of commercial cultivars, we sought to more accurately outline the avocado's domestication history and support or refute previous assumptions of cultivar origin. Uncovering wild germplasm is important to maintaining healthy levels of genetic variation to combat the harmful effects of climate change and the susceptibility of relying on few genomes for the majority of consumed fruits (Bishir and Roberds 1995; Przelomska et al. 2020). These genomes will benefit Latin American germplasm complexes that aim to preserve the avocado's genetic variation and avoid the risk of losing an important component in everyday diets.

2. METHODS

2.1 Sample selection

We searched for *Persea americana* leaf specimens on digital repositories of the New York Botanical Garden (NYBG), the Chicago Field Museum (F), the Missouri Botanical Garden (MO) DNA Bank, and Texas Oklahoma Resource and Consortium of Herbaria (TORCH). These are four of the largest available databases of herbarium leaves within the United States and, taken together, contain enough putatively wild avocado specimens to homogenously sample the avocado's native range. To be sure that the selected samples came from likely wild trees, we sampled from leaves whose metadata indicated that the tree was located either in a forest or roadside away from urban areas or was specifically designated as "wild." Our digital search strategy was to first select as many probable wild specimens from NYBG, F, and MBG that are digitally available, and then use samples from other herbaria on TORCH to homogenously cover the wild geographic range of *P. americana*. Herbaria either shipped complete leaf specimens from which we removed a ~2cm sample for genomic extraction, or they would mail a pre-removed fragment.

2.2 DNA extraction and Illumina library construction

We performed all extraction steps at The Radiocarbon and Isotope Preparation Laboratory and the Modern DNA Laboratory, Anthropology Department, Texas A&M University. Since these specimens are historic, we thoroughly cleaned the tools and workstation with bleach, water, and ethanol in between handling each sample to avoid human or cross contamination. We powdered each specimen in liquid nitrogen using a mortar and pestle. For DNA isolation and purification, we used the Qiagen DNEasy Plant Mini Kit with the following modifications to retain potentially shorter fragments: 1) The initial incubation lasted overnight with gentle rotation at 55°C. 2) After adding buffer P3, samples were placed in a freezer for 10 minutes before centrifugation. 3) Purified DNA was eluted in two 50μl stages after 15 minutes of incubation at 35°C. We quantified DNA using a Qubit Fluorometer, and the 25 samples with the highest yield were selected for further lab work.

We performed all post-extraction lab work at the pre- and post-PCR modern labs at the Laboratory of Molecular Anthropology and Microbiome Research (LMAMR), Oklahoma University. Agilent Tapestation results for each sample showed a range of DNA fragmentation among each sample. We sheared via sonication the samples whose fragmentation was much larger than a mean of 400 bp down to this size. For Illumina library construction, we concentrated samples to 13μl with a SPRI bead cleanup, then used the Kapa Hifi MasterPrep Kit following manufacturer's instructions with end-repair and adapter-ligation solution volumes of 15μl and 27.5μl, respectively. We then did one final bead concentration to 25μl and ran a test qPCR using the Kapa Hifi Hotstart ReadyMix and an IS7 and IS8 primer to discern optimal cycle number. The PCR profile was 3min at 98°C, followed by 35 cycles of 20sec at 98°C, 15sec at 60°C, and 30sec at 72°C, and then one final extension for 1min at 72°C. We then indexed three replicates of each library with separate IS7 and IS8 primers and amplified them with the same profile at either 10 or 13 cycles. Replicates were then pooled and reconcentrated to 30μl and analyzed via Tapestation to confirm their ligated fragment size. We diluted each sample to their correct equimolar concentrations, pooled them, and dried the pool to 30μl via SpeedVac. Finally, we filtered the pooled solution for the correct size range (~300bp) using a Pippin Prep and shipped the filtered sample to the Oklahoma Medical Research Facility for sequencing at a maximum of 10x coverage on a partial Novaseq-S4 lane.

2.3 Sequence read preprocessing

We performed all bioinformatic analyses using the tools available on the Texas A&M High Performance Research Computer. We used SAMtools (Li et al. 2009) and BWA (Li and Durbin 2009) to index the ‘Hass’ avocado cultivar reference genome (GenBank: SDSS00000000.1; Rendon-Anaya et al. 2019). We used AdapterRemoval v2 (Schubert et a. 2016) with default settings on each sample to trim Illumina tru-seq adapters, discard short reads, and merge paired end reads. We then mapped each merged read file as single-end reads to the reference genome using the BWA *aln* command at both default settings and with a relaxed alignment at the ends of reads (-l 1024) to compare the mapped read percentages for both modern and ancient DNA protocols, respectively. We then used SAMtools to convert alignment files to SAM format and obtain mapping statistics followed by the conversion of SAM to BAM format and the filtering of unmapped, low quality (< 20 Phred score), and PCR duplicate reads. mapDamage2.0 (Jónnson et al. 2013) then estimated the chemical damage on the 5’ and 3’ ends of reads, as is done with aDNA read data. We preprocessed 30 additional previously published genomes representing the most important cultivars of each horticultural race with the following changes: 1) we only mapped the truncated pair one reads (matching the herbarium specimens in coverage), 2) they were mapped to the reference genome using BWA *mem* with default parameters, and 3) we omitted the mapDamage analysis.

2.4 Population structure and genetic diversity

Since our sequence data was below 5x coverage for each library, we generated genotype likelihood files (beagle files) for each mapped sample using ANGSD (Korneliussen et al. 2014). We only utilized the genotypes of sites that are shared among at least two-thirds of the population and have a minor allele frequency of at least 5%. With the beagle files, we used

PCAngsd (Meisner and Albrechtsen 2018) to calculate the covariance matrix for each individual, and we use NGSadmix (Skotte et al. 2013) to estimate population structure and admixture among wild avocados at K=2-4. The same was done separately for the 30 published genomes and then for the 55 total herbarium specimen and published genomes taken together to illustrate how cultivar and putatively wild avocado populations are structured and admixed in relation to each other. We then use the realSFS command from ANGSD to calculate the between-population genetic distance (F_{ST}) for each pairwise combination of populations at K=2-4 for the herbarium, published, and combined datasets. Hybrid individuals were assigned to whichever population comprises the majority of that individual's ancestry. We also used realSFS to calculate the the average pairwise nucleotide diversity (π) for all polymorphic loci in each contig. The populations used in the π analysis were identical to those used in the F_{ST} analysis as well as each dataset at K=1. To calculate the final π value for each population, we divided each contig's value by its number of sites, then averaged them. For the ancestral genome, we provided the Hass avocado reference genome and applied "-fold 1" to the realSFS command.

2.5 Identifying escaped cultivars

To tease out potential escaped cultivars in our dataset, we used ngsDist (Vieira et al. 2016) on the combined herbarium specimen and published cultivar beagle file generated by ANGSD to calculate the genetic differentiation (p-distance) for each pairwise combination of herbarium and previously-published cultivar set. We also conducted an outgroup f_3 statistical test for each pairwise combination by using ANGSD '-doPlink' option to generate a plink file for all herbaria and published cultivars plus *Persea donnel-smithi* for an outgroup (n=56). We then applied the popstats.py script (https://github.com/pontussk/popstats; Skoglund et al. 2015) to calculate each outgroup f_3 value (--f3vanilla) with the test f_3(*P. donnel-smithi*, herbarium specimen, published cultivar) for all pairwise combinations of test individuals. We lastly used PCAngsd and NGSadmix again for each horticultural race plus any herbarium specimens that nested in their clusters during the initial population structure analysis. We deemed any herbarium genomes that still nested in the cultivar population to be recently feralized and were therefore pruned from out dataset before further analysis.

2.6 Biogeography and cultivar origin analyses

To understand if the genetic variation in wild avocados may be explained by geography, we used PCAngsd and R to calculate the covariance matrix for both the Central Mexico and the Central and South American herbarium specimen populations. Rather than plotting the principal components together, we performed a multivariate regression analysis with both PC's and the latitude, longitude, and elevation of each individual. We also performed an outgroup f_3 statistical analysis to estimate the geographic origin of important modern cultivar varieties. We used popstats.py again, this time with the test f_3(*P. donnel-smithi*, herbarium specimen, horticultural race population) for all herbarium specimens and cultivar populations, using the same PLINK file generated earlier. With each individual's f_3 statistic for all three races, we created heat maps and looked for geographic regions with excess affinity for a cultivar population (regions whose herbarium specimens have distinctly high f_3 values).

3. RESULTS

3.1 Extraction and sequencing

Our search for avocado herbarium specimens on digital repositories returned several hundred recent (< 60yrs old) leaves collected from trees across the Western Hemisphere. Of these, we selected a total of 40 putatively specimens from 10 herbaria whose metadata suggest they are wild for genome extraction (Table S1). DNA yield ranged from zero to 34 ng/µl, but most (n=39) were between 0.2 and 6, suitable for next generation sequencing. We plotted each specimen's yield against the number of years since its collection and found a weak negative correlation (R^2=0.128, p=0.021; Figure 1a in Appendix 1). This trend is comparable to a previous study analyzing the efficacy of a modified kit-based extraction on herbarium leaves (Marinček et al. 2022), but their yields for recent specimens are higher than those in this study. This could be the product of our extracting a plant species that may retain nucleotides worse under herbarium conditions, or it could reflect the different methods of preservation used by different herbaria. Of these 44 extracted genomes, we selected 25 that had relatively high yield and represented a homogenous sampling of the avocado's putative wild range, including the Caribbean. These samples represent nine countries, from Tamaulipas, Mexico to La Convención, Peru (Table 1 in Appendix 1).

We aligned our herbarium specimen reads (BioProject: PRJNA945882) to the 'Hass' avocado cultivar reference genome (GenBank: SDSS00000000.1; Rendón-Anaya 2019). We used both the default parameters and a relaxed alignment to determine if recent herbarium DNA fragments return a greater percentage of mapped reads if treated like ancient DNA. We found that the default parameters had an average of ~3% higher mapping than the relaxed alignment, which took on average 26 times longer to complete per sample, (Table S2). The percentage of mapped reads for the herbarium avocados ranged from 45% to 85% except for two samples that registered 4% and 16%. Interestingly, when plotting the mapping percentage of specimens above 45% (n=23) against their age, we found a positive correlation (R^2=0.391, p=0.001; Figure 1b in Appendix 1). This trend suggests that somehow brief exposure to herbarium conditions may improve the recovery of high-complexity DNA fragments, and this warrants further research. We sequenced most samples (n=22) at 1-5x coverage, with the remaining three having 0.075-0.22x due to low read counts or mapping percentage. These three are still at acceptable levels for whole-genome skimming and population structure analyses, but we omitted the sample with 0.075x (Ponce_276, Ecuador) from our f_3 analyses due to insufficient coverage. MapDamage results showed no levels of fragment misincorporation on the 3' or 5' ends of mapped reads.

3.2 Population structure

To obtain a general approximation of population structure for wild and commercial cultivar avocados, we used a principal components analysis (PCA) to visualize the distribution of genetic variation among individuals. We ran separate PCA's for the 25 herbarium specimen dataset, the 30 published cultivar genomes dataset, and both groups taken together (Figure 2 in Appendix 1). For the herbarium specimen PCA, we observed two distinct clusters, one of which was concentrated and only included individuals collected in Central Mexico, and the other a wider group of all other samples that appear to cluster roughly based on geographic proximity. For the PCA of the published genomes, we observed three distinct clusters for each of the horticultural races as well as intermittent individuals representing Guatemalan x Lowland (GxL) and Guatemalan x Mexican (GxM) hybrids. When creating a PCA for all 55 samples, we observed that the Central Mexico herbarium group nested with the unadmixed cultivars of the

Mexican variety. The same was true for Panamanian and Dominican herbarium genomes with the Lowland Variety. Herbarium specimens from Mexico, Costa Rica, Peru, and Ecuador clustered around the var. *costaricensis* wild type, and those collected from Honduras and Nicaragua nested between this cluster and the Lowland group. Interestingly, none of the herbarium genomes clustered with the Guatemalan horticultural race, with the exception of an individual collected from Chiapas (Matuda_37384), which appears to nest with the three putative wild samples with predominantly var. *guatemalensis* germplasm.

We further measured population structure and hybridization by calculating admixture graphs for the three datasets, assuming a population number (K) of two to four (Figures 3-5 in Appendix 1). For the herbarium genome dataset, at K=2, there is a clear separation of the two PCA clusters with admixture only occurring in Chiapas to Nicaragua. At K=3 and 4, the Central Mexico population is completely isolated, and there is a moderate degree of genetic continuity between the Southern Mexico/Guatemalan group and the Panama/Dominican Republic group, again in the Nicaragua/Honduras area. Admixture graphs for the 30 published genomes dataset were similar to those reported with SNP markers (Solares et al. 2022) at K=3 and 4. At K=3 the cultivars are separated by horticultural variety, with GxM hybrids and putative wild groups nesting with the Guatemalan population. The transition from K=3 to 4 separates the GxM hybrids from the Guatemalan group. For the combined herbarium and published genomes dataset, K=2 separates the M/Central Mexico group with all other varieties and wild types, with a high degree of admixture among admixed M varieties, GxM hybrids, and herbarium specimens from Chiapas and Guatemala. At K=3, the G and L clusters segregate, and there is extensive admixture between these two populations at Nicaragua, Honduras, and Costa Rica. At K=4, the central PCA cluster including only herbarium genomes and var. *costaricensis* becomes its own population. This group has extensive hybridization between the G and L clusters at Chiapas (including putative wild accessions) and at Honduras/Nicaragua/Panama, respectively.

3.3 Genetic diversity

We calculated nucleotide diversity (π), a sample size-independent estimation of genetic variation, for all populations at K=1-4 for each of the three datasets (Table 2 in Appendix 1). The population that each individual assigned was based on that specimens' highest population identity during each admixture run. Notably, the π value at K=1 for the published cultivar genomes (0.0107) is much smaller than that of the herbarium specimens (0.0138). This %77.5 ratio of cultivar to wild diversity is substantially less than the %91.5 ratio reported in a previous study only utilizing a narrow geographic sample of wild individuals (Chen et al. 2009). However, it should be noted that previous genetic and anthropological research suggest that each avocado variety has its own independent domestication trajectory. A better understanding of diversity reduction associated with domestication would arise from comparing independent varieties and their own separate progenitor populations. The K=3 herbarium population and K=4 joint population including only herbarium genomes from Chiapas/Guatemala/Costa Rica/Nicaragua/South America had the highest π values among all K=3 and 4 populations. The most diverse populations that included published cultivar genomes were those with the Mexican variety, but this number may be inflated since these populations included individuals that were nearly 50/50 hybrids with Guatemalan cultivars. The least diverse groups were those that included the Guatemalan race.

We also calculated the genetic differentiation for all pairwise combination of K=2-4 populations using the fixation index (unweighted F_{ST}; Table 3 in Appendix 1). As expected based on their low levels of admixture with other populations, the herbarium genomes from

Central Mexico had enriched levels of F_{ST} with all other groups (0.238-0.399) compared to F_{ST} tests excluding the Central Mexico group, which were all comparable (0.137-0.194). For the published genomes dataset, K=3 and 4 groups involving the Lowland horticultural race were the most differentiated from other varieties (0.279-0.363). This enrichment for the Lowland variety is likely due to the paucity of clear domesticated hybrids between it and the other two races. By far the smallest F_{ST} value among cultivar populations is between GxM hybrids (mainly representing the relatives of 'Hass,' Solares et al. 2023) and the Guatemalan/Putative Wild groups (0.079), suggesting that all cultivars in these groups should be treated as Guatemalan genomes. Furthermore, the near 50/50 hybrid individuals between the unadmixed Mexican cultivars and the GxM/Guatemalan population should be considered as the true GxM hybrids. The least diverse pairwise combinations of groups from the joint dataset were those between the only-wild/herbarium specimen population with the Guatemalan and Lowland clusters (0.153 and 0.155, respectively). The joint dataset population for the Central Mexico herbarium genomes and the Mexican variety were more distant to other populations (0.269-0.416) than the Mexican race alone was to the other domesticated types (0.237-0.346). In this case, the addition of wild germplasm from Central Mexico, representing the progenitor population for the Mexican variety is further isolating the unadmixed Mexican cultivars from the 50/50 GxM hybrid individuals. This isolation in turn classifies the hybrids as part of the Guatemalan cluster, further differentiating the Mexican and Guatemalan varieties.

3.4 Identifying escaped cultivars

Separating feral/escaped cultivars from purely wild herbarium specimens, particularly those that escaped recently and probably have little wild germplasm, was challenging. We first took an approach to identify potential clones of any of the published cultivar genomes by calculating the genetic distance (P-distance) between all pairwise combinations of herbarium specimen and published cultivar genomes (Table S3). Distance results varied from 0.168 to 0.638, and there were generally lower distance values between pairs that shared an admixture group compared to those that did not. However, none of the pairs returned a distance less than 0.15, used as a threshold for identifying clonal variants in a previous study (Cronin et al. 2020), indicating that none of the herbarium specimens were direct clones of important cultivars. We then took a phylogenetic approach to determine which herbarium genomes shared a greater degree of derived alleles compared to the ancestral state. To carry out this analysis, we used f_3 statistics to measure the shared branch length between pairs of herbarium and cultivar genome in relation to *Persea donnel-smithi*, used as our outgroup. The statistic for all tests of f_3(herbarium specimen, commercial cultivar, *P. donnel-smithi*) ranged from 0.199 to 0.233 (Table S4). While herbarium specimens generally had higher genetic affinity with cultivars that with which they share a cluster, none of the f_3 values were comparably high to warrant discarding any individuals as recently feralized. For our final approach we constructed two separate admixture graphs at K=2, one between Central Mexican herbarium leaves and unadmixed Mexican cultivars, and the other between Honduras/Nicaragua/Panama/Dominican Republic herbarium specimens and Lowland cultivars (Figure 6 in Appendix 1). We found that all cultivar genomes remained in the same population with at least ~50% identity. For the Central Mexico/Mexican Race cluster, we made a PCA to outline these differences (Figure 7 in Appendix 1). To ensure that only wild germplasm is used, we assumed all individuals that nested in cultivar clusters (n=7) were feralized domesticates, and these were pruned from the dataset before further analysis.

3.5 Biogeography Analyses

We created another PCA for the pruned herbarium dataset, discarding potential escaped cultivar genomes (Figure 8a in Appendix 1). Our results still show two main population clusters, one for the Central Mexico individuals and one for all specimens collected in Chiapas and southward. This persistence of population structure suggests that the feral herbarium avocados may have escaped very early on or escaped from local gardens, which are likely to incorporate greater wild germplasm. A K=2 admixture analysis also confirms that the two clusters are well-defined, with a small degree of gene flow in Guatemala, Chiapas, and Nicaragua (Figure 7b in Appendix 1). To better understand how geography may explain the genetic variation in wild avocados, we conducted three multiple regression analyses testing the correlation between the first two principal components and each sample's latitude, longitude, and elevation (Figure 9 in Appendix 1). Results showed a moderate to high significant correlation for all three variables (R^2=0.563, p=0.002; R^2=0.805, p<0.001; and R^2=0.427, p=0.015, respectively). Geography therefore explains a large portion of the genetic variation of wild avocados.

We calculated the f_3 statistic for each truly wild herbarium genome with each K=3 domesticated variety to estimate geographic regions of origin for the three horticultural races. We used the test f_3(*P. donnel*-smithi, wild herbarium sample, horticultural race), where the Mexican variety included only the three unadmixed individuals to avoid Guatemalan germplasm, and the Guatemalan variety included both the Guatemalan and GxM clusters due to their miniscule genetic difference. We then generated a heat map for all three varieties that included each specimen's f_3 value for that variety and their geographic location (Figure 10 in Appendix 1). While this analysis would benefit from a larger sample size, we did observe broad regions with excess affinity for each cultivar group. Wild genomes with enriched f_3 for the Mexican cultivar were found in Central Mexico, with decreasing affinity as one travels south. High f_3 for the Guatemalan race ranges from Veracruz to Northern Nicaragua, although the wild individual from Guatemala appears have the lowest value in this range. The elevated levels in southern Central Mexico may suggest it is a region of hybridization for Mexican and Guatemalan types, or it derive from using reported GxM hybrids as a part of the Guatemalan cultivar test population. The f_3 levels for the Lowland variety cover from Honduras to Panama, as well as Peru. The specimen collected from Santa Lucía, Honduras, has high affinity with all three varieties, but this may be a consequence of its low coverage (0.217x). Interestingly, the Nicaraguan tree from Cerro Mogotón had a distinctly low f_3 statistic for all three horticultural races, and the two individuals from Costa Rica show higher affinity to the Guatemalan cluster than the Lowland.

4. DISCUSSION

This is, to our knowledge, the first reported recovery and analysis of non-modern avocado DNA. Our lower yield compared to a study examining the utility of modern extraction methods for recent herbarium specimens (Marinček et al. 2022) is likely a product of the variation in leaf tissue dependent on a species' phytochemical makeup. Avocado leaf tissue is particularly recalcitrant to DNA extraction, requiring its own specialized extraction protocol to maximize recovery (Nath 2022). We have shown here that a modern laboratory and a kit extraction, minimally modified for the isolation of short DNA fragments, is still suitable for obtaining yields sufficient for WGS studies examining population structure. Our weak negative correlation between yield and specimen age shows that the length of time since an herbarium leaf was collected does affect the likelihood that ample DNA will be present, but we agree that the method of storage for leaves is a more important factor. Many herbaria still use treatments of heat and alcohol, which expedite the degradation process (Bakker et al. 2020). The increase in mapping percentage with the age of the specimen was an unexpected correlation. We encourage further research to determine if brief exposure to herbarium conditions is somehow beneficial to capturing high-complexity DNA fragments. Our near identical mapping results when using either modern or ancient read preprocessing methods alongside the lack of chemical damage at the end of fragments shows that recent herbarium specimens should be treated as modern DNA for bioinformatics.

Our PCA and admixture clustering analyses highlight the differences between wild/feral and commercial cultivar avocado population structure. Commercial breeding restricts the total variation of trees to those originating from specific regions and produces hybrid genomes not found in natural populations. In the absence of breeding, *P. americana* forms two distinct populations, a tight cluster from Central Mexico and a broader cluster from Chiapas and southward. Our admixture analysis demonstrate that the Central Mexico cluster is well isolated from other regions, as it has only a small degree of hybridization in Chiapas, Guatemala, and Nicaragua individuals. This intermediate region likely acts as a zone of hybridization between wild avocado populations. When pruning the dataset for potential escaped cultivars, the population structure and admixture remained largely the same. We therefore propose that wild *P. americana* exists as two relatively isolated populations. There is likely some geographic reproductive barrier that exists in southern Mexico which restricts gene flow from Central Mexico to Chiapas. We hypothesize that the lowland region between the Sierra Madre del Sur and the Sierra Madre de Chiapas is this isolating agent, as the local avocados cultivated in Central Mexico and Guatemala prefer upland environments (Storey et al. 1986; Chanderbali et al. 2013). We require a greater sampling of wild avocados in Mexico and Guatemala to confirm the Isthmus of Tehuantepec as the region that separates the two clusters.

Our cluster analyses for the 30 previously published cultivars were near identical to the study that first sequenced and analyzed them, despite our use of a different *P. americana* reference genome and a bioinformatic method that examines low-coverage genotype likelihoods instead of medium-coverage SNPs (Solares et al. 2022). This similarity illustrates the efficacy of genotype likelihood ratios when assessing population structure, which requires a lower cost for sequencing and is better suited for analyzing potentially low-recovery historic genomes. While there are three clusters reflecting the three horticultural races, the cultivars previously reported as GxM hybrids cluster very close to the pure Guatemalan type and are nested together at K=3. Solares and collegues note that this group of putative GxM hybrids mainly include relatives to the Hass cultivar. Seven domesticate genomes appear to fall under both Mexican and Guatemalan/Hass clusters, including three Mexican (069-02, Bacon, Zutano), two Hass (Fuerte, Pinkerton), and two Guatemalan (Lyon, Anaheim) cultivars. Solares and collegues report on these inconsistencies, and we agree that emerging WGS technology is challenging previous assumptions on the cultivar origin of major domestic lineages.

When we combine both the herbarium and published genome datasets, we show that the wild Central Mexican population nests with the unadmixed Mexican cultivars. The metadata of our definitively wild specimens in this region notably mentions how each tree is located on a steep slope or within a dense forest, further refuting their potential status as an escaped cultivar. We interpret this isolated clustering as an indication that the Mexican horticultural race probably originates from Central/Northeastern Mexico, supporting previous claims (Chanderbali et al. 2013, Chen et al. 2009; Storey et al. 1986). The way in which the sub-Central Mexico population clusters in relation to commercial cultivars paints a different picture. Herbarium individuals that do not cluster with any nearby cultivar populations, and are therefore the best representation of wild germplasm, tend to cluster between the Guatemalan and Lowland groups. This population which segregates at K=4 and includes var. *costaricensis* may represent the progenitor population for both domesticate varieties. However, given that this wild population extends from Chiapas to Peru, the clustering analyses alone are insufficient in determining at any fine scale the geographic origin of either the Guatemalan or Lowland races. One herbarium genome (Matuda_37384, from Siltepec, Chiapas) grouped tightly with the three putative wild Guatemalan types collected from Chiapas (CH-G-07, CH-G-10, CH-G-11). We suggest that these individuals represent either hybrids between the wild sub-Central Mexican avocado population and the Guatemalan variety, or a more recent wild ancestor for Guatemalan avocados.

The three herbarium specimens that rest in between the two wild populations in the joint-dataset PCA are the same as those with substantial admixture between the two main wild avocado populations.

Our estimations of genetic diversity within (π) and between (F_{ST}) wild and cultivar avocado populations highlight the avocado's system of mating in the presence and absence of human intervention. The fact that the genetic differentiation between the two wild populations is greater than all three pairwise combinations of cultivar varieties shows that cultivar hybridization from commercial breeding has led to a decline of isolation between different avocado populations. This disparity would increase with the addition of more LxG and LxM hybrids (e.g., 'Vero Beach' and 'Yon'). Furthermore, the nucleotide diversity within both wild avocado populations is higher than those within each horticultural race with the exception of the Mexican variety, which we suspect will drop considerably when pruning hybrids from the test group. We attribute this reduction of diversity in the transition from wild to domesticate populations to the genetic erosion associated with domestication and clonal propagation, which restricts the number of genomes utilized in commercial farming. Taken together, our results tell the story that during the domestication process, pre-colonial arboriculturists began harvesting a smaller selection of stands within house gardens, which decreased the diversity of avocados grown for human consumption. Over time as the avocado became increasingly isolated from wild germplasm due to more intensive management, its total genetic variation eroded to the levels we measure today. Commercial breeding after Spanish Conquest produced clear hybrids between each variety and driving down the genetic differentiation among all distinct populations. More studies of the avocado's domestication history using ancient genomes are needed to estimate the degree that native cultivars hybridized prior to European contact.

Despite the into domesticated populations through breeding, there is still a ~23% decrease in total diversity from wild to domesticate form in *P. americana*. This a much higher reduction in diversity compared to that calculated by Chen et al. (2009). In their study, they only examined π at four loci across the entire genome. More importantly, their sampling of wild germplasm was not representative of the true wild progenitor population(s), which extends from Chiapas and across all of Central America, following our $f_3$analysis. They also probably sampled feral cultivars, as true wild avocados likely aren't native to the Caribbean, and it makes little sense to have a progenitor of the Mexican variety to grow naturally as far south as Ecuador. We propose that our sampling strategy using herbarium genomes is a more convenient and reliable method of obtaining a homogenous distribution of truly wild trees. Additionally, the presence of low F_{ST} between the sub-Central Mexico wild avocados and the Lowland and Guatemalan variety clusters, which is smaller than the differentiation between it and the Mexican variety and the F_{ST} between the Lowland and Guatemalan varieties, demonstrates its intermediate position between the Guatemalan and Lowland clusters. We propose that this pattern suggests that the sub-Central Mexico avocado population contains the progenitor gene pool to both of these domesticated lineages.

The significant correlation between each wild tree's latitude, longitude, and elevation suggests that the genetic distribution of wild *P. americana* is heavily dependent on environment and geography. Or, at the very least, there is no outside factor that inhibits wild avocados in either population to mating as far as avocado pollen will reach. Our results show that, when both populations are taken together, wild avocados exist as a cline of variation as one travels north to south, west to east, and from high to low elevations. The tolerance to specific levels of heat, moisture, and salinity for specific cultivar varieties probably reflects their geographic origins as areas with similar conditions. In other words, since a primary explanatory factor for the genetic variation in wild avocados is environment, it makes sense that a defining characteristic for each domesticate population is their adaptation to a specific climate regime (Storey et al. 1986).

Our f_3 analysis of each wild tree's affinity for any of the three horticultural races allows us to identify broad regions where each was likely first domesticated. As expected from previous genetic and field studies, the Mexican variety clearly originates from Central Mexico. Whether or not the primarily highland wild individuals contribute the most germplasm to modern commercial cultivars requires a finer sampling of Central Mexico wild avocados. Pinpointing the loci of origin for the Guatemalan and Lowland cultivars is more challenging, particularly because they likely share a wild progenitor population based on our clustering and diversity analyses. The general range of genetic affinity for the Guatemalan variety spans from Veracruz to northern Nicaragua, whereas the Lowland cultivars derive from somewhere between Honduras and Panama. One explanation for the origin of both of these two races is that their domestication began in the centers of their respective high-affinity regions. In this scenario, the Guatemalan cultivar comes from Chiapas/Guatemala and the Lowland cultivar's origin lies in Costa Rica. This model is generally consistent with that proposed by Storey et al. (1986), but the higher f_3 statistic for the Lowland race in Atlantic Coast Panama compared to Pacific Coast Costa Rica indicates that pre-Columbian cultivation and dissemination may have been along the Atlantic Coast.

We offer an alternate scenario for the Guatemalan and Lowland races. With the exception of Stevens_34298, which was not related to any cultivar and may be of a different species, wild trees in Honduras and Northern Nicaragua share a high degree of derived alleles with both horticultural varieties. This joint affinity can be explained by the Guatemalan and Lowland races originating in this restricted region. Here, early horticulturalists grew wild avocados in both upland and lowland ecotones, bringing them to a semi-domesticated state and cementing the signals of domestication origin within this region into the cultivated avocado population. Eventually, these early domesticates disseminated north and south throughout Central America, hybridizing with local wild trees to maintain environmental tolerance. Then, further domestication in these separate regions, likely those proposed as origins in the previous model, develops them into the fully domesticated Guatemalan and Lowland varieties. This model requires further rigorous testing involving a larger sample of wild specimens across Central America and the genomes of ancient domesticated avocados in Honduras and Nicaragua, which may carry traces of both varieties. Regardless, our data does refute claims based on genetic (Chen et al. 2009) and historic (Galindo-Tovar and Arxate-Fernández 2010) data that the Lowland variety was first domesticated in the Maya Lowland. It should be noted that under a model of a single center for domestication for the Guatemalan and Lowland varieties well before the Mayan state, there would be extensive cross-cultural interactions among societies that mainly cultivate lowland or highland-type domesticated avocados. It would therefore still be likely that pre-Columbian lowland Central American societies would introduce the fully domesticated Lowland type to the Yucatec Maya, which are documented by the Spanish chroniclers of that culture area (Cobo [1653]1956; Landa [1590]1978).

5. CONCLUSION

We demonstrate the utility of low-coverage whole genome sequencing and the potential of herbarium leaf genomes to provide key insights into the domestication history of plants. Our herbarium specimens, when pruned of potential escaped cultivars, show that wild *P. americana* is structured into two separate populations, one isolated in Central Meixco and the other a contiguous cluster from Chiapas to South America. The reduction in genetic diversity within populations is also greater than that calculated in a previous study, as ours has a more comprehensive sampling of the wild native range of the species. We found that the genetic variation among all wild individuals is likely dependent on environment and geography. We also identified the most likely geographic region of origin for the three horticultural varieties, supporting claims that the Mexican variety originated in Central Mexico. We however also show that the Guatemalan and Lowland varieties may have originated further south than previously thought and may have even been domesticated simultaneously from one area. Further studies of avocado domestication should examine ancient domesticated avocado genomes to test our new model.

REFERENCES

Arzate-Fernández, A., Elena Galindo-Tovar, M., and Romulo Raggio, F. (2010). West Indian avocado: where did it originate? *International Journal of Experimental Botany* 0031, 203–207. Available at: www.revistaphyton.fund-romuloraggio.org.ar.

Ashworth, V. E. T. M., Kobayashi, M. C., De La Cruz, M., and Clegg, M. T. (2004). Microsatellite markers in avocado (Persea americana Mill.): Development of dinucleotide and trinucleotide markers. *Sci Hortic* 101, 255–267. doi: 10.1016/j.scienta.2003.11.008.

Bakker, F. T., Bieker, V. C., and Martin, M. D. (2020). Editorial: Herbarium Collection-Based Plant Evolutionary Genetics and Genomics. *Front Ecol Evol* 8. doi: 10.3389/fevo.2020.603948.

Betz, V. M. (1999). Cotton, maize, and chocolate. Plant cultivation in Mesoamerica. *Athena Rev 2*, 24–31.

Buckler IV, E. S., Pearsall, D. M., and Holtsiord, T. P. (1998). Climate, plant ecology, and Central Mexican Archaic subsistence. *Curr Anthropol* 39, 152–164.

Cañas-Gutiérrez, G. P., Galindo-López, L. F., Arango-Isaza, R., and Saldamando-Benjumea, C. I. (2015). Diversidad genética de cultivares de aguacate (Persea americana Mill) en Antioquia, Colombia. *Agronomía Mesoamericana* 26, 129. doi: 10.15517/am.v26i1.16936.

Chanderbali, A. S., Albert, V. A., Ashworth, V. E. T. M., Clegg, M. T., Litz, R. E., Soltis, D. E., et al. (2008). Persea americana (avocado): Bringing ancient flowers to fruit in the genomics era. *BioEssays* 30, 386–396. doi: 10.1002/bies.20721.

Chanderbali, A. S., Soltis, D. E., Soltis, P. S., and Wolstenholme, B. N. (2013). “Taxonomy and Botany,” in *The avocado: botany, production, and uses*, eds. B. Schaffer, B. N. Wolstenholme, and A. W. While (Wallingford, UK: CABI), 31–50.

Chen, H., Morrell, P. L., Cruz, M. D. La, and Clegg, M. T. (2008). Nucleotide diversity and linkage disequilibrium in wild avocado (Persea americana Mill.). *Journal of Heredity* 99, 382–389. doi: 10.1093/jhered/esn016.

Cobo, B. (1653). *Historia de Nuevo Mundo*. 1st ed. Madrid: Biblioteca de Autores Españoles.

Cronin, D., Kron, P., and Husband, B. C. (2020). The origins and evolutionary history of feral apples in southern Canada. *Mol Ecol* 29, 1776–1790. doi: 10.1111/mec.15277.

de Landa, D. (1590). *Relación de las cosas de Yucatán.* México: Editorial Porrua.

Dillehay, T. D., Goodbred, S., Pino, M., Vásquez Sánchez, V. F., Tham, T. R., Adovasio, J., et al. (2017). Simple technologies and diverse food strategies of the Late Pleistocene and Early Holocene at Huaca Prieta, Coastal Peru. *Sci Adv* 3, e1602778. Available at: https://www.science.org.

Fuller, D. Q. (2018). Long and attenuated: comparative trends in the domestication of tree fruits. *Veg Hist Archaeobot* 27, 165–176. doi: 10.1007/s00334-017-0659-2.

Furnier, G. R., Cummings, M. P., and Clegg, M. T. (1990). Evolution of the Avocados as Revealed by DNA Restriction Fragment Variation. Available at: https://academic.oup.com/jhered/article/81/3/183/881098.

Galindo-Tovar, M. E., Ogata-Aguilar, N., and Arzate-Fernández, A. M. (2008). Some aspects of avocado (Persea americana Mill.) diversity and domestication in Mesoamerica. *Genet Resour Crop Evol* 55, 441–450. doi: 10.1007/s10722-007-9250-5.

Guzmán, L. F., Machida-Hirano, R., Borrayo, E., Cortés-Cruz, M., Espíndola-Barquera, M. del C., and García, E. H. (2017). Genetic structure and selection of a core collection for long term conservation of avocado in Mexico. *Front Plant Sci* 8. doi: 10.3389/fpls.2017.00243.

Harvard School of Public Health (2022). Avocados. *The Nutrution Source*.

Janzen, D. H., and Martin, P. S. (1982). Neotropical Anachronisms: The Fruits the Gomphotheres Ate. *Science (1979)*, 19–27. doi: 10.1126/science215.4528.19.

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., and Orlando, L. (2013). MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. in *Bioinformatics*, 1682–1684. doi: 10.1093/bioinformatics/btt193.

Kahler, A. L., Sharon, D., Mhameed, · S, Lahav, · E, Lavi, · U, Cregan, P. B., et al. (1996). Contribution from the Agricultural Research Organization, The Volcani Center.

Kistler, L., Newsom, L. A., Ryan, T. M., Clark, A. C., Smith, B. D., and Perry, G. H. (2015). Gourds and squashes (Cucurbita spp.) adapted to megafaunal extinction and ecological anachronism through domestication. *Proc Natl Acad Sci U S A* 112, 15107–15112. doi: 10.1073/pnas.1516109112.

Konrade, L., Shaw, J., and Beck, J. (2019). A rangewide herbarium-derived dataset indicates high levels of gene flow in black cherry (Prunus serotina). *Ecol Evol* 9, 975–985. doi: 10.1002/ece3.4719.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324.

Mac Neish, Richard S. (1964). Ancient Mesoamerican civilization. *Science* 143, 531–537.

Marinček, P., Wagner, N. D., and Tomasello, S. (2022). Ancient DNA extraction methods for herbarium specimens: When is it worth the effort? *Appl Plant Sci* 10. doi: 10.1002/aps3.11477.

Meisner, J., and Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics* 210, 719–731. doi: 10.1534/genetics.118.301336.

Przelomska, N. A. S., Armstrong, C. G., and Kistler, L. (2020). Ancient Plant DNA as a Window Into the Cultural Heritage and Biodiversity of Our Food System. *Front Ecol Evol* 8. doi: 10.3389/fevo.2020.00074.

Rendón-Anaya, M., Ibarra-Laclette, E., Méndez-Bravo, A., Lan, T., Zheng, C., Carretero-Paulet, L., et al. (2019). The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. *Proc Natl Acad Sci U S A* 116, 17081–17089. doi: 10.1073/pnas.1822129116.

Sand Korneliussen, T., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. Available at: http://www.biomedcentral.com/1471-2105/15/356.

Schele, L. (1974). "Observations on the cross motif at Palenque," in *Primera mesa redonda de Palenque*, ed. R. M. Green (Pebble Beach, CA: Robert Louis Stevenson School, Pre-Columbian Art Research), 41–61.

Shahbandeh, M. (2023). Global avocado production 2000-2021. *statista*.

Shahbandeh, M. (20232). Avocado production worldwide 2021, by country. *statista*.

Simon, M. F., Mendoza Flores, J. M., Liu, H. L., Martins, M. L. L., Drovetski, S. V., Przelomska, N. A. S., et al. (2022). Phylogenomic analysis points to a South American origin of Manihot and illuminates the primary gene pool of cassava. *New Phytologist* 233, 534–545. doi: 10.1111/nph.17743.

Skoglund, P., Mallick, S., Bortolini, M. C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M. L., et al. (2015). Genetic evidence for two founding populations of the Americas. *Nature* 525, 104–108. doi: 10.1038/nature14895.

Skotte, L., Korneliussen, T. S., and Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics* 195, 693–702. doi: 10.1534/genetics.113.154138.

Smith, C. E. (1966). Archeological Evidence for Selection in Avocado. *Econ Bot* 20, 169–175.

Smith, C. E. (1969). Additional Notes on Pre-Conquest Avocados in Mexico. *Econ Bot* 23, 135–140.

Solares, E., Morales-Cruz, A., Balderas, R. F., Focht, E., Ashworth, V. E. T. M., Wyant, S., et al. (2022). Insights into the domestication of avocado and potential genetic contributors to heterodichogamy. *G3 Genes|Genomes|Genetics*. doi: 10.1093/g3journal/jkac323.

Soltis, P. S. (2017). Digitization of herbaria enables novel research. *Am J Bot* 104, 1281–1284. doi: 10.2307/26641647.

Spengler, R. N., Petraglia, M., Roberts, P., Ashastina, K., Kistler, L., Mueller, N. G., et al. (2021). Exaptation Traits for Megafaunal Mutualisms as a Factor in Plant Domestication. *Front Plant Sci* 12. doi: 10.3389/fpls.2021.649394.

Statista Research Department (2022). Major exporters of avocados worldwide 2020. *statista*.

Storey, W. B., Bergh, B., and Zentmyer, G. A. (1986). The Origin, Indigenous Range, and Dissemination of the Avocado. *California Avocado Society* 70, 127–133.

Tulig, M., Tarnowsky, N., Bevans, M., Kirchgessner, A., and Thiers, B. M. (2012). Increasing the efficiency of digitization workflows for herbarium specimens. *Zookeys* 209, 103–113. doi: 10.3897/zookeys.209.3125.

Vieira, F. G., Lassalle, F., Korneliussen, T. S., and Fumagalli, M. (2016). Improving the estimation of genetic distances from Next-Generation Sequencing data. *Biological Journal of the Linnean Society* 117, 139–149. Available at: https://academic.oup.com/biolinnean/article/117/1/139/2440246.

Weiß, C. L., Schuenemann, V. J., Devos, J., Shirsekar, G., Reiter, E., Gould, B. A., et al. (2016). Temporal patterns of damage and decay kinetics of dna retrieved from plant herbarium specimens. *R Soc Open Sci* 3. doi: 10.1098/rsos.160239.

APPENDIX A: FIGURES AND TABLES

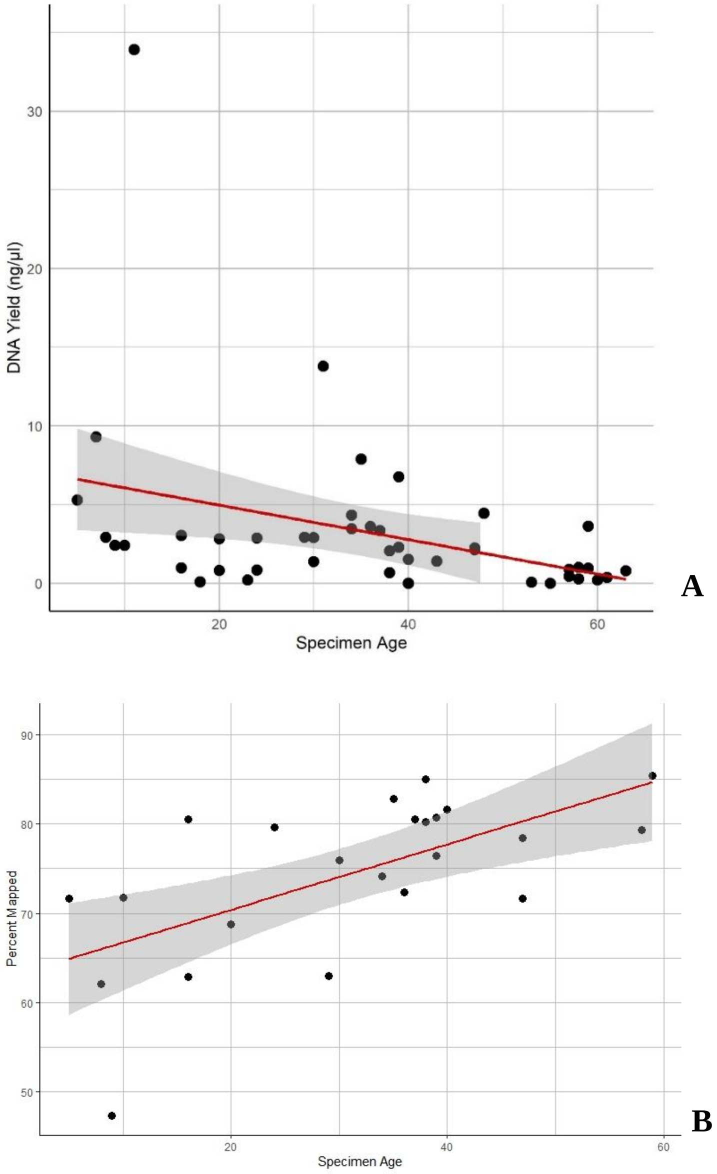

Figure 1. Regression analyses of specimen age versus yield and percent mapped
A. Linear regression of all 45 extracted herbarium specimens, plotting their age since collection versus DNA yield, measured via Qubit Fluorometer. R^2=0.128, p=0.021.
B. Linear regression of the 25 samples selected for whole-genome sequencing excluding those with less than 45% mapping, plotting their age versus the proportion of reads mapped. R^2=0.391, p=0.001.

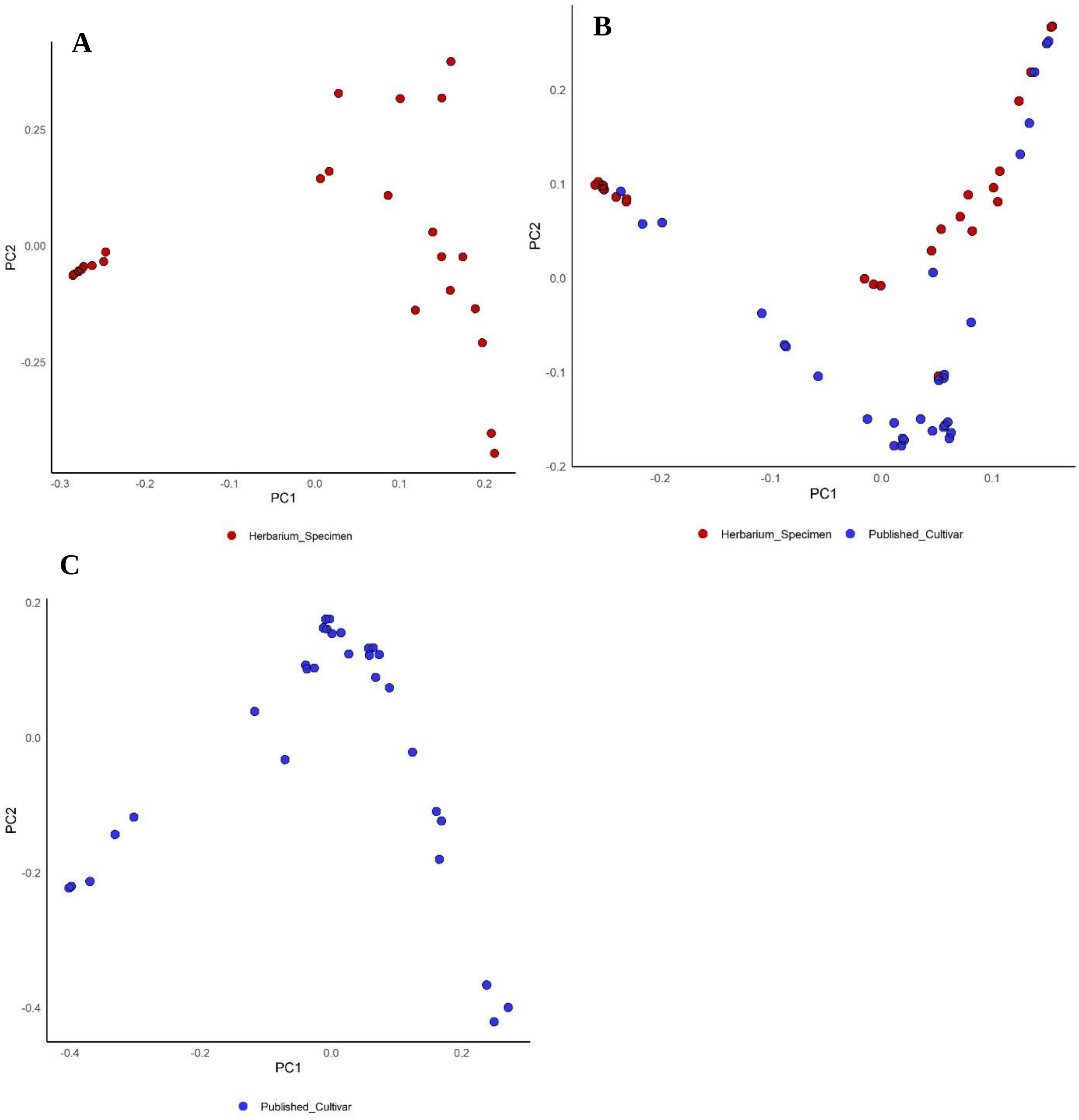

Figure 2. PCA's of each genome dataset
Principle components analyses of the 25 herbarium genomes (**A**), 30 previously published cultivar genomes (**B**), and both datasets taken together (**C**), visualizing the general population structure for wild and domesticated avocado.

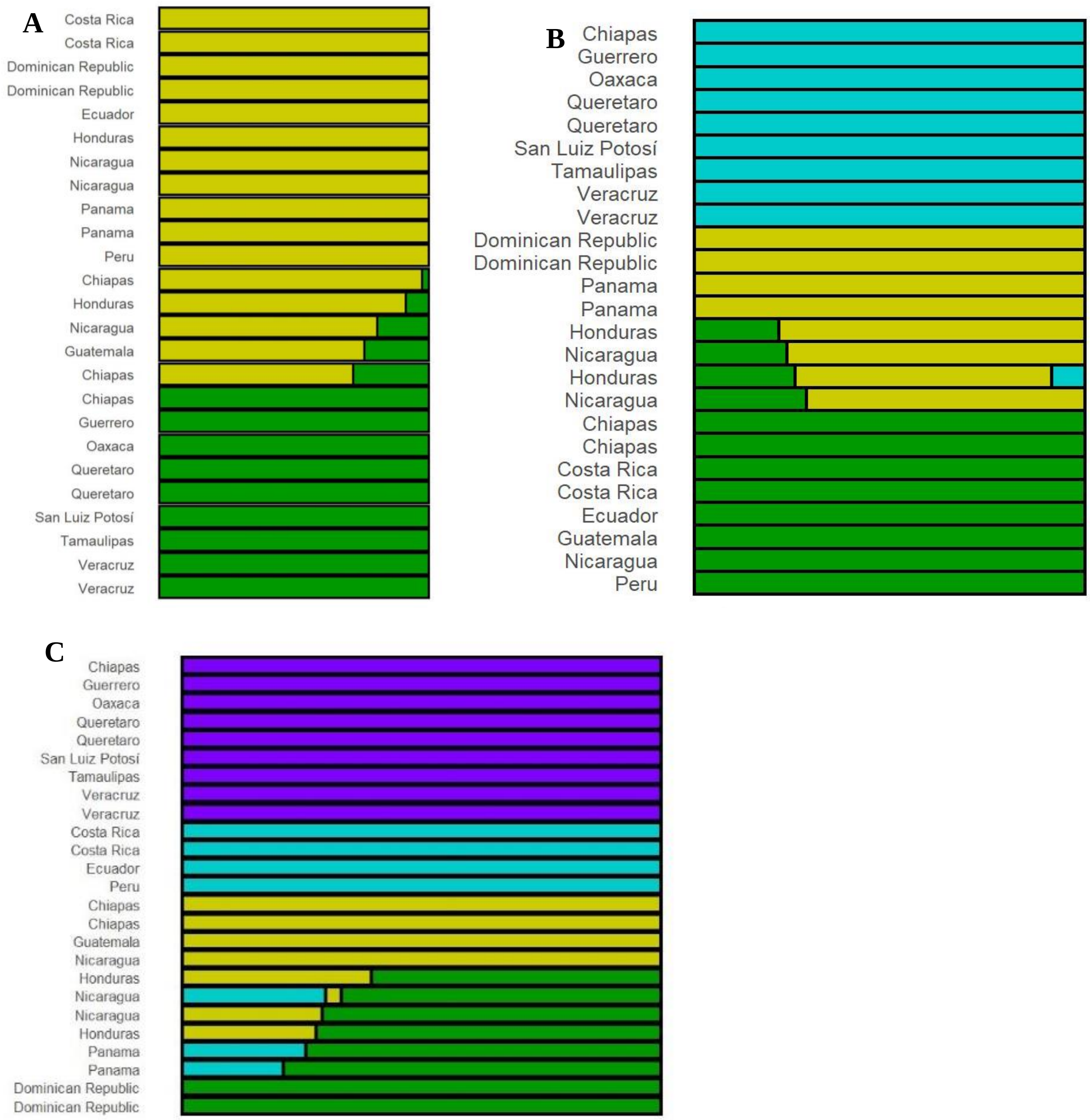

Figure 3. Admixture graphs of the herbarium dataset at K=2-4
Results of ngsAdmix for the 25 herbarium avocados, illustrating the proportion of each sample's genome that belongs to one or more populations, at K=2, 3, and 4 (**A**, **B**, and **C**, respectively).

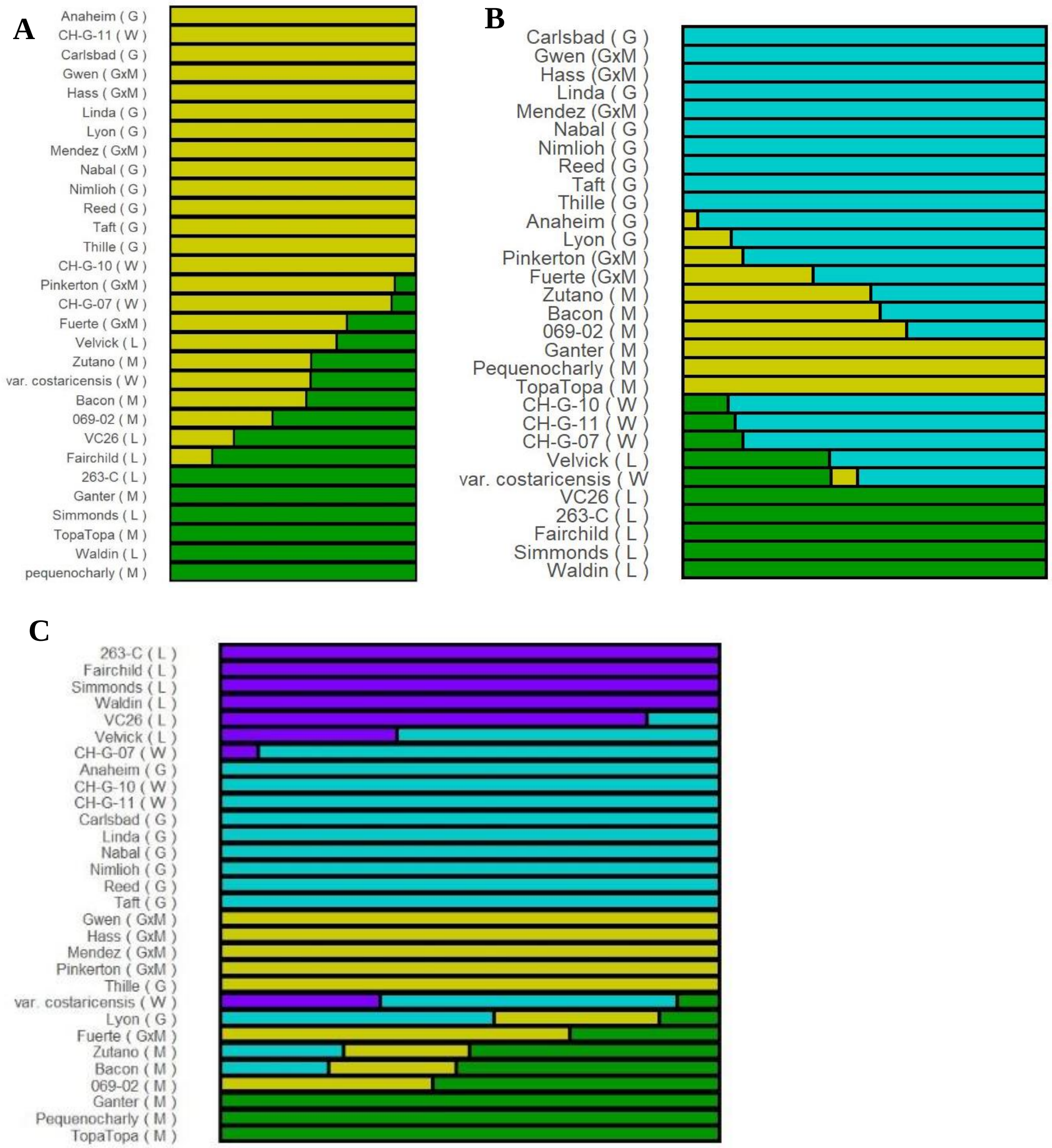

Figure 4. Admixture graphs of the published genome dataset at K=2-4
Results of ngsAdmix for the 30 published avocado cultivar dataset, illustrating the proportion of each sample's genome that belongs to one or more populations, at K=2, 3, and 4 (**A**, **B**, and **C**, respectively). M=Mexican variety, G=Guatemalan variety, GxM=Guatemalan x Mexican hybrid ('Hass' relatives), L=Lowland Variety, W=putative wild.

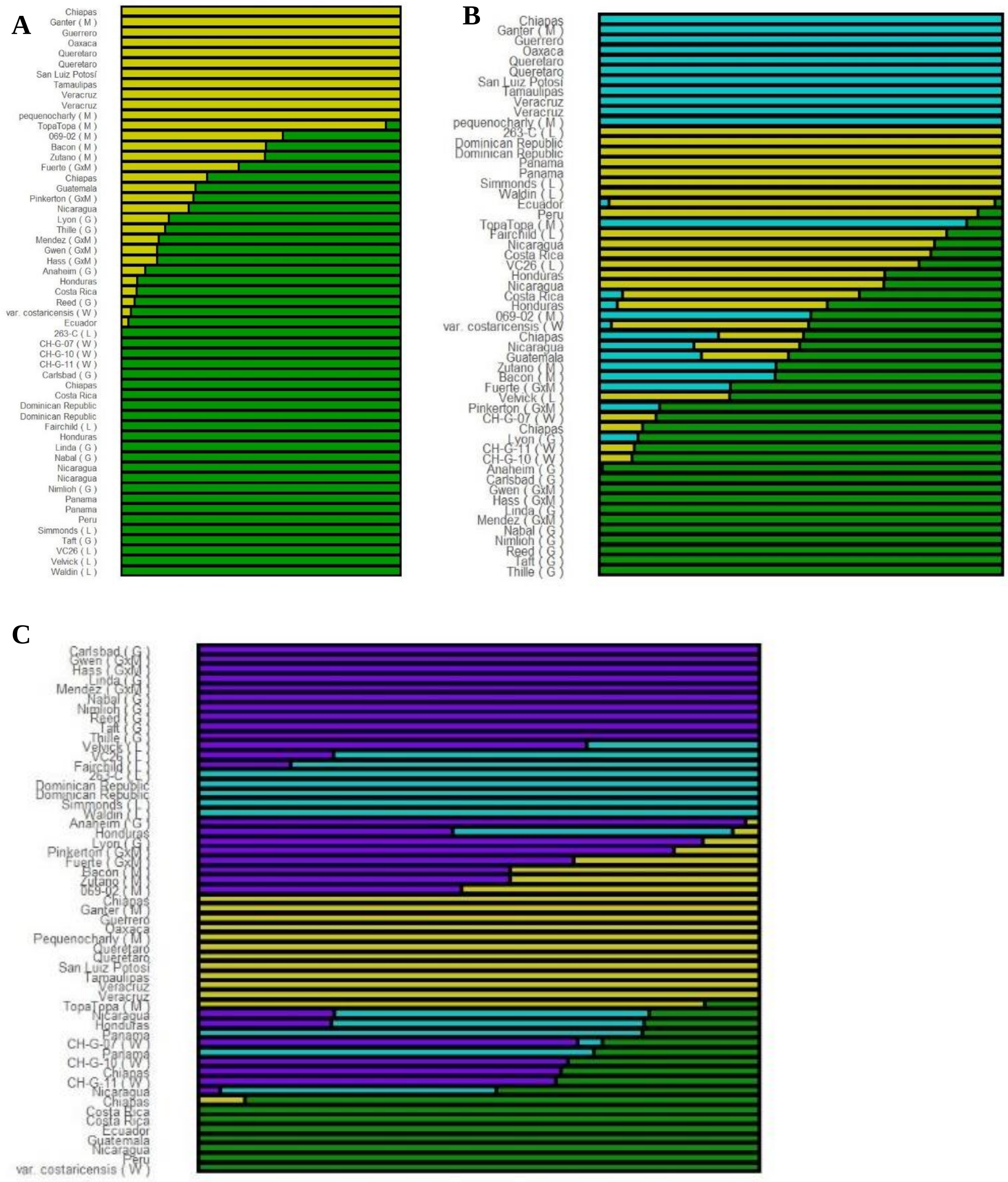

Figure 5. Admixture graphs of the joint dataset at K=2-4
Results of ngsAdmix for both the 25 herbarium specimens and the 30 published avocado cultivars datasets taken together, illustrating the proportion of each sample's genome that belongs to one or more populations, at K=2, 3, and 4 (**A**, **B**, and **C**, respectively). M=Mexican variety, G=Guatemalan variety, GxM=Guatemalan x Mexican hybrid ('Hass' relatives), L=Lowland Variety, W=putative wild.

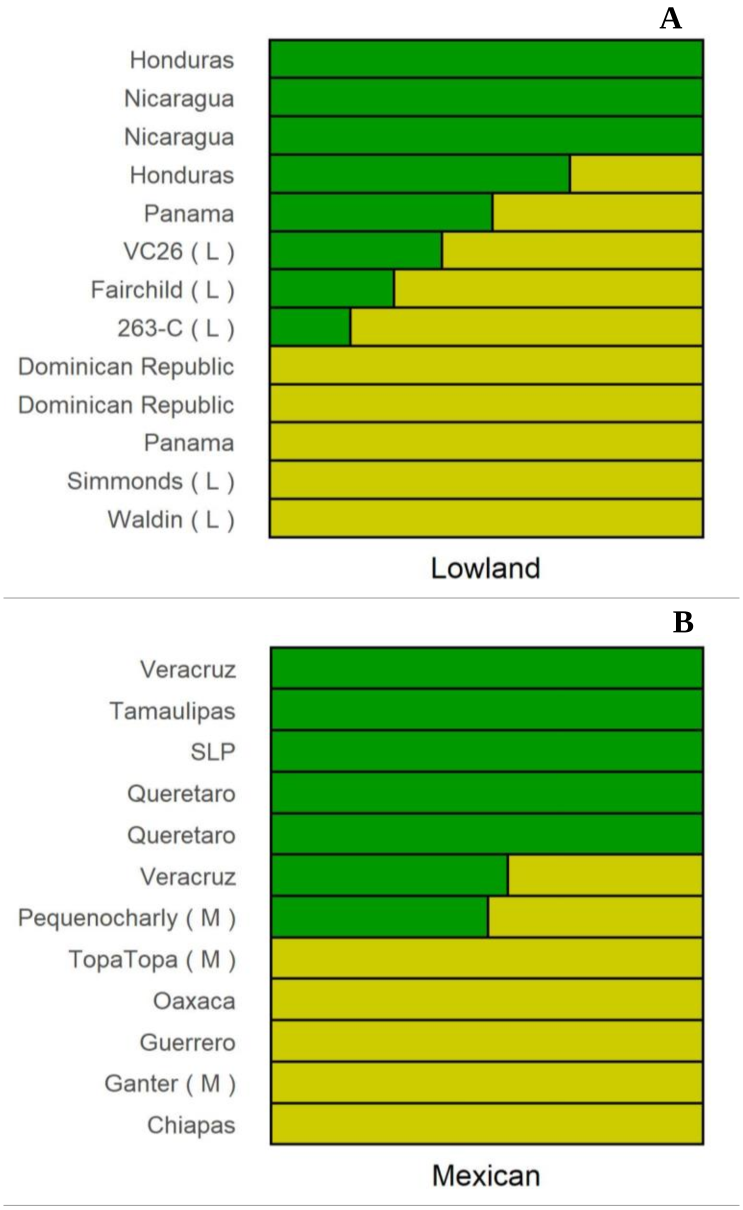

Figure 6. Admixture graphs at K=2 for all individuals that clustered with the Guatemalan and Mexican varieties
Results of ngsAdmix for herbarium specimens and Lowland (**A**) and Mexican (**B**) commercial cultivars that nest together during clustering analyses, identifying potential escaped/feral cultivars within the herbarium dataset. M=Mexican cultivar, L=Lowland cultivar.

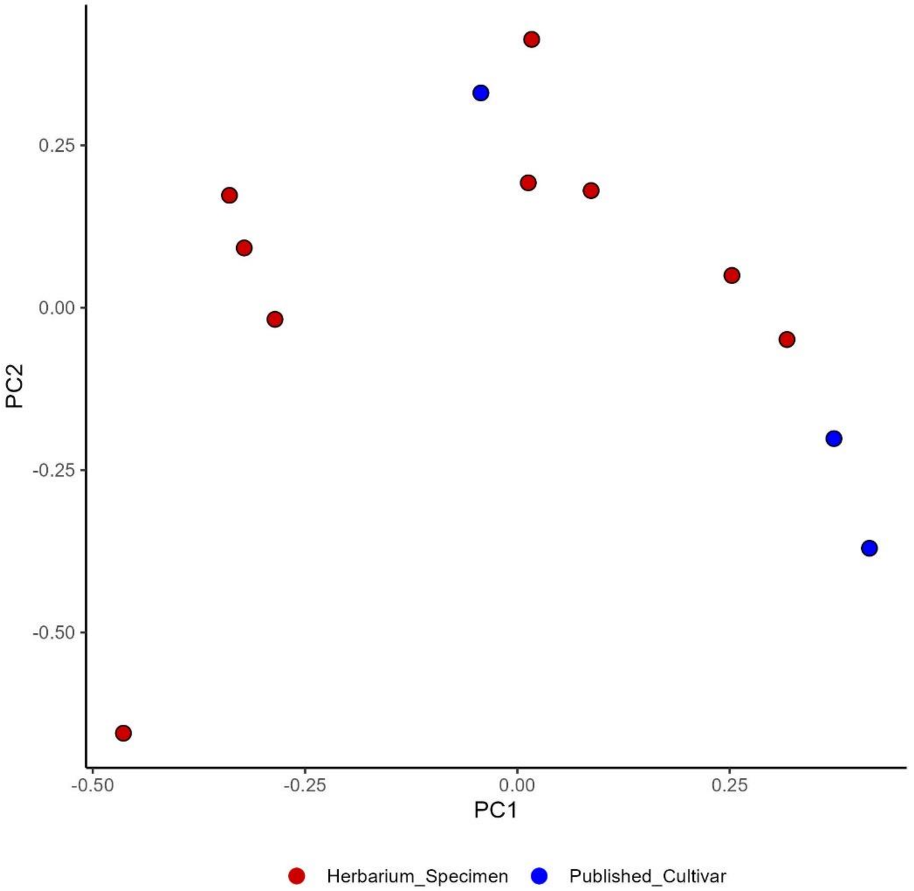

Figure 7. PCA of all individuals that clustered with the Mexican variety

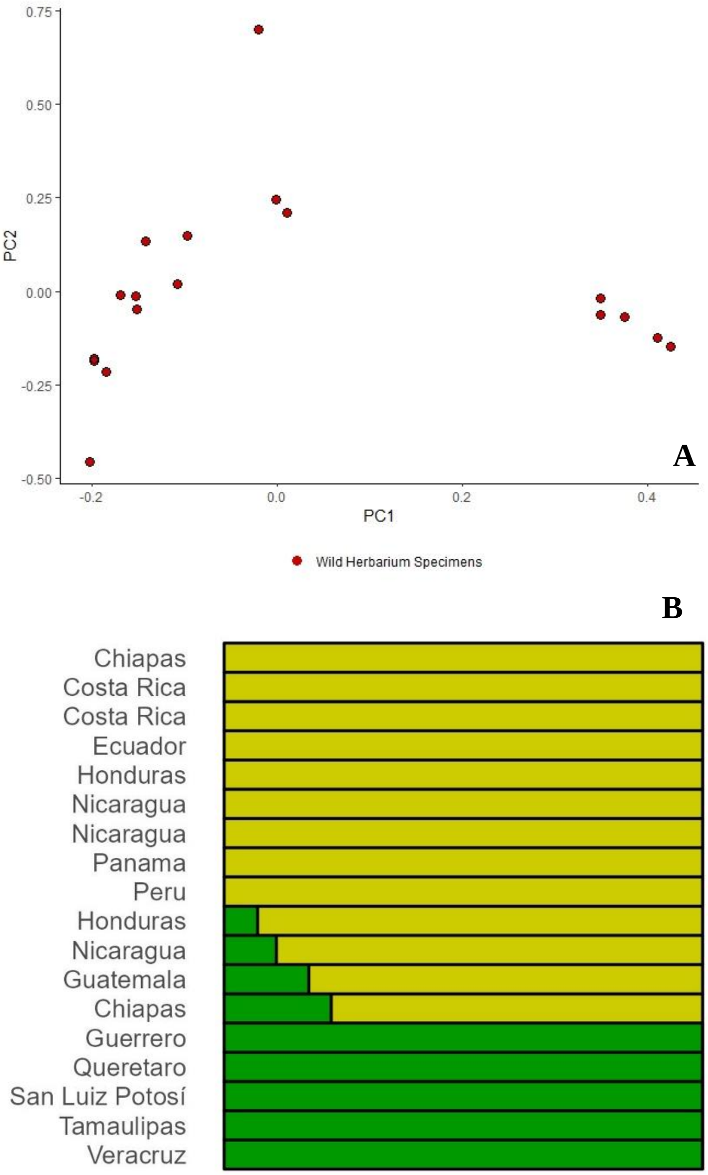

Figure 8. PCA and admixture graph of the herbarium individuals confirmed as wild
PCA (**A**) and admixture graph (**B**) of the 18 likely wild herbarium samples, based on genetic differentiation from domesticated varieties.

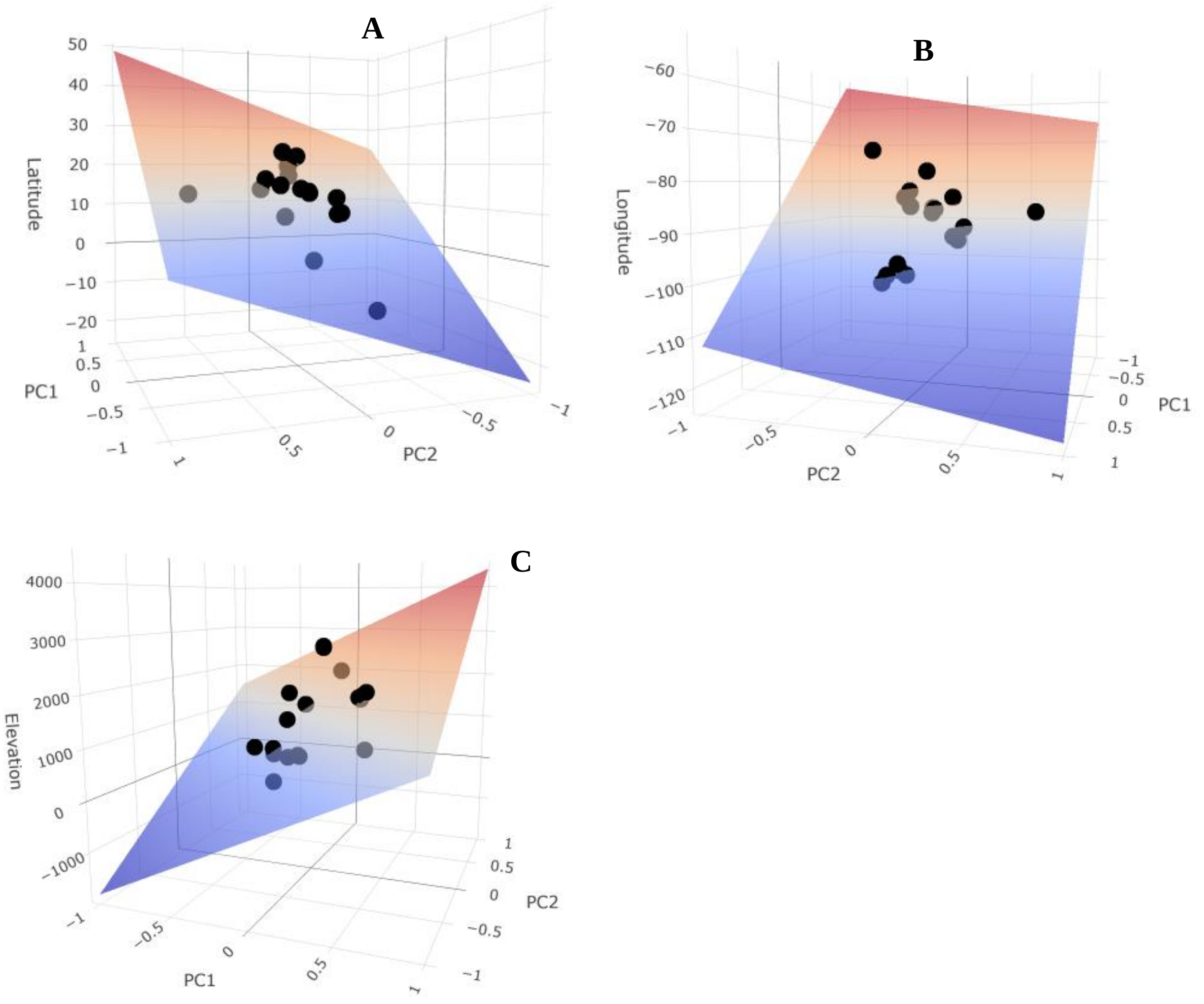

Figure 9. Multivariate regression analyses of each wild genome with geographic characteristics versus their first two principle component values
Regression results for the 18 likely wild herbarium genomes, testing the correlation between both principal components of each genome with their latitude (**A**, R^2=0.563, p=0.002), longitude (**B**, R^2=0.805, p<0.001), and elevation (**C**, R^2=0.427, p=0.015).

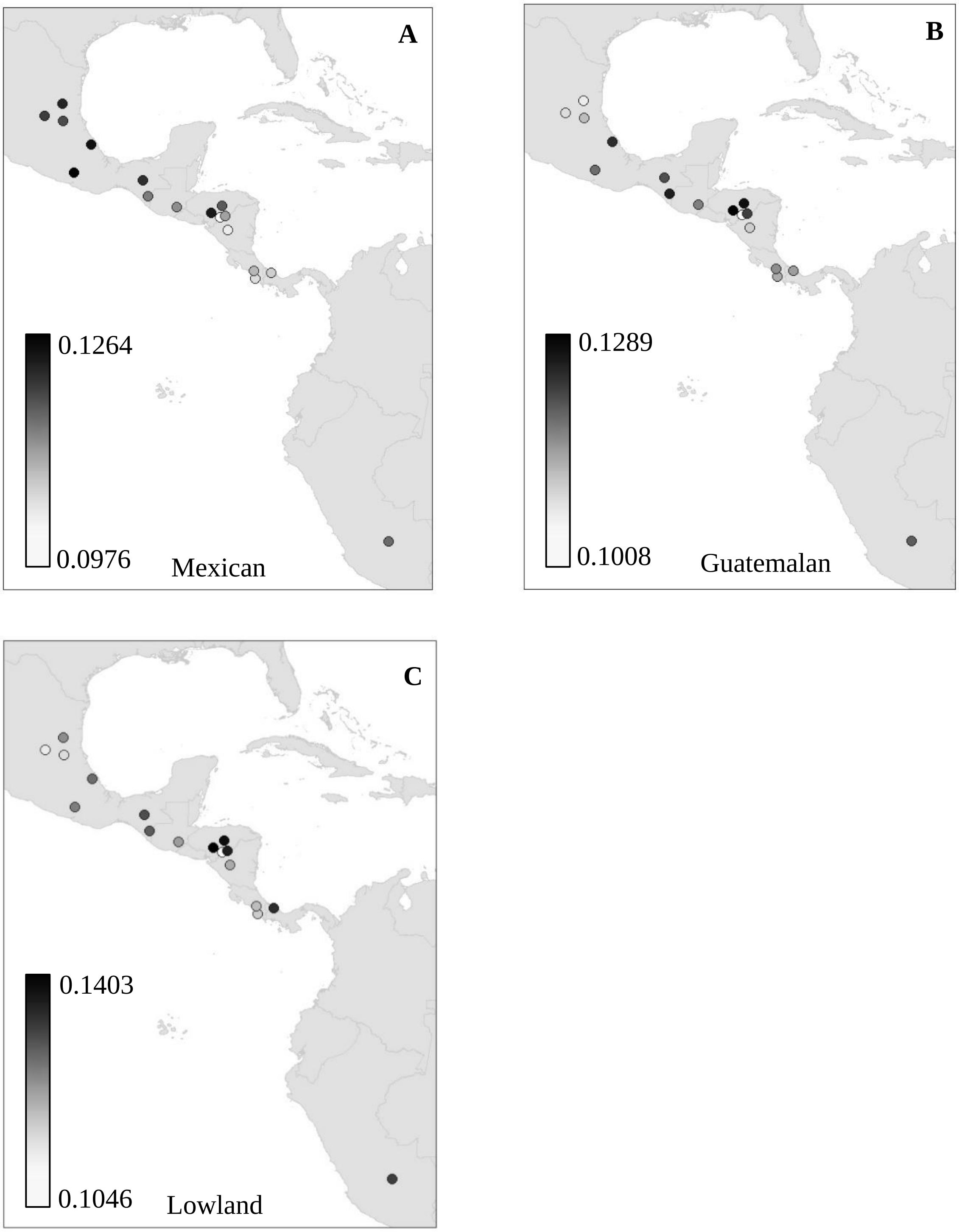

Figure 10. f3 statistical analysis of each wild genome and their affinity to each of the horticultural varieties
Heat maps illustrating each likely wild herbarium specimen’s genetic affinity for the Mexican (**A**), Guatemalan (**B**), and Lowland (**C**) horticultural varieties (high f3=high affinity).

Table 1. Relevant sequenced herbarium specimen metadata

Col Date	Collector	Coll No.	Country	State/Department	Decimal Lat	Decimal Long	Elev (m)	Q20 cov
2/28/1976	Manuel G. Zola Baez	189	Mexico	Veracruz	19.959	-97.2027	420	2.93025
3/13/1965	D. E. Breedlove	9306	Mexico	Chiapas	16.8222	-92.5082	2011.7	1.43633
4/1/1964	E. Matuda	37384	Mexico	Chiapas	15.47722	-92.2728	1500	1.80262
4/4/1987	R. Fernandez Nava	3825	Mexico	Queretaro	21.1365	-99.5751	1650	2.43988
10/23/1994	J. L. Panero, E. Marique, I. Calzada	5186	Mexico	Oaxaca	17.1992	-97.9875	1875	1.72716
2/25/1989	P. M. Peterson, C. R. Annable	7105	Panama	Bocas del Toro	9.1872	-82.2124	15	3.11579
8/6/1993	M. Ponce	276	Ecuador	Napo	-1.07	-77.6	450	0.075952
1/24/1993	Karsten Thomsen	568	Costa Rica	Puntarenas	8.72	-83.52	350	2.88372
11/13/1983	Viveros & Casas	194	Mexico	Guerrero	17.37043	-98.344	1800	1.66058
3/20/1984	T. Zanoni, J. Pimentel, R. Garcia	29315	Dominican Republic	Samana	19.28	-69.17	178	4.98131
6/19/1984	T. Zanoni, M. Mejía, J. Pimentel, R. García	30704	Dominican Republic	Cordillera Central	18.68	-70.13	550	2.89483
4/17/2013	W. D. Stevens & O. M. Montiel	34298	Nicaragua	Nueva Segovia	13.76031	-86.4018	2070	4.96966

Table 1 (cont.)

1/27/2014	**W. D. Stevens**	**34535**	**Nicaragua**	**Matagalpa**	**12.68528**	**-85.7514**	**595**	**6.9728**
1/11/2015	W. D. Stevens & O. M. Montiel	35580	Nicaragua	Nueva Segovia	13.81764	-85.982	1230	2.9006
4/13/1985	F. Alvarado Flores	152	Honduras	Olancho	14.685	-86.2306	460	2.72434
4/29/1985	Z. Nolasco	170	Honduras	Francisco Morazan	14.11472	-87.1097	1800	0.217066
2/28/1986	Gordon D. McPherson	8518	Panama	Colon	9.5	-79.666	50	3.30926
2/23/2003	Brad Boyle	7427	Mexico	Tamaulipas	23.0451	-99.263	1750	0.179687
2/22/2007	Y. Ramirez-Amezcua, E. Carranza	851	Mexico	Querétaro	21.6325	-99.1956	790	3.60273
3/7/2003	F. García	4025	Mexico	San Luis Potosí	22.03542	-100.731	1950	5.65643
5/1/2018	Ronald L. Jones	10235	Costa Rica	San José	9.347455	-83.6334	740	2.19156
5/24/2007	G. Calatayud, H. Coasaca, M. Luza, N. Anaya, M. Callalli, F. Zamora	4113	Peru	La Convencion	-12.7733	-72.6172	909	5.36643
7/15/1999	J. Morales	303	Guatemala	Jalapa	14.5833	-89.9167	2700	1.09609
1/20/1976	S. A. Reyes	83	Mexico	Veracruz	19.683	-96.919	1850	1.13001
4/30/1988	D. Breedlove	67047	Mexico	Chiapas	16.7572	-92.7139	2740	1.19897

Table 2. Nucleotide diversity values within each analyzed population

Herbarium Specimen K	Population	*π*	Published Cultivars K	Population	*π*	Both Datasets K	Population	*π*
1	all	0.013793	1	all	0.010657	1	all	0.012228
2	Central Mexico	0.011896	2	M, L	0.013468	2	all others	0.010898
2	b-Central Mexico	0.012684	2	G, GxM, W	0.008597	2	M, Central Mexico	0.0121
3	Peru, Costa Rica, Nicaragua, Guatemala, Chiapas	0.01353	3	L	0.010183	3	G, GxM, W, Chiapas	0.009184
3	Dominican Republic, Panama, Nicaragua, Honduras	0.011654	3	M	0.012739	3	L, Dominican Republic, Panama, Ecuador, Peru, Nicaragua, Honduras, Costa Rica	0.011688
3	Central Mexico	0.011896	3	G, GxM, W	0.008152	3	M, Central Mexico	0.011954
4	Dominican Republic, Nicaragua, Honduras, Panama	0.011654	4	M	0.012864	4	Ecuador, Costa Rica, Peru, Guatemala, Nica, var. costaricensis, Chiapas	0.013327
4	Chiapas, Guatemala, Nicaragua	0.01262	4	GxM,G	0.007066	4	M, Central Mexico	0.011954
4	Peru, Ecuador, Costa Rica	0.012507	4	G, W	0.008848	4	L, Dominican Republic, Panama, Nicaragua, Honduras	0.010451
4	Central Mexico	0.011896	4	L	0.010183	4	M, G, GxM, W	0.008721

Table 3. Fixation index values between each analyzed population

Herbarium Specimens K	FST	Herbarium Populations	Published Cultivars K	FST	Published Populations	Both Datasets K	FST	Both Datasets Populations
2	0.347155	Central Mexico - Sub-Central Mexico	2	0.153335	M, L - G, GxM, W	2	0.284009	M, Central Mexico - all else
3	0.137216	Ecuador, Peru, Costa Rica, Chiapas, Guatemala - Carib, Panama, Honduras, Nicaragua	3	0.336053	L - M	3	0.192327	G, GxM, Chiapas, W - L, Dominican Republic, Panama, Ecuador, Peru, Nicaragua, Honduras, Costa Rica
3	0.302704	Ecuador, Peru, Costa Rica, Chiapas, Guatemala - Central Mexico	3	0.312877	L - G, GxM, W	3	0.329316	G, GxM, Chiapas, W - M, Central Mexico
3	0.398774	Dominican Republic, Panama, Honduras, Nicaragua - Central Mexico	3	0.236665	M - G, GxM, W	3	0.362273	L, Dominican Republic, Panama, Ecuador, Peru, Nicaragua, Honduras, Costa Rica - M, Central Mexico

Table 3 (cont.)

4	**0.192627**	**Dominican Republic, Panama, Honduras, Nicaragua - Chiapas, Guatemala**	**4**	**0.229**	**M - GxM, G**	**4**	**0.269246**	**Ecuador, Costa Rica, Peru, Guatemala, Nicaragua, var. *costaricensis*, Chiapas - M, Central Mexico**
4	0.170984	Dominican Republic, Panama, Honduras, Nicaragua - Ecuador, Peru, Costa Rica	4	0.252339	M - G, W	4	0.155515	Ecuador, Costa Rica, Peru, Guatemala, Nicaragua, var. costaricensis, Chiapas - L, Dominican Republic, Panama, Nicaragua, Honduras
4	0.398774	Dominican Republic, Panama, Honduras, Nicaragua - Central Mexico	4	0.346066	M - L	4	0.153256	Ecuador, Costa Rica, Peru, Guatemala, Nicaragua, var. costaricensis, Chiapas - G, GxM, W

Table 3 (cont.)

4	**0.19424**	**Chiapas, Guatemala - Ecuador, Peru, Costa Rica**	**4**	**0.079344**	**GxM, G - G, W**	**4**	**0.415561**	**M, Central Mexico - L, Dominican Republic, Panama, Nicaragua, Honduras**
4	0.237708	Chiapas, Guatemala - Central Mexico	4	0.362634	GxM, G - L	4	0.343358	M, Central Mexico - G, GxM, W
4	0.339177	Ecuador, Peru, Costa Rica - Central Mexico	4	0.27946	G, W - L	4	0.281935	L, Dominican Republic, Panama, Nicaragua, Honduras - G, GxM, W