Tree Genetics & Genomes (2023) 19:42
https://doi.org/10.1007/s11295-023-01616-8

ORIGINAL ARTICLE

# Pleistocene-dated genomic divergence of avocado trees supports cryptic diversity in the Colombian germplasm

**Jhon A. Berdugo-Cely[1,2] · Andrés J. Cortés[3,4] · Felipe López-Hernández[3] · Paola Delgadillo-Durán[1] · Ivania Cerón-Souza[1] · Paula H. Reyes-Herrera[1] · Alejandro A. Navas-Arboleda[3] · Roxana Yockteng[1,5]**

Received: 29 October 2022 / Revised: 27 June 2023 / Accepted: 15 August 2023


## Abstract

Genomic characterization of ex situ plant collections optimizes the utilization of genetic resources by identifying redundancies among accessions, capturing cryptic variation, establishing reference collections, and ultimately assisting pre-breeding and breeding efforts. Yet, the integration of evolutionary genomic analyses is often lacking when studying the biodiversity of crop gene pools. Such is the case in the avocado, *Persea americana* Mill., an iconic American fruit tree crop that has seen an unprecedented expansion worldwide because of its nutritional properties. However, given a very restricted number of commercial clones, avocado plantations are becoming more vulnerable to diseases and climate change. Therefore, exploring new sources of evolutionary novelty and genetic diversity beyond the commercial varieties derived from traditional genetic pools in Mexico and Central America is imperative. To fill this gap, we aimed to characterize the genomic diversity of Colombian avocado trees. Specifically, we constructed reduced representation genomic libraries to genotype by sequencing 144 accessions from the Colombian National genebank and 240 materials from local commercial orchards in the Colombian northwest Andes. We merged the resulting reads with available sequences of reference genotypes from known avocado groups (also named as races), Mexican, Guatemalan, and West Indian, to discover 4931 SNPs. We then analyzed the population structure and phylogenetic diversity, and reconstructed evolutionary scenarios, possibly leading to new genetic groups in Colombian germplasm. We detected demographic stratification despite evidence of intergroup gene flow. Besides the classical three avocado groups, we found an exclusive Colombian group with a possible genetic substructure related to the geographical origin (Andean and Caribbean). Phylogenetic and ABC demographic modeling suggested that the Colombian group evolved in the Pleistocene before human agriculture started, and its closest relative from the three recognized races would be the West Indian group. We conclude that northwest South America offers a cryptic source of allelic novelty capable of boosting avocado pre-breeding strategies to select rootstock candidates well adapted to specific eco-geographical regions in Colombia and abroad.

**Keywords** *Persea americana* Mill. · GBS-derived SNP markers · Population genomics · Avocado races · Colombian genetic group

Communicated by J.P. Jaramillo-Correa

✉ Andrés J. Cortés
acortes@agrosavia.co

✉ Roxana Yockteng
ryockteng@agrosavia.co

[1] Corporación Colombiana de Investigación Agropecuaria–AGROSAVIA, Centro de Investigación Tibaitatá, Km 14 Vía Mosquera, Mosquera, Colombia

[2] Corporación Colombiana de Investigación Agropecuaria–AGROSAVIA, Centro de Investigación Turipaná, Km 14 Vía Montería-Cereté, Cereté, Colombia

[3] Corporación Colombiana de Investigación Agropecuaria–AGROSAVIA, Centro de Investigación La Selva, Km 7 vía Rionegro–Las Palmas, Rionegro, Colombia

[4] Facultad de Ciencias Agrarias–Departamento de Ciencias Forestales, Universidad Nacional de Colombia–Sede Medellín, Medellín, Colombia

[5] Institut de Systématique, Evolution, Biodiversité-UMR-CNRS 7205, Muséum National d´Histoire Naturelle, Paris, France 75005

Published online: 14 September 2023

## Introduction

Studying evolutionary and genetic variation in crop species is a significant research avenue for discovering novel attributes that may increase worldwide food and nutrient requirements (Ramirez-Villegas et al. 2022). Yet, climate change jeopardizes global crop production, potentially worsening malnutrition, poverty, and sustainable development, especially in developing countries (Peng et al. 2020). In this context, avocado (*Persea americana* Mill., Lauraceae) arises as a highly nutritious fruit tree crop that is becoming a top commodity worldwide (Sommaruga and Eldridge 2021). However, current avocado plantations rely extensively on the monoclonal propagation of a few commercial varieties, either as rootstocks or scions. In the long term, these low-diversity plantations may prove unsustainable due to the lack of a variable genetic pool capable of responding to biotic and abiotic stresses in the face of climate change (Ingvarsson and Dahlberg 2019). Thus, avocado pre-breeding efforts require exploring the centers of diversity to find new sources of genotypes and alleles capable of enriching the genetic bases of key traits, including the adaptative potential to different agroecosystems (McCouch 2004).

The genus *Persea* Mill. evolved as two different subgenera, *Persea* and *Eriodaphne*, each containing 12 and 11 species, respectively (van der Werff 2002; Pironon et al. 2020). The *Persea* subgenus, to which *P. americana* belongs, is highly diverse in Mesoamerica and South America. Therefore, these regions are considered diverse hotspots for avocados and tree species from the Lauraceae family (Pironon et al. 2020). The avocado tree has been dated to at least 9000 YBP (Years Before the Present) and was first used and selected in the Tehuacan Valley in the State of Puebla in Mexico at around 8000 YBP (Smith 1966, 1969). Based on this evidence, several authors proposed this place as the avocado's center of origin (Arumuganathan and Earle 1991; Galindo-Tovar et al. 2007; Alcaraz and Hormaza 2007; Piperno 2011; Calderón-Vázquez et al. 2013). Interestingly, the oldest archaeobotanical record for this species in Colombia is from the Calima Region in the middle Cauca Valley, and was also dated within the same geological period, $7830 \pm 140$ YBP. This opens the question of where the Colombian avocado originated and how it dispersed to the region (Piperno 2011). In pre-Columbian times, the obligate route for exchanging crops and wild relatives between Mesoamerica and South America was through northwest South America, either by foot through the Darien Gap and the Isthmus of Panama or by canoes across the Gulf of Morrosquillo and the Lesser Antilles (Larranaga et al. 2021). Thus, when the Spanish arrived in America during the Conquest, the distribution of avocados already ranged from Mesoamerica to Ecuador and Peru, and its dispersion was further reinforced by the establishment of human populations following antique pre-Columbian commercial routes (Bergh and Ellstrand 1986; Wolters 1999; Galindo-Tovar et al. 2008). Despite this compelling archeological evidence, the study of the genomic diversity of this species in the northwest of South America, including Colombia, remains in its infancy (Chen et al. 2009; Galindo-Tovar and Arzate-Fernández 2010).

The avocado species currently comprises three horticultural groups (also known as races) that differ in origin, genetic diversity, and horticultural characteristics. These groups are classified as Guatemalan (*P. americana* var. *guatemalensis* (L.) Wms.), Mexican (*P. americana* var. *drymifolia* (Schlecht. et Cham.) Blake) and West Indian (*P. americana* var. *americana* Mill.) (Rendón-Anaya et al. 2019). The Guatemalan group originated in the mid-altitude highlands of Guatemala and is characterized by having small fruits and late fruit maturity. The Mexican group originated in the mid-altitude highlands of Mexico and is known for its early fruit maturity and cold tolerance. In contrast, the West Indian group originated in southern Mexico and Central America's lowlands and has larger fruits with low oil content (Rendón-Anaya et al. 2019). Some reports prefer referring to this last group as "lowland" (Galindo-Tovar and Arzate-Fernández 2010; Solares et al. 2023), yet we stick to the more conventional naming of West Indian. The renaming partly obeys the fact that this group has open questions about its origin and dispersion routes (Galindo-Tovar and Arzate-Fernández 2010). For instance, Spanish chroniclers described avocado trees with West Indian characteristics in northern South America, including Colombia, Ecuador, and Peru, both in mountainous regions as well as on the Pacific Coast and the Amazon basin (Galindo-Tovar and Arzate-Fernández 2010).

Several genetic studies of avocados have supported the three major recognized groups and their evolutionary origin using molecular genetic markers such as ESTs (Expressed Sequence Tags), SSRs (Single Sequence Repeats), and SNPs (Single-Nucleotide Polymorphism) (Gross-German and Viruel 2013; Rubinstein et al. 2019; Ge et al. 2019a; Talavera et al. 2019). In these studies, the Mexican and Guatemalan groups appear more closely related to each other than the lowland West Indian group. Furthermore, several avocado germplasm banks have characterized their accessions at the genetic level to identify the ancestry of the conserved genotypes. These include the Venezuelan gene bank-INIA-CENIAP (Ferrer-Pereira et al. 2017), the US National gene bank repository (SHRS ARS USDA) in Miami (Boza et al. 2018), and the Spanish gene bank (Cañas-Gutiérrez et al. 2019). More recently, the first genome assembly of

an avocado cultivar (Hass) provided a reference sequence of 980 Megabases-Mb (Rendón-Anaya et al. 2019). This genomic information reinforced the substructure of the major avocado groups and provided evidence for the hybrid origin of the commercially famous Mexican × Guatemalan var. Hass (Rendón-Anaya et al. 2019). However, a recent analysis suggests that the Hass variety has a complete ancestry to the Guatemalan group (Solares et al. 2023). The latter report also determined the genome sequence of another avocado cultivar, the Gwen variety (Solares et al. 2023), the first to be successfully assembled to a chromosome level.

Although the growing genomic resources of avocados promise to aid the conservation, discovery, and breeding of new commercial varieties, avocado germplasm from the South American tropics is still poorly characterized (Cañas-Gutiérrez et al. 2022). Furthermore, local accessions and seedling rootstocks in the region lack proper group classification, obscuring the dawn of the nursery plant material, and jeopardizing its optimum resilient and sustainable deployment into already complex geographies. Eventually, nurseries could apply early marker screening at seedling saplings before grafting for racial ancestry, genetic value, disease resistance, and against detrimental alleles (Reyes-Herrera et al. 2020). Still, until then, the lack of genetic traceability also challenges fruit yield and quality and, therefore, overall profitability. Although native avocado production in Colombia is mainly for internal consumption of fresh and processed fruit, the northern Andes also presents favorable agro-climatic conditions (from 1600 to 2200 m above sea level–masl) for producing and exporting avocado cv. Hass. The international avocado market is particularly profitable at the end of the year when worldwide demand is high for guacamoles during Thanksgiving, Christmas, and New Year's Eve festivities, as well as during the Super Bowl in early February, yet the offer is unsatisfied because of a valley in the harvest at producing countries but Colombia (Rios Castaño and Tafur Reyes 2003). Recent exports to Europe, Japan, the USA, and other countries totaling 146 million (M) USD prove this point and reinforce Colombia's potential as a supplier of avocado markets abroad (Navarro-Villa 2022). Colombian production ranks second with 876.7 thousand tons (FAO 2022), and Mexico remains the top producer and consumer (Galindo-Tovar et al. 2011; FAO 2022).

Despite the avocado's undeniable nutritional and commercial value, their plantations are becoming more vulnerable to diseases and long-term climate change effects due to a restricted number of clones used as commercial rootstocks and scions. Therefore, this study aimed to leverage local avocado gene pools from northwest South America by combining the reported SNP allelic diversity of classical avocado groups with the reduced representation genotyping of native accessions and varieties from the Colombian avocado germplasm bank, as well as different seedling rootstocks from commercial orchards in the northwestern Andes of Colombia, a significant producer hub of cv. Hass for exportation. Specifically, we focused on the following research questions: (1) Does the Colombian avocado germplasm belong to the recognized genetic groups from Mesoamerica, or does it correspond to a new gene pool of *P. americana*? (2) What is the genetic group identity of the rootstocks used at commercial orchards in the Colombian northwest Andes? and (3) What is the best evolutionary scenario to explain the genetic origin of the Colombian avocado germplasm based on plausible dispersal routes from Mesoamerica? Exploring the questions above would inform whether Colombian avocado resources provide novel evolutionary and allelic diversity beyond the classical groups from Mexico and Central America.

## Material and methods

### Plant material and reference genomic sequences

This study analyzed the genetic diversity of 456 avocado samples (*P. americana*) from two sources: the avocado collection of the Colombian Germplasm Bank (CGB) ($n = 144$) and Seedling Rootstocks (SR) ($n = 240$) of commercial orchards from the northwest Andes of Colombia, a main avocado-producing region (Table S1). Additionally, sequences of genomic resources of 71 individuals from the study of Talavera et al. (2019) representing the three recognized avocado groups (Guatemalan, GU; Mexican, ME; and West Indian, WI) and hybrids among these groups were included as reference materials (RM). Finally, the genome sequence of *P. schiedeana* was used as an outgroup (O). The latter is a sample obtained from the high-altitude germplasm bank of "Fundación Salvador Sánchez Colín" (CICTAMEX, S.C) located at La Cruz Experimental Center in Coatepec Harinas, Mexico (Rendón-Anaya et al. 2019).

The CGB is managed by AGROSAVIA (Colombian Agricultural Research Corporation) at the Palmira research station in the province of Valle del Cauca (3°32′22.0″N 76°18′13.0″W, 1000 masl). Of the 144 conserved accessions, 16 ("Bacon", "Booth_5", "Booth_7", "Booth_8", "Choquette", "Duke_7", "Edranol", "Fuerte", "Hass", "Lula", "Pollock", "Reed", "Simmonds", "Topa Topa", "Waldin", and "Zutano") are commercial varieties or cultivars (Commercial Materials, CM), and 128 correspond to Colombian native and criollo trees (native × introduced hybrids) (Fig. 1). Some of these accessions were sampled from contrasting agroclimatic regions, particularly areas with high-humidity soils, where it is probable to find tolerant genotypes to *Phytophthora cinnamomi* (Rodríguez-Henao et al. 2017), a major threat in avocado plantations. Following the International Bioversity

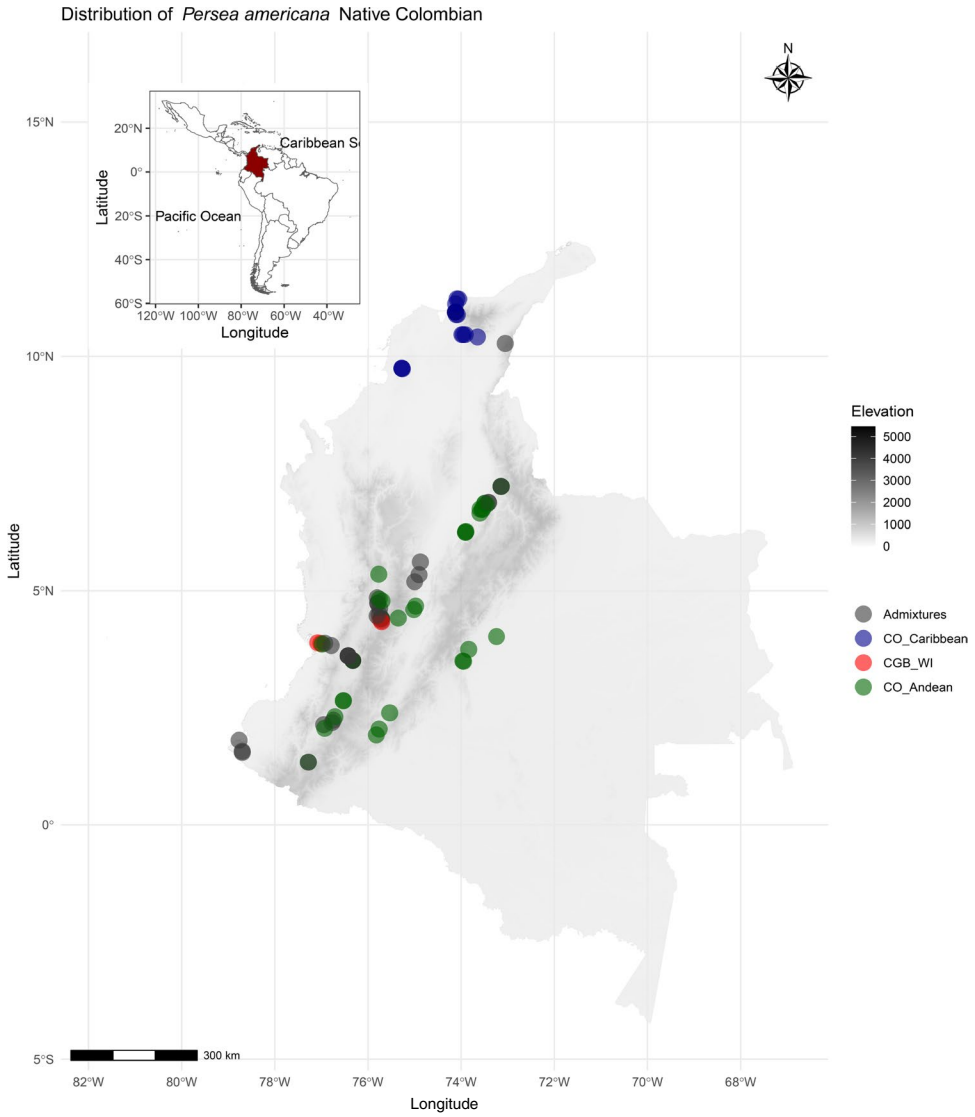



**Fig. 1** Geographic location of the avocado genotypes conserved in the CGB with passport data. The color of the samples represents the assignation of these samples to one specific genetic cluster (Colombian Andean and Colombian Caribbean, West Indian-WI, or hybrids/admixtures) based on the following ancestry analysis results

manual for avocados, this germplasm has been characterized by juvenile morphological and botanical traits (IPGRI 1995). The SR had 240 seedling rootstocks from eight commercial orchards in the Antioquia province, Colombia's primary avocado-producing region (Table S1). This germplasm was previously screened using SSR markers (Cañas-Gutierrez et al. 2019; Reyes-Herrera et al. 2020).

## Molecular genetic characterization of CGB and SR materials

### DNA isolation

We extracted genomic DNA from foliar tissues of 144 accessions conserved in the CGB and from root tissues of 240 seedling rootstocks (SR) using the DNeasy Plant Mini Kit (QIAGEN, Germany) with the following modifications: first, we added 450 µL of the AP1 solution to an equal volume of 20% SDS (sodium dodecyl sulfate), and we incubated the samples for 30 min at 65 °C and froze for 60 min at − 20 °C. DNA quality was verified by electrophoresis in 1% agarose gels dyed with Sybr Safe (Invitrogen), and the DNA concentrations were measured in a Qubit® 2.0 fluorometer (Life Technologies).

### Genomic libraries construction and sequencing

Each avocado DNA sample was double digested using *PstI* (CTGCA-37 °C) and *ApeKI* (G/CWGC-75 °C) enzymes, following the recommendations of New England BioLabs. Genomic libraries of 300 base pairs (bp) were constructed from the digested DNA using the NEBNext Ultra DNA kit for Illumina (E7103L). We used the 96 NEBNext Multiplex Oligos kit indexes for Illumina (E6609L) to identify the sequences corresponding to each sample. Each genomic library was quantified through fluorescence in a Qubit 2.0, and

the average size of the fragments of some random samples was determined by digital electrophoresis on a Tape Station 4200 (Agilent). Finally, the genomic libraries were diluted to a concentration of 10 nM and pooled in groups of 96 samples. The pooled samples were sequenced using a *paired-end* strategy with 150 sequencing cycles in Illumina HiSeq X equipment (Macrogen, Inc. Korea). The demultiplexing of sequences for each genotype was implemented using the *bcl2fastq Illumina* software (Illumina 2022). Raw data sequences were deposited in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) database under the Bioproject number PRJNA878519.

## SNPs identification for the CGB, SR, RM, and O samples

To discover SNPs among all avocado samples, we made the SNP calling using the pipeline reported by Osorio-Guarín et al. (2020). In summary, for all sequence files of each analyzed sample (downloaded and generated in this study), we used the *FastQC* software (Andrews 2010) to verify the quality of the Fastq files. The adapter's primers and sequences with a quality of less than Q30, and a length of less than 50 bp were filtered using *Trim-Galore* software (Krueger 2012). The *Burrows-Wheeler Aligner (BWA)* software (Li and Durbin 2009) was used to align the sequences against the *P. americana* var. Hass cultivar reference genome version 2 (~913 Mb) (Rendón-Anaya et al. 2019). We used *Picard* (BROAD Institute 2022b) and *the Genome Analysis ToolKit-GATK* (BROAD Institute 2022a) software to remove PCR duplicates, correct mapping quality assignment, and realign and recalibrate the reads. The SNP calling was done using the *Unifiedgenotyper* option of the *GATK* algorithm. Finally, we used the *VCFtools* software (Danecek et al. 2011) for maintaining the SNP markers with a Minimum Allele Frequency (*MAF*) of 5%, a minimum of 2X of sequencing depth (*SD*) per position, keeping biallelic SNPs and excluding InDels (insertion-deletion). Samples and SNPs with percentages higher than 20% (genotype level) and 5% (marker level) of missing data were excluded.

Because the nucleotide sequences were gathered from different tissues (leaves and roots) as well as from published genomic resources (Rendón-Anaya et al. 2019; Talavera et al. 2019), we differentially implemented a final filter for polymorphic SNPs discovery to get three goal-oriented datasets with contrasting SNP density (Table 1). Summary statistics, such as Mean Depth of Sequencing (*MDS*), nucleotide diversity (*Pi*), transition/transversion ratio (*Ts/Tv*), *MAF*, and Polymorphism Information Content (*PIC*), were determined in dataset 1 using *VCFtools* software (Danecek et al. 2011) and *R-SNPready* package (Granato et al. 2018) in *R* software (R project 2022).

We relied on three datasets and approaches, as described below, to answer the research questions proposed in this study about the classification of the Colombian avocado germplasm given the classical races (ME, GU, and WI), as well as the presence and evolutionary origin of potential new genetic groups.

## Unsupervised non-parametric genetic clustering

Population stratification was characterized using dataset 1, which included 4931 SNPs identified in 199 samples (Table 1), using two complementary non-parametric approaches from the unsupervised machine learning paradigm: partitioning and hierarchical clustering. We had to perform a preliminary step to reduce the dimensionality of the SNP dataset through principal component analysis (PCA) (Alhusain and Hafez 2018; Foote et al. 2019) using the *R-glPca* function in the *R-adegenet* package (Jombart and Ahmed 2011). To find the optimum number of principal components needed as input in the clustering analysis, we carried out the Tracy-Widom test (Tracy and Widom 1994; Patterson et al. 2006) to obtain the statistical significance of each component employing the *Eigenstrat* function performed in *R-AssocTest* (Wang et al. 2020) as in Foote et al. (2019).

Once dimensionality was reduced in 21 principal components (66.28% accumulated variance), we determined the optimal number of genetic pools or *K* groups for each

**Table 1** Avocado SNPs datasets generated to implement the goal-oriented analyses proposed in this study

| Number of joint datasets | Datasets included | Number of samples after filters | Number of polymorphic markers | Implemented analysis |
|---|---|---|---|---|
| 1 | CGB + RM | 199 | 4931 | Population structure/ancestry in CGB/evolutionary demographic scenarios |
| 2 | CGB + RM + O | 203 | 3899 | Phylogenetic |
| 3 | SR + Pure RM and CGB[a] | 289 | 227 | Ancestry in SR |

*CGB* Colombian germplasm bank, *SR* seedling rootstocks from commercial orchards, *RM* reference materials, *O* outgroup

[a]Pure RM and CGB genotypes were identified from dataset 1; these genotypes in the ancestry inference analysis presented an ancestry higher than 80% for each genetic cluster detected in this study (GU, ME, WI, Colombian Caribbean, and Colombian Andean)

unsupervised non-parametric clustering algorithm. Then, we conducted in the *R* software two complementary methods. First, we ran the *NbClust* (Charrad et al. 2014) algorithm, an internal measure's function that integrates 30 indexes to determine the optimal *K* value based on *ward.D2* method and the Euclidean distance, evaluating three ($K=3$) to eight ($K=8$) putative genetic pools. Second, we ran the *optCluster* (Sekula et al. 2017), an improvement of the traditional *clValid* algorithm (Brock et al. 2008) that optimizes both the *K* score, and the clustering approach based on internal stability and biological measures. In this sense, the *optCluster* function validated at once all clustering algorithms by each approach using cross-entropy and genetic algorithm methods, using weighted Spearman footrule distance (Kumar and Vassilvitskii 2010) across three ($K=3$) to eight ($K=8$) putative genetic pools. We explored the optimal number of clusters using available non-parametric algorithms, such as partitioning and hierarchical clustering. For partitioning clustering, we used *K*-means (MacQueen 1967; Lloyd 1982), Partitioning Around Mmedoids (PAM), and Clustering Large Applications (CLARA) (Kaufman and Rousseeuw 1990) algorithms. For hierarchical clustering, we ran Aglomerative Nesting (AGNES) and Divisive Analysis (DIANA) algorithms (Kaufman and Rousseeuw 1990) using the *ward.D2* method.

## Maximum likelihood phylogenetic reconstruction

The phylogenetic analysis was conducted using dataset 2, which included 3899 SNPs for 202 avocado samples (Table 1) and an outgroup (O), the close relative species *P. schiedeana*. Although this species could potentially hybridize with avocados, it behaved as an outgroup for our analyses (Ashworth and Clegg 2003). In a preliminary test (data not shown), in which we also used *Phoebe bournei* as an outgroup, *P. schiedeana* was more related to this species than avocados. Therefore, we decided to use only *P. schiedeana* as an outgroup because we could recover more SNPs. The alignment of nucleotide sequences across all target SNPs was visualized and verified in Geneious software (Kearse et al. 2012). The phylogenetic tree was reconstructed using the Maximum Likelihood (ML) method in the *PhyML* software (Guindon et al. 2010) with 1000 bootstrap replicates.

## Genetic structure and contemporary directional migration rates among avocado genetic clusters

To quantify the genetic differentiation among clusters detected in the unsupervised non-parametric genetic cluster ($K=4$) and phylogenetic ($K=5$) analyses, we conducted inferences of molecular variance (AMOVA) using dataset 1. We calculated the global *Phi* scores (like $F_{ST}$, both provide population differentiation summary statistics) using 1000 permutations in the *R-poppr* package (Kamvar et al. 2014). Pairwise $F_{ST}$ values were calculated in the *R-dartR* package (Gruber et al. 2018) to compare genetic clusters, while the genetic diversity was measured through observed (*Ho*) and expected (*He*) heterozygosity and inbreeding coefficient statistics ($F_{IS}$) calculated in the *R-SNPready* software (Granato et al. 2018). Genotypes from the CGB with geo-referenced information (Fig. 1) were used to implement Mantel correlation tests with 1000 permutations between genetic differentiation and geographic distance within and between pairs of clusters. The *R-geosphere* package (Hijmans et al. 2021) was utilized to gather geographic distances. The *R-ape* package via the *mantel.test* function was used for the explicit test. Two diagrammatic versions were drawn in the same *R* software depicting genetic distance *vs.* geographic distance (km), as well as decaying correlograms against distance classes, the latter using the *mantel.correlog* functionality also from the *R-ape* package.

Finally, we conducted a Bayesian Markov Chain Monte Carlo (MCMC) analysis on dataset 1 using the *BayesAss* software v.*3.05* (Wilson and Rannala 2003) to estimate contemporary and directional gene flow migration rates (*m*) among the four $K=4$ (CO, ME, GU, and WI) and five $K=5$ (ME, GU, WI, Colombian Andean, and Colombian Caribbean) genetic groups detected in the unsupervised non-parametric genetic clustering and phylogenetic analyses, respectively. We used the *BA3SNP* option (Mussmann et al. 2019). We set the parameters for migration rates (*m*: 0.15 in both analyses), allele frequencies (*a*: 0.3 in both analyses), and inbreeding coefficients ($F_{IS}$: 0.01 in $K=4$ and 0.005 in $K=5$) to achieve an optimal acceptance rate between 20% and 60%, as recommended in the manual. We ran three independent analyses for each *K* using different seed values (100, 200, and 300) evaluated in 10 M iterations and 1 M burn-in, sampling every 1000 steps. We checked the convergence of the results and calculated the mean, standard deviation, and 95% Confidence Intervals (CI) of gene flow over the three independent runs, using the *Tracer* v.*1.7.2* software (Rambaut et al. 2018). We considered gene flow migration rates with CI over 0 as significant. We drew the pairwise $F_{ST}$ and *m* results using the *R-qgraph* package (Epskamp et al. 2012).

## Ancestry inferences

To determine which groups of avocados are spanned by the CGB and SR genotypes, we ran two independent ancestry analyses using datasets 1 and 3, because each matrix presented a different number of SNPs, 4931 and 227, respectively (Table 1). These datasets included genotypes with high genetic ancestry (> 80%) to one of the genetic groups (ME, GU, WI, and CO, the Colombian group) determined in the unsupervised non-parametric genetic cluster analysis. In the case of the CO group, we selected samples representing

the two possible genetic sub-populations detected in the Colombian collection (Andean and Caribbean) through phylogenetic analysis. We used samples with a clear genetic ancestry for these five genetic clusters as reference genotypes to implement a supervised analysis in *ADMIXTURE* software (Alexander and Lange 2011). These analyses were run for a *K* of five (ME, GU, WI, Colombian Andean, and Colombian Caribbean) using a cross-validation (CV) of 10. The CGB and SR samples with > 80% ancestry were assigned to a specific genetic cluster, and mixed samples (≤ 80%) were cataloged as admixtures/hybrids based on their admixture proportion (*e.g.*, Colombian Andean × Colombian Caribbean).

## Reconstruction of evolutionary demographic scenarios

We implemented Approximate Bayesian Computation (ABC) using the *DIYABC Random Forest v.1.0* software (Collin et al. 2021), an extended version of *DIYABC v.2.1.0*. This tree-based classification method uses supervised machine learning to classify evolutionary relationship scenarios and estimate their parameter robustness based on backward simulations. With that in mind, we simulated 120,000 sets for the 4931 selected loci of dataset 1 (Table 1), 10,000 trees per scenario, and 1000 as the number of out-of-band testing samples to estimate the historical parameters given the scenario with the highest posterior probability. Because there is no experimental estimate of the SNP mutation rate available for avocados, the mean mutation rate followed an *a priori* uniform distribution. Each locus had a possible range of two allelic states, as expected for bi-allelic SNPs in a diploid species. Moreover, we set the divergence time as the number of generations to the most recent common ancestor as equal to or higher than in this order: $t_3 \geq t_2 \geq t_1$. First, we considered five groups: Mexican (ME), Guatemalan (GU), West Indian (WI), and the two identified Colombian groups (Andean and Caribbean). However, trial runs never showed convergence (data not included). Therefore, we simplified the analysis using previous parameters into four groups: the Mesoamerican groups (i.e., ME, GU, and WI) and the Colombian group (CO) with accessions from the two subgroups (*i.e.*, Andean and Caribbean). This strategy showed convergence.

For these four genetic groups, we constructed 18 different demographic scenarios, grouping them into four families of hypotheses. The null hypothesis (scenario 1) was that all four groups (ME, GU, WI, and CO) diverged at the same time. In contrast, the other three families of hypotheses tested if CO split early, late, or nested compared with the other three recognized groups. The first family of hypotheses (nine scenarios, named from scenarios 2 to 10) reconstructed an early divergence of CO compared to ME, GU, and WI. The second family of hypotheses (three scenarios, named from scenarios 11 to 13) reconstructed a demographic history where CO diverged in a nested manner compared to ME, GU, and WI. Finally, the third family of hypotheses (five scenarios, from 14 to 18) reconstructed a demographic history where CO diverged later compared with ME, GU, and WI groups (Fig. S1). These scenarios reflect a uniform distribution for the possible genealogies. In other words, instead of only inputting likely topologies based on the results from previous archeological reports and genetic outputs from this work, we opted to embrace a broader spectrum of *prior* genealogies. Additionally, rather than recovering a single optimum scenario, we identified a family of probable genealogies with comparable *posterior* probabilities. This strategy is an updated utilization of the Bayesian thinking, which can concurrently handle multiple genealogies with similar likelihoods without limiting itself to a parametric singularity of the posterior distribution.

After selecting the three best scenarios according to the number of votes, we converted from the number of generations to years the resultant posterior distributions for the divergence times $t_1$, $t_2$, and $t_3$ (*i.e.*, summarized by mean, standard deviation, and CI at 0.05 and 0.95). This transformation relied on the generation time ($g$), the expected average time in years between two consecutive generations, which correlates with the onset of the reproductive phase. Since the avocado species is a perennial woody tree crop (Miller and Gross 2011) with a variable reproductive phase depending on environmental and genetic factors, the conversion from the number of generations to absolute time was not trivial and had to acknowledge multiple expectations of the generation time $g$. We utilized three different calibration points to account for this variability in generation times across studies. A conservative generation time of $g = 10$ years represented the earliest age of fruiting observed among 11 pairwise cultivar crosses (*i.e.*, ranging between three and 14 years) (Thorp et al. 2015). However, we also utilized generation times of $g = 30$ years and $g = 50$ years because avocado trees in nature may take longer to reach sufficient fecundity to produce enough viable fruits and offspring (Krome 1956; Goodall et al. 1970; Zuazo et al. 2021). Ultimately, these multiple conversion points enabled considering uncertainty thresholds for explicit comparisons with geological profiles and archeological records.

We finally compared the timing of divergence between pairs of lineages across the best models with two hypotheses on the evolution and genetic diversification of avocados, both inspired by archeological data and Spanish chroniclers' descriptions (Galindo-Tovar et al. 2007; Galindo-Tovar and Arzate-Fernández 2010). The subgenus *Persea* probably diversified as a product of available

habitats in the Pleistocene (*i.e.*, from 2,580,000 to 11,700 YBP) (Galindo-Tovar et al. 2007). If so, avocado dispersion and diversification would have to be mediated by big mammals and hunter-gatherers humans as they migrated from Mesoamerica to northern South America through Central America and the Isthmus of Panama (Galindo-Tovar et al. 2007). In contrast, if the dispersal and diversification of avocados occurred more recently during agriculture and village life, we would expect divergence times during the Holocene, the most recent period of the Quaternary that began approximately 11,650 YBP until today (Krome 1956; Goodall et al. 1970; Galindo-Tovar et al. 2007; Miller and Gross 2011; Thorp et al. 2015; Zuazo et al. 2021).

## Results

### SNP discovery in avocado panels

The sequencing of 384 avocado genotypes from CGB and SR generated over 3,132,393,704 sequence reads (Table S1), which were combined with available sequences from 71 reference genotypes (Talavera et al. 2019), and the outgroup, *P. schiedeana* (Rendón-Anaya et al. 2019). In CBG and SR genotypes, the number of DNA sequences obtained was similar from leaf (CGB-3 M to 52 M) and root (SR-1.3 M to 56 M) tissues (Table S1). However, regarding the SNP calling, the matrices obtained from leaf DNA sequences had the highest number of SNPs. In contrast, matrices including sequences from root tissues had many missing data and a low number of polymorphisms. Therefore, creating a single dataset was inviable due to significant differences in SNP numbers between leaf and root tissue sequences. We then generated three different SNP datasets depending on the goal of each analysis, each including different samples (CGB, SR, RM, and O). The SNPs numbers ranged from 227 in dataset 3, which included 289 genotypes from SR and CGB, as well as RM materials from ME, WI, GU, Colombian Andean, and Colombian Caribbean groups, with high ancestry (> 80%) for each genetic cluster, to 4931 SNPs in dataset 1, which consisted of 199 genotypes from CGB and RM genotypes (Table 1).

The global SNP summary statistics were calculated using dataset 1, which had the highest SNP number (4931) mapped in 8082 contigs of the Hass genome reference assembly. This dataset had a *Ts/Tv* proportion of 1.54, with 2988 *Ts* and 1943 *Tv* substitutions. These SNPs presented an *MDS* mean of 55, ranging from 15 (Ctg0042_632283 SNP) to 245 (Ctg1142_29090); while the nucleotide diversity (*Pi*) presented a mean of 0.275 ranging from 0.064 (Ctg0037_1227237) to 0.500 (Ctg0882_23451). On the other hand, the polymorphisms had a mean *MAF* of 0.255, ranging from 0.100 (Ctg1752_17921) to 0.500 (Ctg0066_1021952), and the SNPs were highly informative, with a mean *PIC* of 0.283, ranging between 0.160 (Ctg1752_17921) and 0.380 (Ctg0066_1021952).

### Avocado genetic stratification was consistent with four genetic groups

The two unsupervised non-parametric clustering approaches consistently showed four clusters as genetic groups (Table S2 and Table S3). Initially, we obtained three principal components, where the first component explained 26.77%, the second component explained 12.98%, and the third component explained 6.93% of the variance from filtered SNPs. The percentage of variation explained by the first three principal components is a typical value in allogamous species for dimensionality reduction analysis using PCA from SNPs, as demonstrated in the initial study by Talavera et al. (2019), who employed the same dimensionality reduction algorithm as ours, obtaining comparable partitions across components. However, we performed the unsupervised training analysis using all the first 21 principal components suggested by the Tracy-Widom test, which accounts for 66.28% of the total variance, reserving only the first two components for 2D visualization of the clustering results. Based on the first 21 principal components, the *NbClust* function suggested four optimal clusters by running the *ward.D2* method (Table S2).

Meanwhile, independent of the clustering approach or validation method as cross-entropy and genetic algorithms, the *optCluster* function recurrently suggested four optimal clusters. PAM was the best partitional clustering method (*Cross-Entropy score*: 25.82, *Genetic Algorithm score*: 25.50), and AGNES was the best hierarchical method (*Cross-Entropy score*: 20.82, *Genetic Algorithm score*: 20.82). Therefore, the *optCluster* function suggested PAM as the best clustering algorithm compared to AGNES.

The avocado genotypes were grouped into four clusters that we identified based on available passport data as follows: GU cluster ($n = 43$), CO cluster ($n = 79$), ME cluster ($n = 19$), and WI cluster ($n = 58$) (Fig. 2). The PCA biplot showed the CO group separated in the first component, while the other three clusters consisted of ME, GU, and WI groups from Talavera et al. (2019). The CO cluster was exclusively composed of Colombian native accessions, while the other three groups included RM genotypes with complete ancestry of GU, ME, and WI groups, as well as RM genotypes with admixtures between Mesoamerican groups (*e.g.*, GU × WI, GU × ME, and ME × WI), commercial materials (CM), and

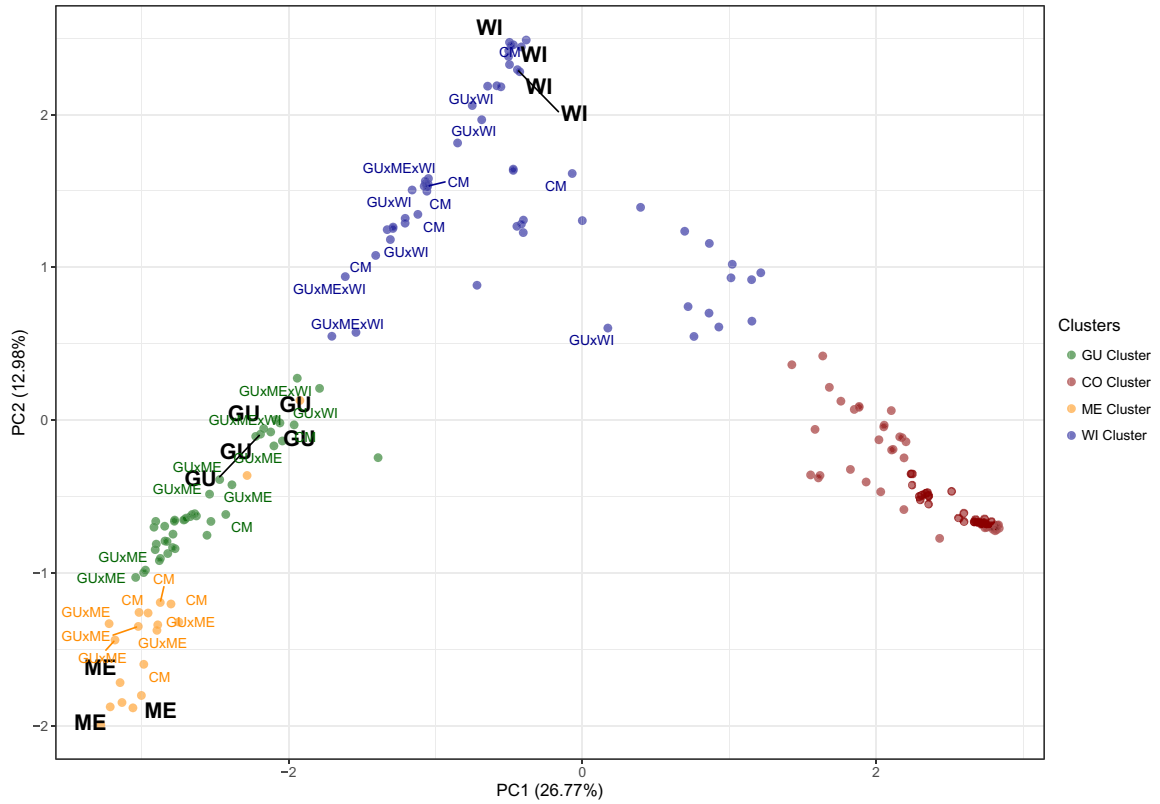



**Fig. 2** Principal component analysis (PCA) biplot showing the genetic diversity of 199 avocado genotypes of CGB and RM analyzed with 4931 SNPs. The colors indicate the assignment to a genetic group based on the Partitional Around Medoids (PAM). Abbreviations are as follows: Guatemalan, GU cluster; Mexican, ME cluster; West Indian, WI cluster; and Colombian, CO cluster. The black letters indicate the three recognized avocado groups (GU: Guatemalan, ME: Mexican, and WI: West Indian) based on the results of Talavera et al. (2019). The varieties or cultivars conserved in CGB are marked as CM (Commercial Materials)

accessions conserved in the CGB with high ancestry of ME ("Bacon", "Duke_7", "Fuerte", "Zutano", "Topa Topa", and VAPA_066), GU ("Edranol", "Hass", "Reed", "Simmonds", QUIQUI_029, and TACO_047), and WI ("Booth_5", "Booth_7", "Booth_8", "Choquete", "Lula", "Pollock", "Waldin", and other 31 Colombian CGB accessions (Fig. 2 and Table S3).

## Phylogenetic analysis supported stratification across four major genetic pools and suggested a substructure in the Colombian genetic group

A phylogenetic tree was constructed using the SNP alignment data of 203 avocado genotypes. The tree consisted of four well-supported clades corresponding to the reference groups and a new clade comprising mainly Colombian genotypes (CO group) from the Colombian germplasm bank (Fig. S2). However, some accessions did not fall within a single clade. Based on the clustering analysis, they corresponded to hybrids or admixtures. Thus, we created a new phylogenetic analysis excluding those hybrids or admixtures, with 92 samples and 3899 SNPs (Fig. 3). The study confirmed four clades with bootstrap support (BS) of 100% that match the clustering analyses (Fig. 1). The clade that first diverged corresponded to samples belonging to the ME group. The accessions from the GU group formed the next clade in position, followed by the clade with the WI materials. The CO and WI groups were related in two supported clades in the most derived phylogeny positions. The CO clade was divided into two subgroups. Based on passport data, the avocado genotypes mainly from the Andean region (Colombian Andean: Cauca, Huila, Quindio, Risaralda, Santander, Meta, and Tolima provinces) were regrouped in a well-supported clade (BS = 100%), while a second weaker-supported clade (BS = 43%) regrouped samples from the Caribbean region (Colombian Caribbean: Bolivar, Cesar, and Magdalena provinces). From this filtered dataset (without hybrids), seven accessions from the Colombian germplasm fell in the Mesoamerican clades, four genotypes ("Waldin", QUIBU_11, VADU_10, and VADU_14) in the WI clade, one ("Hass") in the GU clade, and two ("Topa Topa", and "Duke_7") in the ME clade (Fig. 3).

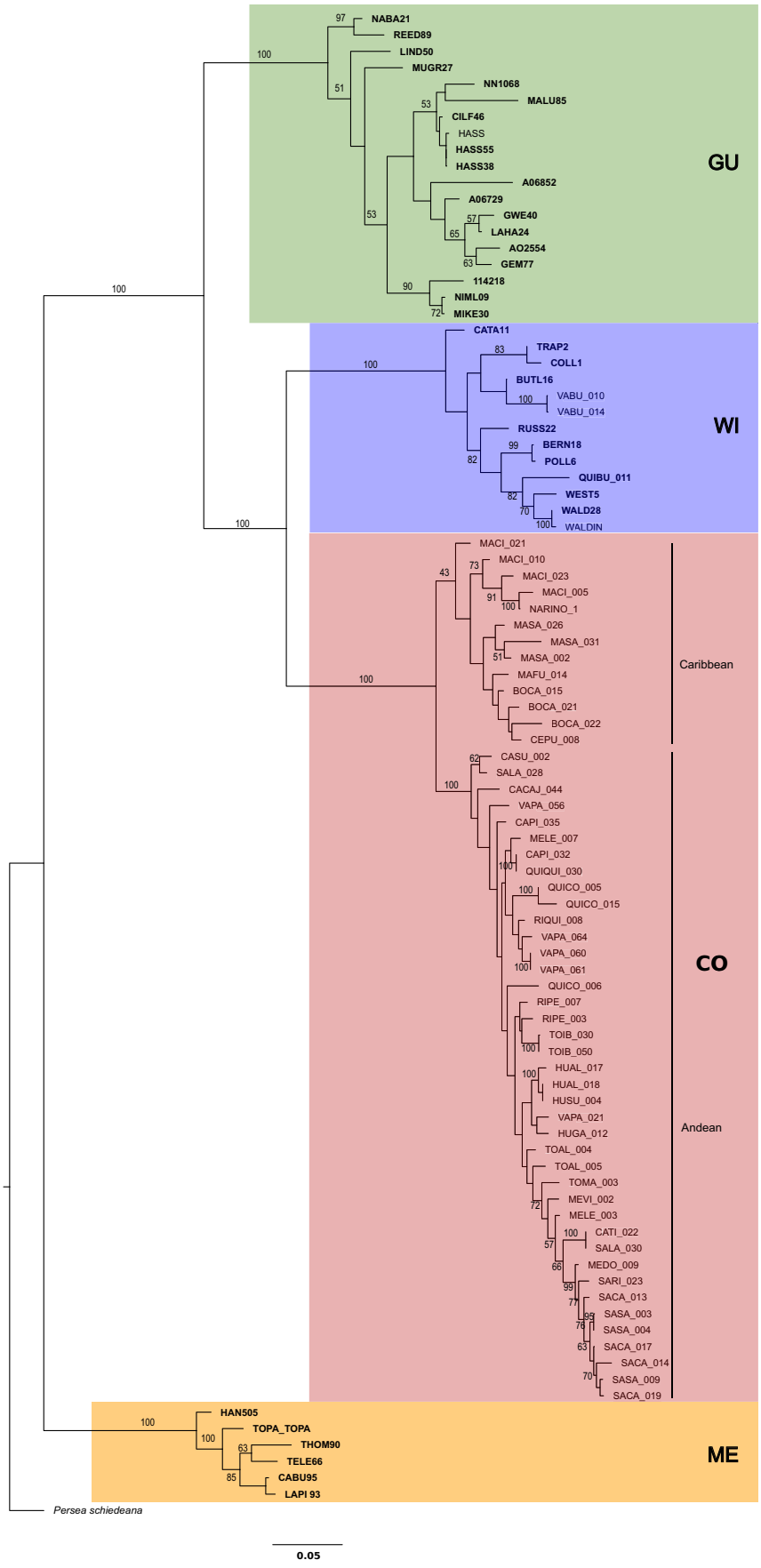



**Fig. 3** Maximum likelihood phylogenetic tree showing the relationships with 3899 SNPs among 91 avocado accessions (> 80% of ancestry for each genetic group) from CGB (Colombian Andean and Colombian Caribbean), Reference Materials (RM) of Mexican (ME), Guatemalan (GU), and West Indian (WI) ancestry, and the *P. schiedeana* outgroup. The values on nodes are the percentage of the bootstrap values for 1000 replicates. Abbreviations are as follows: ME: Mexican, GU: Guatemalan, WI: West Indian group, and CO: Colombian genetic groups. Andean: CGB genotypes from the Andean region of Colombia, Caribbean: CGB genotypes from the Caribbean region of Colombia

## Genetic divergence was concordant with botanical and spatial differentiation

The AMOVAs for the $K=4$ and $K=5$ analyses showed that most of the genetic variation was found among genetic groups ($K=4$: 54.86% and $K=5$: 55.10%), with 45.13% ($K=4$) and 44.89% ($K=5$) of variation within the clusters. This result indicated a solid genetic structure and significant genetic differentiation ($p=0.001$) between the different avocado genetic groups detected in $K=4$ (*Phi*: 0.548) and $K=5$ (*Phi*: 0.551). The pairwise $F_{ST}$ values indicated that the CO group had significant ($p=0.001$) genetic differentiation from the ME ($F_{ST}$ of 0.49), GU ($F_{ST}$ of 0.47), and WI ($F_{ST}$ of 0.30) groups in the $K=4$ analysis (Fig. 4a). The Andean and Caribbean Colombian groups also showed genetic differences ($F_{ST}$ of 0.14), but lower compared to the differences between the Colombian groups against ME (pairwise $F_{ST}$ of 0.48 and 0.50 for Caribbean and Andean), GU (pairwise $F_{ST}$ of 0.46 and 0.48 for Caribbean and Andean), and WI (pairwise $F_{ST}$ of 0.27 and 0.31 for Caribbean and Andean) groups in the $K=5$ analysis (Fig. 4c). In both analyses using $K=4$ and $K=5$, the genetic differences among the Mesoamerican groups were intermediate and presented values between 0.21 (ME vs. GU) and 0.30 (WI vs. ME) (Fig. 4a, c). These results suggest a strong genetic structure and differentiation between the avocado genetic groups, particularly the Colombian and Mesoamerican groups.

The gene flow according to directional migration rates ($m$) calculated for $K=4$ (Fig. 4b) and $K=5$ (Fig. 4d) in BayesAss converged and did not show differences between the several simulations implemented for each dataset. The results suggested that gene flow among the different genetic groups of avocados was generally low ($<0.1$), except for the directional gene flow between the ME and GU groups (ME → GU: 0.15 in $K=4$ and 0.16 in $K=5$). Also, some low but significant directional gene flow was detected between ME and WI (ME → WI: 0.04 in both $K$ analyses), WI and Colombian Caribbean (WI → Caribbean: 0.02 in $K=5$), and Colombian Andean and Caribbean (Andean → Caribbean: 0.03 in $K=5$). Additionally, significant bidirectional gene flow scores were detected between WI and GU (GU → WI: 0.02 and WI → GU: 0.05 in both $K$ analyses), as well as between WI and CO (WI → CO: 0.08 and CO → WI: 0.01 in $K=4$), and WI and Colombian Andean (WI → Andean:

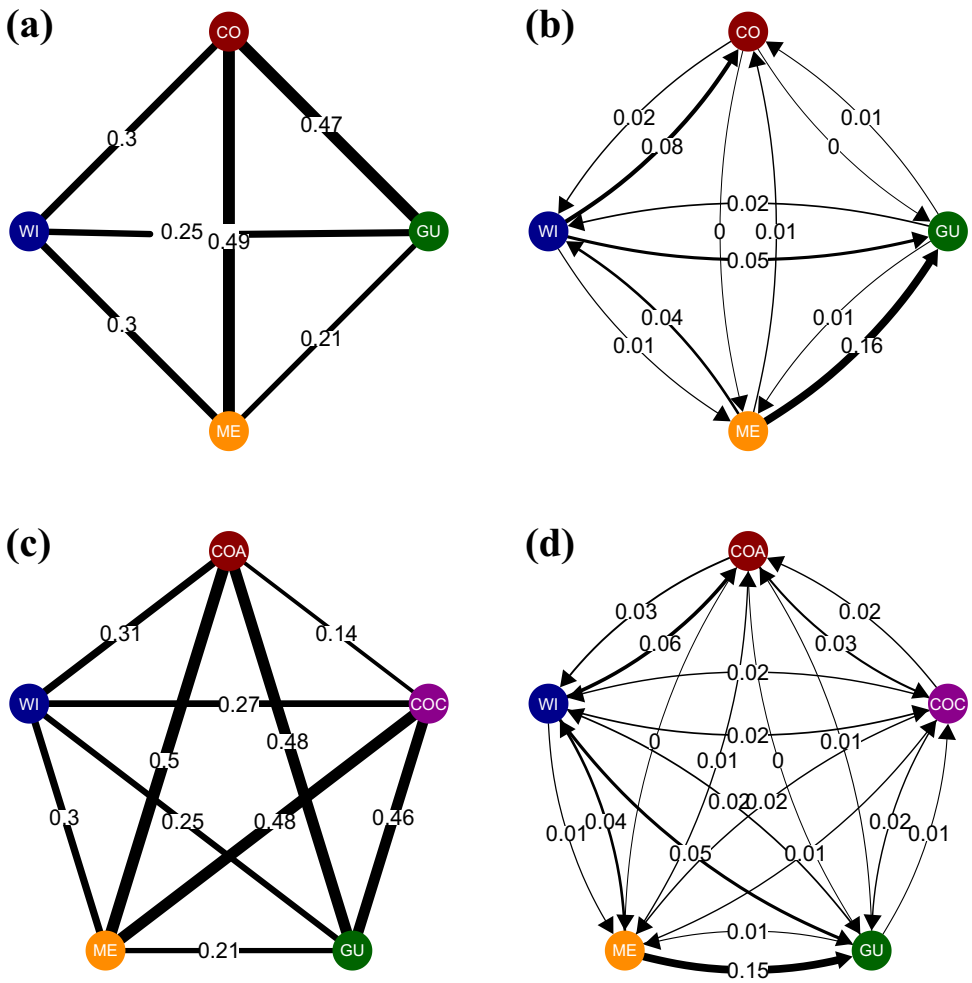



**Fig. 4** Pairwise $F_{ST}$ **a, c** and contemporary migration rates $m$ **b, d** values among the four ($K=4$) and five ($K=5$) avocado genetic clusters determined by population genetic structure and phylogenetic analysis. In the Figure, GU corresponds to the Guatemalan genetic group, CO corresponds to the Colombian group, ME corresponds to the Mexican genetic group, WI corresponds to the West Indian genetic group, COA corresponds to the Colombian Andean genetic group, and COC corresponds to the Colombian Caribbean genetic group

0.06 and Andean → WI: 0.02 in $K=5$) (Fig. 4b and d). These results indicate that gene flow among different avocado genetic groups is limited, and there is some degree of isolation and differentiation between these groups.

In addition, we explored the genomic geographic divergence within the Colombian sub-populations, and their relationship with the WI group, using the samples with geo-referenced information. We implemented explicit correlations of genetic distance with geographic gradients via Mantel tests between pairs of clusters. Here, we were able to recover a significant ($p<0.05$) trend of intra-group isolation by distance (IBD) only when admixed individuals were included, as expected under rampant mobility (Fig. S3 and Fig. S4). However, the IBD trend vanished when the analysis was limited to the purified WI and Colombian Andean and Caribbean clusters. This result may suggest that the same underlying process that helps shape admixture could boost regionalized geographic structure.

The overall population of avocados exhibited intermediate levels of *Ho* (0.22) and *He* (0.27), with a positive value of $F_{IS}$ (0.19). When comparing the genetic groups identified in the $K=4$ and $K=5$ analyses, the Mesoamerican groups WI (*Ho:* 0.28 and *He:* 0.24), GU (*Ho:* 0.24 and *He:* 0.20), and ME (*Ho:* 0.22 and *He:*0.19) were generally more heterozygous than the Colombian group (CO) in the $K=4$ analysis (*Ho:* 0.18 and *He:* 0.16). In the $K=5$ analysis, Colombian Andean (*Ho:* 0.18 and *He:* 0.16) and Caribbean (*Ho:* 0.18 and *He:* 0.15) subgroups exhibited similar levels of heterozygosity. Negative $F_{IS}$ values were observed for all genetic groups; however, the CO group had higher negative $F_{IS}$ values (−0.08 in $K=4$) compared to the GU (−0.18), WI (−0.14), and ME (−0.12) genetic groups. Furthermore, the $K=5$ analysis identified distinct $F_{IS}$ values between the Colombian Andean (−0.13) and Caribbean (−0.20) subgroups (Table 2).

## Limited Mesoamerican genetic ancestry in the Colombian germplasm supported the presence of a new avocado group in Colombia

The CGB conserves avocado genotypes with ancestry from all avocado genetic groups, with 55.8% having complete ancestry (pure genotypes) to the five clusters recovered in previous analyses. Most CGB genotypes had genetic ancestry from Andean (34.1%) and Caribbean (13.2%) groups native to Colombia. A minor proportion (8.5%) of CGB genotypes presented pure ancestry of WI (5.4%: "Pollock", "Waldin", QUIBU_011, QUIPI_001, QUIPI_003, VABU_010, and VABU_014), ME (1.6%: "Duke_7" and "Topa Topa"), and GU (1.6%: "Reed" and "Hass"). The other 44.2% of the CGB genotypes were admixed or hybrids with diverse admixture patterns (18 possible combinations of hybrid backgrounds). The highest proportion of avocado hybrids had mixed ancestry from WI × GU (10.1%), WI × Colombian Andean (4.7%), and WI × GU × Colombian Andean (3.1%). Meanwhile, NATU_001 and CANO_008, some of the few accessions from the CGB reported as tolerant to *P. cinnamomi* (Rodríguez-Henao et al. 2017), were admixed (53% WI × 38% Colombian Andean × 8% Colombian Caribbean for NATU_001, and 63% WI × 22% GU × 14% Colombian Andean for CANO_008 (Fig. 5a, Table S3, and Table S4). On the contrary, in the SR, most genotypes (75.7%) had a pure ancestry of all genetic groups except the ME group. The genotypes with pure genetic ancestry were distributed in the Colombian Caribbean (37.7%), Andean (11.3%), GU (23.4%), and WI (3.5%) groups. The remaining genotypes were admixed (24.24%), the majority having a Colombian ancestry of Caribbean × Andean (13.4%) and Andean × Caribbean (4.3%) (Fig. 5b and Table S4).

**Table 2** Genetic diversity summary statistics (mean and range) in the overall avocado population and in the four and five genetic clusters detected in this study

| Analyses | Genetic groups | Number of samples[a] | *Ho* | *He* | $F_{IS}$ |
|---|---|---|---|---|---|
| $K=4$ | GU | 43 | 0.24 (0.10:0.35) | 0.20 (0:0.50) | −0.18 (−0.74:0.51) |
| | ME | 19 | 0.22 (0.10:0.32) | 0.19 (0:0.50) | −0.12 (−0.65:0.46) |
| | WI | 58 | 0.28 (0.14:0.37) | 0.24 (0:0.50) | −0.14 (−0.53:0.41) |
| | CO | 79 | 0.18 (0.09:0.33) | 0.16 (0:0.50) | −0.08 (−0.97:0.46) |
| | Total | 199 | 0.22 (0.09:0.37) | 0.27 (0.06:0.50) | 0.19 (−0.36:0.68) |
| $K=5$ | GU | 43 | 0.24 (0.10:0.35) | 0.20 (0:0.50) | −0.18 (−0.74:0.51) |
| | ME | 19 | 0.22 (0.10:0.32) | 0.19 (0:0.50) | −0.12 (−0.65:0.46) |
| | WI | 58 | 0.28 (0.14:0.37) | 0.24 (0:0.50) | −0.14 (−0.53:0.41) |
| | Caribbean | 17 | 0.18 (0.09:0.25) | 0.15 (0:0.50) | −0.20 (−0.66:0.41) |
| | Andean | 62 | 0.18 (0.11:0.33) | 0.16 (0:0.50) | −0.13 (−1.07:0.28) |
| | Total | 199 | 0.22 (0.09:0.37) | 0.27 (0.06:0.50) | 0.19 (−0.36:0.68) |

*He* expected heterozygosity, *Ho* observed heterozygosity, $F_{IS}$ inbreeding coefficient

[a]Number of samples clustered by PAM ($K=4$), and ML and Admixture ($K=5$) approaches

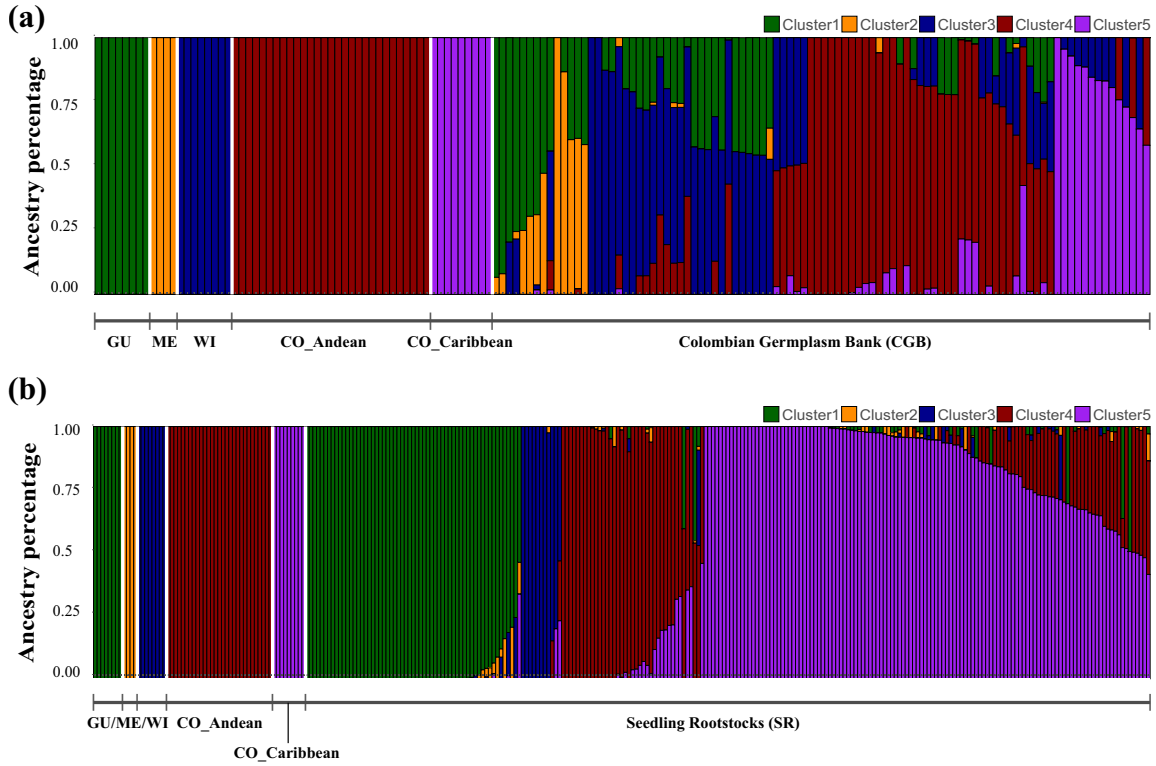


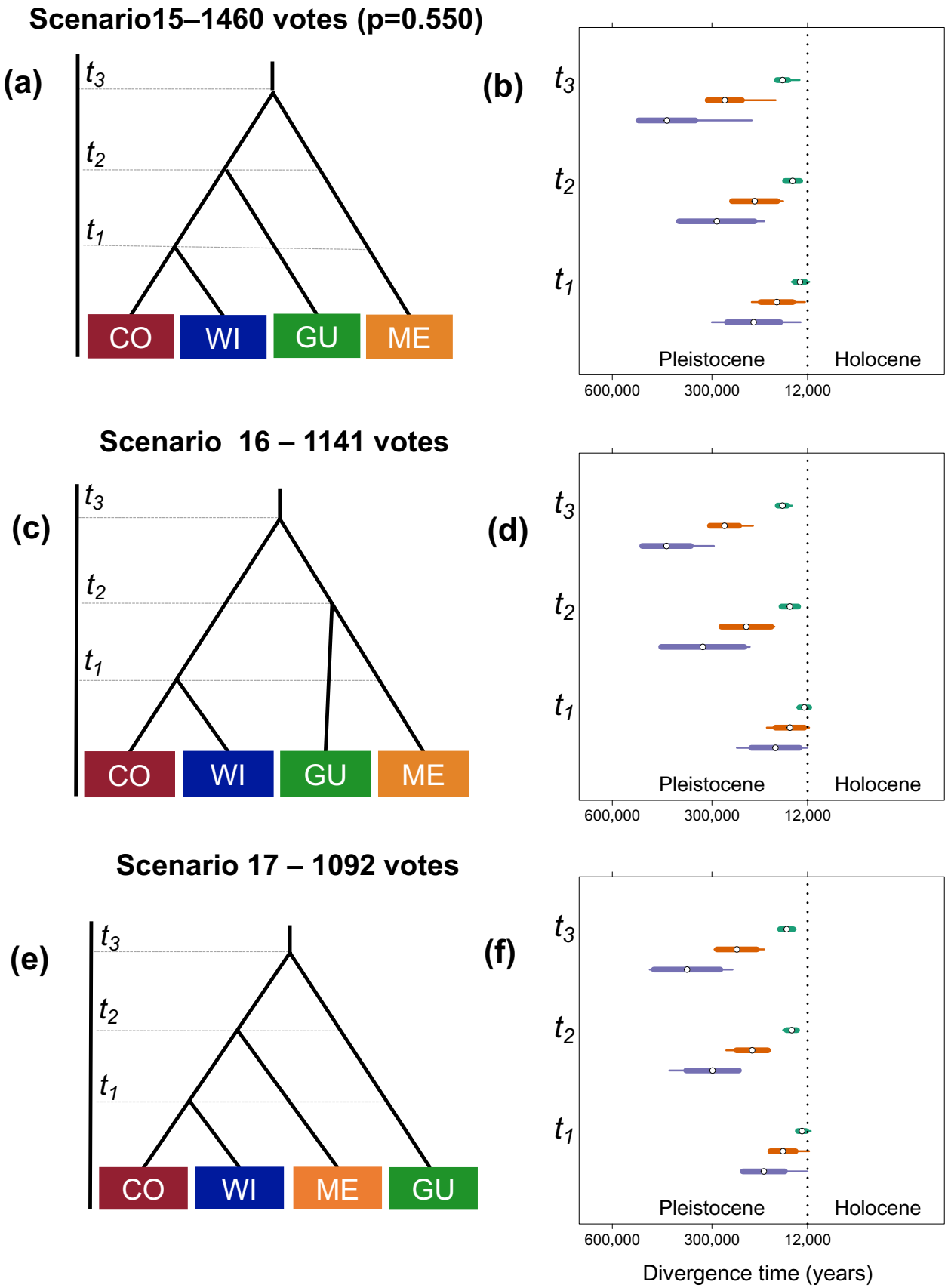

**Fig. 5** Barplot ancestry analysis results implemented in the CGB and SR germplasm using genotypes with complete ancestry backgrounds of Mexican (ME), Guatemalan (GU), West Indian (WI), Colombian Andean, and Colombian Caribbean Colombian (CO) genetic groups. **a** CGB ancestry analysis barplot. **b** SR ancestry analysis barplot

## The Colombian group diverged during the Pleistocene after the ME, GU, and WI splits

We used the DIYABC random forest method to test the best scenario of evolutionary genealogical relationships among groups explaining the evolution of the Colombian group (CO) against the other groups (*i.e.*, ME, GU, and WI) (Fig. S1). From all 18 scenarios, scenario 15 fitted best our data with 1460 votes out of 10,000 random trees and a posterior probability of 0.550. The topology of this scenario is the same as the phylogenetic analyses (Fig. 6a and Fig. S2). The subsequent two best scenarios were 16 and 17, with 1141 votes and 1092 votes, respectively (Fig. 6b, c). All three scenarios corresponded to the hypothesis of later divergence for CO compared to the ME, GU, and WI groups. The best three scenarios showed that CO and WI are sister clades with a recent common ancestor between them. The difference among the three best scenarios was whether ME or WI had the earliest divergence (Fig. 6a, c, d).

According to the best three scenarios with the higher number of votes, all four genetic groups diverged during the Pleistocene regardless of the generation time ($g$) used as calibration. This result supports that divergence is ancient and occurred before the Holocene when agriculture and village life started in the Americas. Thus, in the best scenario (scenario 15 with 1460 votes and $p = 0.550$), the median divergence time of the ME group ($t_3$) occurred between 87,079 YBP (95% CI: 36,020–95,300 at $g = 10$) and 435,393 YBP (95% CI: 180,100–476,502 at $g = 50$). Then, the median divergence time of the GU group ($t_2$) occurred between 57,070 YBP (95% CI: 28,490–80,680 at $g = 10$) and 285,351 YBP (95% CI: 142,450–403,398 at $g = 50$). Finally, the later median divergence time between WI and CO ($t_1$) occurred between 34,879 YBP (95% CI: 6570–60,058 at $g = 10$) and 174,393 YBP (95% CI: 32,850–300,291 at $g = 50$) (Fig. 6b).

In the case of the second-best scenario (scenario 16 with 1,141 votes), the oldest median divergence time between two ancestral avocado clades ($t_3$) occurred between 87,351 YBP (95% CI: 58,790–95,828 at $g = 10$) and 436,755 YBP (95%: 293,950–479,139 at $g = 50$). Moreover, the divergence between ME and GU ($t_2$) occurred between 65,500 YBP (95% CI:37,220–86,756 at $g = 10$) and 327,499 YBP (95% CI: 186,100–433,781 at $g = 50$). In parallel, the divergence between the CO and WI ($t_1$) occurred between 21,746 YBP (95% CI: 2,440–44,988 at $g = 10$) and 108,732 YBP (95% CI: 12,200–224,941 at $g = 50$) (Fig. 6d).


Scenario15–1460 votes (p=0.550)
(a)
t3
t2
t1
CO
WI
GU
ME
(b)
t3
t2
t1
Pleistocene
Holocene
600,000
300,000
12,000
Scenario 16 – 1141 votes
(c)
t3
t2
t1
CO
WI
GU
ME
(d)
t3
t2
t1
Pleistocene
Holocene
600,000
300,000
12,000
Scenario 17 – 1092 votes
(e)
t3
t2
t1
CO
WI
ME
GU
(f)
t3
t2
t1
Pleistocene
Holocene
600,000
300,000
12,000
Divergence time (years)

◂**Fig. 6** Evolutionary relationships of avocado groups reconstructed by the DIYABC random forest methodology. **a** The most probable evolutionary scenario (scenario 15) of avocados in Mesoamerica and northern South America showed 1460 votes and a posterior probability of 0.550. This scenario suggests an earlier divergence of the Mexican (ME) group at $t_3$, followed by the Guatemalan (GU) group divergence at $t_2$ and a later divergence between the Colombian group (CO) and the West Indian (WI) group at $t_1$. **b** The second-most probable scenario (Scenario 16) had 1141 votes. This scenario suggests a parallel evolution of two ancient clades that diverged at $t_3$. One clade split at $t_2$, originating ME and GU. The other clade split at $t_1$, forming CO and WI. (**c**) The third most probable scenario (Scenario 17) had 1092 votes. This scenario suggests an earlier divergence of the GU at $t_3$, followed by the ME divergence at $t_2$ and a later divergence between the CO and WI at $t_1$. The divergence times are not scaled. **d**–**f** The divergence time in years of the best three scenarios, 15, 16, and 17, respectively, assuming 10 years (green), 30 years (orange), and 50 years (violet) of generation time ($g$). The white circles represent the mean estimates, the thick hash marks indicate the standard deviation, and the thin hash marks indicate the confidence interval at 95%. The gray vertical dot line separates the Pleistocene and the Holocene within the Quaternary era. The Pleistocene started 2,000,000 years ago and ended 12,000 years ago. At that time, the Holocene started until the present

Finally, in the third-best scenario (scenario 17 with 1092 votes), the median divergence time of the GU group ($t_3$) occurred between 74,968 YBP (95% CI: 47,570–97,945 at $g = 10$) and 374,840 YBP (95% CI: 327,850–487,474 at $g = 50$). Then, the median divergence time of the ME group ($t_2$) occurred between 59,626 YBP (95% CI: 46,240–85,654 at $g = 10$) and 298,130 YBP (95% CI: 231,200–428,268 at $g = 50$). Finally, the most recent median divergence time between WI and CO ($t_1$) occurred between 28,726 YBP (95% CI: 2420–37,120 at $g = 10$) and 143,631 YBP (95% CI: 12,100–185,599 at $g = 50$) (Fig. 6f).

## Discussion

Our study analyzed two collections of Colombian avocado genotypes: (1) the Colombian Germplasm Bank (CGB), which contains both native avocados and commercial varieties, and (2) seedling rootstocks (SR) from commercial orchards in the Antioquia province, a major Colombian avocado-growing region. The comparison of genome-reduced sequences of the Colombian germplasm with avocados from Mesoamerica, including the Mexican (ME), Guatemalan (GU), and West Indian (WI) groups, as well as hybrids among these, reveals that the majority of CGB and SR accessions belong to a new Colombian genetic group (CO), divided into Andean and Caribbean sub-populations. While some CGB and SR accessions belonged to the classical Mesoamerican groups, this discovery highlights untapped genetic diversity in northwest South America that could be highly beneficial for tree crop improvement.

### Enabling genomic resources for avocado rootstocks

This study does not report a unique marker dataset due to significant differences in SNP coverage for sequences gathered from leaf and root tissues. Consequently, we had to split the reference mapping and SNP calling steps by tissue type among sequences to generate three datasets for each research question. Regardless of the final number of SNPs in each dataset, these variants enable us to approach the research questions at different levels of resolution. The efficiency of nucleic acid extraction processes varies depending on the tissue's properties, such as physical characteristics (*e.g.*, lignification level) and chemical composition (*e.g.*, presence of secondary metabolites). Therefore, when utilizing root tissue for genomic approaches, it is recommended to sample tissues with consistent and optimized conditions (*e.g.*, dryness and size) (Fisk et al. 2010; Zeng et al. 2015; Oliveira et al. 2015; Miller et al. 2017). Alternatively, stooling or layering (Knight et al. 1927; Webster 1995) could also be employed to facilitate the collection of leaf tissue from rootstocks. These propagation approaches could convey more effective DNA extraction methods for avocado rootstocks by inducing leaf tissue formation rather than exclusively relying on the root system.

### The Colombian avocado germplasm does not belong to the classical Mesoamerican groups but is consistent with a new gene pool of *P. americana*

All analyses implemented in this study reinforce the well-described differentiation among the classical Mesoamerican groups (ME, GU, and WI), where ME and GU are genetically closer than WI (Rubinstein et al. 2019; Ge et al. 2019b; Talavera et al. 2019; Wienk et al. 2022; Solares et al. 2023). However, our inferences went one step further. We demonstrated a consistent and robust genetic differentiation of the previously reported Mesoamerican groups from a new genetic group composed exclusively of genotypes native to northwest South America, *i.e.*, the Colombian (CO) group. This novel genetic group further exhibits a genetic substructure associated with the eco-geographic origin, allowing the differentiation of avocado genotypes from Colombia's Andean and Caribbean regions. Cañas-Gutierrez et al. (2019) also report signals of population structure between avocados from the North and the South of Colombia by scoring SSRs markers in samples of the CGB and some commercial cultivars from the Antioquia province. Various genetic summary statistics support this cryptic population structure. For instance, concerning gene flow patterns, the Colombian group had low gene flow with ME and GU but higher gene flow from the WI group. At the local level, these

results suggest that the gene flow in the Caribbean group comes from the Andean and WI groups. The low gene flow and significant genetic differentiation linked with geography may indicate ecological divergence, an *ad hoc* hypothesis that deserves further testing (Cañas-Gutierrez et al. 2019).

Meanwhile, another genetic summary statistic associated with different levels of population stratification was heterozygosity (*i.e.*, correlated with negative $F_{IS}$ scores). The Colombian group presents lower heterozygosity values than the Mesoamerican groups, which could reflect a bottleneck during colonization. High levels of heterozygosity in this species are expected to be reinforced by outcrossing (ranging from 74% to 96%) due to the avocado's peculiar floral biology behavior known as synchronous protogynous dichogamy. In this system, the female flower function predates the male flower, which minimizes selfing and promotes genetic admixture (Borrone et al. 2008; Solares et al. 2023). Alternatively, the positive selection of highly heterozygous individuals in specific environments may be a signature of heterosis in hybrids with mixed ancestries among groups, a pattern already reported by Reyes-Herrera et al. (2020). Finally, heterozygosity may as well vary by other exogenous forces such as tree distribution within orchards, edaphoclimatic variability, and agronomic management (Borrone et al. 2008; Sánchez-González et al. 2020). Although the underlying cause of divergence and its correlates may remain elusive, the Colombian groups reported in this study still provide new insights into avocados' phylogenetic diversity, population structure, and the potential drivers of dispersal.

Finally, this is the first report of the Colombian (CO) avocado group, probably associated with the absence of native avocado samples from Colombia in previous studies (Ge et al. 2019a, b; Rubinstein et al. 2019; Solares et al. 2023; Talavera et al. 2019; Wienk et al. 2022). As the sampling of native avocado trees across the Americas continues increasing, we may be able to identify other genetic clusters locally adapted to distinctive agroecological regions. For example, Solares et al. (2023) recently included some samples from Costa Rica, a new possible genetic group that also seems to be under differentiation from the prevalent Mesoamerican trinity. Therefore, future studies require embracing potential comprehensive centers of avocado genetic diversity beyond Mesoamerica while better clarifying the origin of such possible novel groups.

## Seedling rootstocks from commercial orchards result from rampant admixture among genetic groups with potential heterotic effects

The CGB and SR germplasm exhibited high numbers of pure genotypes with ancestries predominantly from the Colombian groups, with the Caribbean ancestry being more frequent in commercial genotypes and the Andean ancestry in the CGB collection. This result could be the effect of the distinct sampling approaches employed by the two collections. The SR primarily includes rootstocks used in commercial orchards (Cañas-Gutiérrez et al. 2019), while the CGB comprises native "criollo" genotypes, some of which may have a natural tolerance to *P. cinammomi* (Rodríguez-Henao et al. 2017). Moreover, 5% of the CGB and 23% of the SR populations display pure ancestries of the classical Mesoamerican groups. This result suggests that the seedlings used for commercial plantations also have a proportion of Mesoamerican ancestry in their genetic background. Our finding aligns with the expectation that plant genotypes utilized in these commercial plantations could result from modern introductions (Cañas-Gutiérrez et al. 2019).

Previous genetic structure reports have shown recurrent evidence of hybridization and introgression among the classical Mesoamerican groups (Rubinstein et al. 2019; Ge et al. 2019b; Talavera et al. 2019; Wienk et al. 2022; Solares et al. 2023). Therefore, it will not be surprising that if Mesoamerican ancestries were recovered from seedling rootstocks in Colombia, they could as well harbor signatures of admixture. Indeed, 44% and 24% of genotypes of the CGB and SR populations were classified as hybrids. Within the CGB, some hybrids displayed ancestries from Mesoamerican groups, such as WI × GU (Rodríguez-Henao et al. 2017). The remaining hybrids (20.2%) have backgrounds between the Colombian Andean and other groups, although hybrids with ancestries between Andean and Caribbean are relatively scarce in the CGB (4.6%). In contrast, most hybrids in the SR population have backgrounds between the Colombian Andean and Caribbean groups (18%). These findings indicate that the commercial seedlings used for grafting may result from hybridization within the Colombian germplasm and other Mesoamerican groups. This result is interesting, especially considering that current rootstock planting material in Colombia is usually gathered from seeds of commercial varieties such as "Waldin", "Duke_7", "G_755", and "Lula" as well as some Colombian genotypes like La Torre, Tumaco, and Villagornona, in addition to "criollo" native trees (Ríos-Castaño et al. 2005; Bernal and Díaz 2020). The fact that most rootstocks used in Colombia's main avocado production region have admixed origins may result in novel heterotic combinations that could assist adaptation to the country's specific agroecological conditions (a hypothesis already embraced by Reyes-Herrera et al. (2020) in the same plantations using SSR markers). Although insightful for rootstock breeding efforts, this finding does not yet provide clues on the evolutionary origin of the Colombian genotypes, a matter we fully embrace in the next section.

## The Colombian lineage of avocados diverge from the West Indian group during the Pleistocene

The best demographic model from our study (scenario 15) is concordant with the recent genome-wide phylogeny from Solares et al. (2023) and with our phylogeny. Both results suggest that the Mexican lineage diverged first. However, the subsequent best scenarios (scenarios 16 and 17) open the question about whether the Guatemalan group experienced the second most ancient divergence or if the Mexican and Guatemalan groups are both sister clades of the West Indian and Colombian groups.

Despite these undefined divergence nodes, the 95% confidence intervals of their timing suggest that the evolution of avocado, its long-distance dispersal, and possible adaptive divergence started ~ 430,000 YBP, a more recent estimate compared to the divergence time of 1.3 M YBP calculated from the wide-genome phylogeny of Solares et al. (2023). However, both calculations match the climatic fluctuations of the Pleistocene, a period from 2.5 M to 11,700 YBP. During the Pleistocene, the Central American isthmus had already uplifted completely, joining North America and South America and enabling the massive migration interchange of fauna, flora, and humans among continents (Galindo-Tovar et al. 2008). Thus, our data are consistent with long-distance dispersal mediated by big mammals such as the giant ground slots that inhabited the Americas during the Pleistocene. These species probably consumed avocados, dispersed their seeds, and generated isolation of specific populations for long periods before any human did (Diamond 1999; Barlow 2002; Chen et al. 2009).

More recently, the West Indian and the Colombian groups share a common ancestor before the onset of the Holocene (*i.e.*, 34,879 YBP in scenario 15, 21,746 YBP in scenario 16, and 28,726 YBP in scenario 17). Our results suggest that the four main groups of *P. americana* were already differentiated when the first humans arrived in the USA. During that epoch, the dispersion of avocado seeds and propagules may have happened by spontaneous growth from leftovers of the first hunter-gatherers' humans that reached America soon after the last glacial maximum (LGM) ~ 16,000 to 13,000 YBP. They probably consumed avocados in Mesoamerica as they migrated southward, reinforcing their dispersion (Wiersum 1997; Diamond 1999; Galindo-Tovar and Arzate-Fernández 2010). Consequently, our evidence sustains the hypothesis that tropical trees such as avocados developed together and enabled the establishment of American cultures (Galindo-Tovar and Arzate-Fernández 2010). Avocados could be one of the first trees recurrently used in situ across different places in the tropical America, ultimately promoting sedentary life and agriculture development (Smith 1966, 1969; Galindo-Tovar and Arzate-Fernández 2010).

Meanwhile, the first human activities in northwest South America, nowadays Colombia, date around 11,000 YBP (Oyuela-Caycedo 2008). Thus, given the archeological evidence, Colombia must have been a mandatory trading path bringing Mesoamerican and South American plant crops (Oyuela-Caycedo 2008). Specifically, trading activities by diverse human cultures extended from Mexico and Honduras toward northern South America. Archeological evidence points towards in situ utilization of the Mexican avocado group around 8000 YBP in the Tehuacan Valley (Smith 1966, 1969). Further archeological evidence dates the avocado seeds in the Moche Valle of Peru between 2500 and 1800 YBP, and on the Peruvian Pacific coast around 1500 YBP. Likewise, the Caral culture from Peru consumed avocados even before maize 1200 years ago (Heiser 1979; Pozorski 1979; Skidmore 2005). The presence of cassava in Tabasco (Mexico) 4600 years ago offers evidence for recurrent bidirectional trading, including plant resources from the Amazon basin, a plausible pattern for avocados, too (Healy 1978; Pope et al. 2001; Marcus 2003).

Our genetic data agree with a common ancestor between the West Indian and the Colombian groups. Archeological and historical evidence suggests that the West Indian group originated in the lowlands of Yucatan and Belize. According to Galindo-Tovar and Arzate-Fernández (2010), migration may have started as follows: Maya groups living in this area utilized West Indian avocados and migrated eastward, reaching Honduras with evidence of avocado consumption by the Papayeca culture around 1200–1000 YBP. Finally, when the Spanish arrived on the Caribbean coast of South America, they described the presence of avocado trees in Yaharo (Colombia) in 1519. Other chronicles described avocados in the Peruvian Amazonia in 1542 and Ecuador in 1748, matching West Indian characteristics (Galindo-Tovar and Arzate-Fernández 2010). Overall, the three best evolutionary demographic scenarios and the geographic distribution of the genetic groups fit archeological data and chroniclers' descriptions (Galindo-Tovar et al. 2008; Galindo-Tovar and Arzate-Fernández 2010). Interpreting the genetic data in the light of archeological and documentary evidence supports long-distance exchange promoted by human groups in areas where avocado races were already established. Such germplasm exchange in the last 11,000 years may have produced, until present times, the high levels of admixture observed in *P. americana* accessions and the isolation by distance pattern that arises when hybrid accessions are considered (*i.e.*, "rampant" mobility hypothesis). Although the evolutionary history between Mexican and Guatemalan groups remains open and is beyond the scope of the current sampling, our study reinforces the origin and adaptation of avocados to middle-high altitudes in Mexico and Guatemala.

Morphological and genomic convergence of the avocado fruit parsimoniously supports a common ancestor between

the West Indian and Colombian groups, which opens new questions for future studies. For instance, did the pre-adaptation of the lowland West Indian group from Central America enable the colonization of South America? After colonization, did the West Indian group radiate across northern South America, and did it re-adapt to new local conditions diverging in the Colombian groups? Recent genome-wide evidence supporting wild avocados from Costa Rica as a probable distinct ecotype arising from ancient hybridization between the Mexican group and other antique lineages opens further exciting questions (Solares et al. 2023). Did different groups reach South America and hybridize with the nascent sub-populations? Future research efforts aiming to test the above hypotheses will require extensive sampling within Colombia and surrounding areas, covering all distribution ranges, such as Costa Rica in Mesoamerica and Ecuador and Peru in the north of South America. We hypothesize ad hoc that as this study supports a new genetically differentiated sub-population in Colombia, expanding characterization to those countries could reveal further primary or secondary centers of diversity.

Extensive molecular sampling also enables compelling reconstructions of the demographic, evolutionary, and selective processes interplayed during the avocado colonization of South America and its subsequent divergence and adaptation. After all, genomic heterogeneity obliges us to acknowledge that multiple, sometimes conflicting signatures differently imprint discrete positions across the genome (Ellegren and Galtier 2016). It is, therefore, practical to move from a pre-conception that coerces all genetic markers into a ubiquitous underlying process. Instead, embracing more credible genomic thinking requires concurrently acknowledging the existence of porous genome sections more prone to exhibit gene flow (Stölting et al. 2013) compared to low-recombining genomic islands of divergence (Wolf and Ellegren 2017), particularly among groups. This shift in the interpretation of the scale and drivers of heterogeneity from the population level into the molecular fine-tuning is already appreciated in several crop plant species (Cortés et al. 2018), avocados not being the exception. For instance, Rendón-Anaya et al. (2019), comparing $F_{ST}$- and $d_{xy}$ statistics, determined genomic divergence profiles among avocado groups from Mesoamerica, especially against the Hass background. Also considering the three traditional Central American groups, Solares et al. (2023) performed selective sweep and $F_{ST}$- and π-inspired divergence mapping, focusing on the Gwen varietal and the two main heterodichogamy flowering types. Interestingly, authors found gene enrichments for stress response, terpene production, and metabolic processes linked to interracial divergence. These processes could as well underlie the separation of the Colombian group.

Therefore, it remains to be assessed via genomic scans whether the Colombian sub-populations, as part of their divergence and adaptation across northern South America, recruited genomic standing variation dating back to its Mesoamerican ancestors or recalled novel variants that arose after the formation of the isthmus of Panama. Our current data may offer some tendencies in this regard. Reports by Rendón-Anaya et al. (2019) and Solares et al. (2023) utilized whole-genome resequencing (WGR), which confers more coverage of highly recombined regions experiencing insufficient linkage disequilibrium, a common trend in long-lived allogamous tree species (Kelleher et al. 2012). Unfortunately, so far, we have been unable to carry out those detailed genome-wide reconstructions of the inter-racial divergence and selection profiles against the new Colombian subgroups because the NCBI-uploaded assembly of Rendón-Anaya et al. (2019) lacked the pseudomolecule information. A new reanalysis of our dataset using chromosome-level information of the recent Gwen's reference genome by Solares et al. (2023) could provide more resolution for divergence and selective sweep mapping, which may, in turn, sustain the distinctiveness of the Colombian clade.

## Perspectives

The novel sub-population from northwest South America indicates that previously unexplored avocado resources may have persisted via divergent selection across isolated pockets of cryptic diversity at secluded hills and valleys of the northern Andes and the Caribbean savanna. Therefore, further habitat-based population-guided collections (Castañeda-Álvarez et al. 2016) are crucial to prioritize the conservation of avocado genetic resources and better characterize these isolated pockets of diversity (López-Guzmán et al. 2021), both lowland and mid-altitude. Notably, we recommend new sampling trips at the foothills of the Sierra Nevada de Santa Marta, the massif of Montes de María, and the vicinity of Santa María la Antigua del Darién (southeast of the Isthmus of Panama). The former, an isolated mountain range in northern Colombia separated from the Andean range, is where avocados were first described during colonial times (Reyes-Herrera et al. 2020). The second is a reminiscence of an antique orchard of splendid local avocado trees. Yet abandoned decades ago by the conflict and the plagues, its plus tree survivors are candidates to source adaptation and resistance. The latter corresponds to the first settlement founded by conquistadors in mainland America (*ca.* in 1510), initially named Dariena but deserted a couple of decades after due to the unforgiving climate and the fierceness of the indigenous people. All three are likely sources of isolated exotic variation for avocados, without excluding other similar hotspots of cryptic avocado diversity in Ecuador and even south of the Huancabamba depression, a critical biogeographical barrier for Andean plant taxa in the northern Andes of Peru (Weigend 2002).

### Novel genetic resources will boost avocado breeding

Modern platforms for allelic discovery (Hickey et al. 2017) would benefit by targeting cryptic pockets of genetic diversity and habitat-based population-guided collections (Castañeda-Álvarez et al. 2016), as suggested here for avocado resources in the northern Andes. This way, avocado rootstock pre-breeding efforts will not rely exclusively on exogenous diversity, likely without sufficient pre-adaptations to the local conditions of northern South America. Ultimately, the identification, conservation, and utilization of novel adaptive sources (Cortés et al. 2020) among native avocados and related wild species will enable diversifying rootstock selection by offsetting the winnowing effect (McCouch 2004) of clonal tree propagation in natural genetic variation (Ingvarsson and Dahlberg 2019).

Classical diversifying rootstock selection of multi-clonal genotypes and half-sib families (Fernández-Paz et al. 2021) may be boosted by genome-enabled predictions targeting discreet genetic pools (Arenas et al. 2021), such as the Colombian cluster described in this study. Early screening at nurseries could target racial constitution, and desired categorical and quantitative trait scores in adulthood. Predictive markers of the demographic substructure linked with local stresses may also assist biotic resistance (Guevara-Escudero et al. 2021). Multi-clonal and seedling rootstock breeding from local genetic resources (Mickelbart and Arpaia 2002) could broaden the genomic basis of avocado adaptation in regions with contrasting and complex ecologies, such as the northern Andes. Families and farmers also keep a reservoir of native avocado trees in their backyards as traditional orchards and living fences (Galindo-Tovar et al. 2008). These old avocado trees may too source novel rootstock adaptations (Cañas-Gutiérrez et al. 2022), so far unseen in current commercial plantations (Reyes-Herrera et al. 2020). Yet, satisfying the growing demand (Reyes-Gómez et al. 2023) for high-quality avocados (Astudillo-Ordóñez and Rodríguez 2018) with uniform superior rootstocks is an urgent requirement in current nurseries and plantations (Ernst 1999), for which diversified clonality may be more appealing (Ingvarsson and Dahlberg 2019). Even in this case, cryptic Colombian avocado variation may provide the desired adaptive allelic combinations and broad sense heritiabilities to be clonally propagated as a panel of elite rootstock genotypes.

Meanwhile, rootstock breeding could be parallelized with genetic screening for innovative fruit quality traits to modernize and diversify the fresh avocado market beyond Hass. It will not be surprising that arcane avocado gene pools and natural segregation continue sourcing new fruit varieties, as demonstrated by the unexpected discovery of Hass-like cultivars such as Carmen, Gem, Gwen, Maluma, and Méndez (Kremer-Köhne 1999). Regional consumption in Central and northern South America may even offer unexploited market targets for non-Hass fruit types because consumers in those regions are historically used to commercialize and consume "criollo" avocados, often experiencing a Hass aversion. After all, market preference in northern South America tends to prioritize lowland West Indian large and watery fruits, many of which are unknown outside local villages.

### Bridging genomic divergence with morphological differentiation

Morphological analyses in fruit tree crops like avocados are fundamental to corroborate the horticultural racial distinction among differentiated genetic clusters. Each classical avocado group exhibits specific characteristics in its morphology, phenology, ecology, and adaptation (Barrientos Priego 2010). However, cryptic populations could be geographically isolated across different agro-climatic conditions without notorious phenotypic differentiation (Campos-Rojas et al. 2008), as expected at the initial phases of allopatric divergence (Coyne and Orr 2004). Definitory morphological characteristics appropriate for and exclusive to the Colombian population are still lacking, yet they may convey a clearer racial subdivision. López-Galé et al. (2022) suggested using fruit and seed traits to racially discriminate seed-donor "criollo" avocados. Oil quality could as well be an indicator to discern Colombian genotypes (Campos-Rojas et al. 2008).

Even without morphological distinctiveness, cryptic genetic diversity from the Colombian population can still offer standing variation for adaptation to regions where avocado plantations are expanding, while coping with in situ climate change effects (Aitken and Whitlock 2013). Future studies may rely on species distribution modeling (SDM) using current and forecasted climate data, as per geo-referencing of herbaria specimens and germplasm (López-Hernández and Cortés 2019), to predict the future potential distribution of avocado groups (Franklin 2010; Peterson et al. 2011), as well as their local genetic adaptation (Cortés et al. 2022). This approach will enable the prediction of the most suitable environmental conditions for avocado plantations in light of the oncoming conditions as part of an assisted gene flow platform targeting long-lived perennial crops such as fruit trees (Aitken and Bemmels 2016). Current ecological preferences (temperature, humidity, altitude) for planting avocados from different races (Ashworth and Clegg 2003) may be uncoupled with the most dramatic oncoming climatic scenarios. Facilitation and antagonistic biotic interactions could as well be distorted under changing climate. In both cases, the Colombian populations will very likely provide novel alternatives. For instance, the Colombian group closer to the West Indian lowland race seems better adapted to humid tropical areas, a candidate reservoir for rootstock selection of resistance to soil-borne diseases, including *P. cinammomi* (Gross-German and Viruel 2013).

In short, although this study is the first to unveil the Colombian avocado population as a genetically distinct

cluster, its phenotypic novelty is yet to be determined. Newly available reference alignments for avocado have increased the resolution for genomic scans (Solares et al. 2023) and haplotype resolved phasing (Nath et al. 2022), which could, in turn, assist the discovery of previously unnoticed segregating trait variation via reverse genetics. Meanwhile, underutilized allelic combinations within the Colombian cluster can source avocado rootstock adaptive breeding and cultivar diversifying selection.

## Conclusions

This study identified two ancient Colombian genetic clusters of avocados beyond the three traditionally recognized races. These cryptic groups were genetically different and presented minor heterozygosity scores, with possible origin from the West Indian group (alternatively, although less parsimonious, it could also be the case that the West Indian radiate from the Colombian group). The Colombian sub-populations are in two distinct geographic regions, the Andes and the Caribbean, which may reflect divergent local adaptation after the initial colonization from Mesoamerica during the Pleistocene. Exploring the avocado genetic resources in South America will allow identifying genotypes with superior characteristics adapted to diverse agro-ecologies, which may source the selection of new cultivars and rootstock genotypes.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11295-023-01616-8.

**Acknowledgements** We thank the owners and technicians of the eight avocados "Hass" orchards for donating rootstocks tissue, and the field assistants H.M Arias, J.M. Bedoya, K.Y. Calle, L.E. Cano, E. Carranza-Hernández, S.A. Guzmán, J.A. Henao, L.M. Mejía, A.M. Otálvaro, A.N. Sánchez, and H.D. Yepes for sampling roots during 2015–2017, in collaboration with V. Velásquez-Zapata and L. Patiño. The *Sistema General de Regalías*-funded project, under which rootstocks (SR) were sampled, would not have been possible without the vision of M. Londoño, J.M. Cotes, C.M. Gómez-Osorno, P.J Tamayo-Carmona and L.N, Martínez-Caballero, and the collaboration from G.P. Cañas-Gutiérrez, A.P. Clavijo, M. Casamitjana-Causa, O.A. Delgado-Paz, C.A. Díaz-Diez, J. Díaz-Montano, G. Higinio, C.M. Holguín, M. Latorre, L. Muñoz-Baena, P.E. Rodríguez-Fonseca, T.M. Rondón-Salas, A. Sánchez, and S.M. Sepúlveda-Ortega. We are also in debt to the team that maintains the avocado genebank held by AGROSAVIA, specifically A. Caicedo, A. Hernandez, E. Rodriguez, and D. Cañar. Recognition also goes to A. Elbakyan for facilitating retrieving of the non-open access (OA) publications referenced in this work. An early version of this work was improved thanks to insights from A. Barrientos, J.I. Hormaza, M.F. Martínez, P. Manosalva, J. Patel, and G. Wilkie during the IX World Avocado Congress held on September 2019 in Medellín (Colombia). The authors also appreciate the feedback given after the oral presentation of this work in the XI Colombian Botanical Congress held on November 2022 in Villavicencio (Colombia). AGROSAVIA's Capacity Building Division is recognized for financing C-SI's enrolment in the latter. Finally, we would like to express our gratitude to the reviewers and the editor of this document, as their comments have enabled us to create a more robust document.

**Author contribution** N-AAA, C-SI, YR, CAJ, and R-HP designed the sampling of the seedling rootstocks (SR) and CGB collections, and N-AAA led tissue sampling of the seedling rootstocks (SR). B-CJA and D-DP performed DNA extraction and genomic library preparation. B-CJA carried out bioinformatic analyses to discover SNPs and filtered and prepared input datasets. B-CJA, CAJ, L-HF, C-SI, R-HP, and YR made data analyses and interpreted results. All authors wrote and approved the submitted version.

**Funding** Open Access funding provided by Colombia Consortium. This research was supported by the Ministerio de Agricultura y Desarrollo Rural (MADR) of Colombia under the funds TV17 and TV21-22 in the project "*Diseño e implementación de un Plataforma de genotípificación para el Banco de Germoplasma Vegetal de Colombia conservado por AGROSAVIA*" with 1001386 code. The sampling of rootstocks was possible thanks to a grant from *Sistema General de Regalías* (SGR-Antioquia) awarded to N-AAA, contract number 1833. CAJ also acknowledges Vetenskapsakademien (VR) grant 2022–04411. Samples were collected under Permiso Marco 1466–2014 of AGROSAVIA.

**Data availability** We made available this study in Spanish to promote further discussion among researchers and producers within Colombia and the region (Supplementary File S1). This published article and supplementary material include all data generated or analyzed during this study. The data for this study was submitted to NCBI under BioProject: PRJNA878519, which contains 384 raw Illumina sequencing data. Published sequencing data used in this study from Talavera et al. (2019) were from NCBI PRJNA564105 number. The sequences of *P. schiedeana* used as an outgroup were downloaded from NCBI with SRR8295605 accession number published by Rendón-Anaya et al. (2019).

## Declarations

**Conflict of interest** The authors declare no competing interests.



## References

Aitken SN, Bemmels JB (2016) Time to get moving: assisted gene flow of forest trees. Evol Appl 9:271–290. https://doi.org/10.1111/EVA.12293

Aitken SN, Whitlock MC (2013) Assisted gene flow to facilitate local adaptation to climate change. Annu Rev Ecol Evol Syst 44:367–388. https://doi.org/10.1146/annurev-ecolsys-110512-135747

Alcaraz ML, Hormaza JI (2007) Molecular characterization and genetic diversity in an avocado collection of cultivars and local Spanish genotypes using SSRs. Hereditas 144:244–253. https://doi.org/10.1111/J.2007.0018-0661.02019X

Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinforma 12:1–6. https://doi.org/10.1186/1471-2105-12-246/FIGURES/3

Alhusain L, Hafez AM (2018) Nonparametric approaches for population structure analysis. Hum Genomics 12:1–12. https://doi.org/10.1186/S40246-018-0156-4/TABLES/2

Andrews S (2010) FastQC a quality control tool for high throughput sequence data. Babraham bioinformatics. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 23 Feb 2022

Arenas S, Cortés AJ, Mastretta-Yanes A, Jaramillo-Correa JP (2021) Evaluating the accuracy of genomic prediction for the management and conservation of relictual natural tree populations. Tree Genet Genomes 17:12. https://doi.org/10.1007/s11295-020-01489-1

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Report 9(3):208–218. https://doi.org/10.1007/BF02672069

Ashworth VETM, Clegg MT (2003) Microsatellite markers in avocado (Persea americana Mill.): genealogical relationships among cultivated avocado genotypes. J Hered 94:407–415. https://doi.org/10.1093/JHERED/ESG076

Astudillo-Ordóñez CE, Rodríguez P (2018) Parámetros fisicoquímicos del aguacate Persea americana Mill. cv. Hass (Lauraceae) producido en Antioquia (Colombia) para exportación. Cienc Tecnol Agropecuaria 19:383–392. https://doi.org/10.21930/rcta.vol19_num2_art:694

Barlow CC (2002) The ghosts of evolution: nonsensical fruit, missing partners, and other ecological anachronisms. Basic Books

Barrientos Priego AF (2010) El aguacate. Biodiversitas 88:1–7

Bergh B, Ellstrand N (1986) Taxonomy of the avocado. Calif Avocado Soc 70:135–146

Bernal JA, Díaz CA (2020) Actualización tecnológica y buenas prácticas agrícolas (BPA) en el cultivo de aguacate.Corporación Colombiana de Investigación Agropecuaria - AGROSAVIA., Mosquera. Segunda edición. https://doi.org/10.21930/agrosavia.manual.7403831

Borrone JW, Olano CT, Kuhn DN et al (2008) Outcrossing in Florida avocados as measured using microsatellite markers. J Am Soc Hortic Sci 133:255–261. https://doi.org/10.21273/JASHS.133.2.255

Boza EJ, Tondo CL, Ledesma N et al (2018) Genetic differentiation, races and interracial admixture in avocado (Persea americana Mill.), and Persea spp. evaluated using SSR markers. Genet Resour Crop Evol 65(4):1195–1215. https://doi.org/10.1007/S10722-018-0608-7

BROAD Institute (2022a) Genome Analysis Toolkit Variant Discovery in High-Throughput Sequencing Data. https://gatk.broadinstitute.org/hc/en-us. Accessed 23 Feb 2022

BROAD Institute (2022b) Picard tools. https://broadinstitute.github.io/picard/. Accessed 23 Feb 2022

Brock G, Pihur V, Datta S, Datta S (2008) clValid: An R package for cluster validation. J Stat Softw 25:1–22. https://doi.org/10.18637/JSS.V025.I04

Calderón-Vázquez C, Durbin ML, Ashworth VE et al (2013) Quantitative genetic analysis of three important nutritive traits in the fruit of avocado. J Am Soc Hortic Sci 138:283–289. https://doi.org/10.21273/JASHS.138.4.283

Campos-Rojas EE, Espíndola-Barquera M de la C, Mijares-Oviedo P (2008) Diversidad del género Persea y sus usos. Fundación Salvador Sánchez Colín CICTAMEX, S.C. Coatepec, Harinas. México. 59 p.

Cañas-Gutierrez GP, Arango-Isaza RE, Saldamando-Benjumea CI (2019) Microsatellites revealed genetic diversity and population structure in Colombian avocado (Persea americana Mill.) germplasm collection and its natural populations. J Plant Breed Crop Sci 11:106–119. https://doi.org/10.5897/JPBCS2018.0792

Cañas-Gutiérrez GP, Sepulveda-Ortega S, López-Hernández F et al (2022) Inheritance of yield components and morphological traits in avocado cv. Hass from "criollo" "elite trees" via half-sib seedling rootstocks. Front Plant Sci 13

Castañeda-Álvarez NP, Khoury CK, Achicanoy HA et al (2016) Global conservation priorities for crop wild relatives. Nat Plants 2(4):1–6. https://doi.org/10.1038/nplants.2016.22

Charrad M, Ghazzali N, Boiteau V, Niknafs A (2014) Nbclust: An R package for determining the relevant number of clusters in a data set. J Stat Softw 61:1–36. https://doi.org/10.18637/jss.v061.i06

Chen H, Morrell PL, Ashworth VETM et al (2009) Tracing the geographic origins of major avocado Cultivars. J Hered 100:56–65. https://doi.org/10.1093/JHERED/ESN068

Collin FD, Durif G, Raynal L et al (2021) Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest. Mol Ecol Resour 21:2598–2613. https://doi.org/10.1111/1755-0998.13413

Cortés AJ, Skeen P, Blair MW, Chacón-Sánchez MI (2018) Does the genomic landscape of species divergence in Phaseolus beans coerce parallel signatures of adaptation and domestication? Front Plant Sci 9:1–18. https://doi.org/10.3389/fpls.2018.01816

Cortés AJ, Restrepo-Montoya M, Bedoya-Canas LE (2020) Modern strategies to assess and breed forest tree adaptation to changing climate. Front Plant Sci 11:1–12. https://doi.org/10.3389/fpls.2020.583323

Cortés AJ, López-Hernández F, Blair MW (2022) Genome–environment associations, an innovative tool for studying heritable evolutionary adaptation in orphan crops and wild relatives. Front Genet 13:1–14. https://doi.org/10.3389/fgene.2022.910386

Coyne JA, Orr HA (2004) Speciation. Sinauer Associates, Inc., Sunderland, Massachusetts, U.S.A

Danecek P, Auton A, Abecasis G et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156. https://doi.org/10.1093/BIOINFORMATICS/BTR330

Diamond J (1999) Guns, germs and steel: the Fates of Human Societies. W. W. Norton & Company, New York

Ellegren H, Galtier N (2016) Determinants of genetic diversity. Nat Rev Genet 17:422–433. https://doi.org/10.1038/nrg.2016.58

Epskamp S, Cramer AOJ, Waldorp LJ et al (2012) qgraph: Network visualizations of relationships in psychometric data. J Stat Softw 48:1–18. https://doi.org/10.18637/JSS.V048.I04

Ernst AA (1999) Micro cloning: a multiple cloning technique for avocados using micro containers. Rev Chapingo Ser Hortic 5:217–220

FAO (2022). FAOSTAT Crops and livestock products. https://www.fao.org/faostat/en/#data/QCL/visualize. Accessed 30 Jun 2022

Fernández-Paz J, Cortés AJ, Hernández-Varela CA, Mejía-de-Tafur MS, Rodriguez-Medina C, Baligar VC. (2021) Rootstock-mediated genetic variance in cadmium uptake by juvenile cacao (Theobroma cacao L.) genotypes, and its effect on growth and physiology. Front Plant Sci 23(12):777842. https://doi.org/10.3389/fpls.2021.777842

Ferrer-Pereira H, Raymúndez MB, Pérez-Almeida I (2017) Análisis morfoanatómico foliar en grupos hortícolas de aguacate (Persea americana) depositados en la colección del INIA-CENIAP, Venezuela. Trop Subtrop Agroecosyst 20(2):315–328. https://www.redalyc.org/pdf/939/93952506012.pdf

Fisk MC, Yanai RD, Fierer N (2010) A molecular approach to quantify root community composition in a northern hardwood forest — testing effects of root species, relative abundance, and diameter. Can J For Res 40(4):836–841. https://doi.org/10.1139/X10-022

Foote AD, Martin MD, Louis M et al (2019) Killer whale genomes reveal a complex history of recurrent admixture and vicariance. Mol Ecol 28:3427–3444. https://doi.org/10.1111/MEC.15099

Franklin J (2010) Mapping Species Distributions: Spatial Inference and Prediction (Ecology, Biodiversity and Conservation). Cambridge University Press, Cambridge https://doi.org/10.1017/CBO9780511810602

Galindo-Tovar M, Arzate-Fernández A (2010) West Indian avocado: where did it originate? Aguacate Antillano: ¿dónde se originó? Phyton: Int J Exp Bot 79:203–207

Galindo-Tovar ME, Arzate-Fernández AM, Ogata-Aguilar N, Landero-Torres I (2007) The avocado (Persea americana, Lauraceae) crop in Mesoamerica: 10,000 years of history. Harv Pap Bot 12:325–334

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

Galindo-Tovar ME, Milagro-Pérez PA, Alejandre-Rosas JA et al (2011) RELACIONES GENÉTICAS DEL AGUACATE (Persea americana Mill.) EN SIETE MUNICIPIOS DEL CENTRO DE VERACRUZ, CARACTERIZADAS CON MICROSATÉLITES. Trop Subtrop Agroecosyst 13:339–346

Ge Y, Zang X, Tan L et al (2019a) Single-molecule long-read sequencing of avocado generates microsatellite markers for analyzing the genetic diversity in avocado germplasm. Agronomy 9:512. https://doi.org/10.3390/AGRONOMY9090512

Ge Y, Zhang T, Wu B et al (2019b) Genome-wide assessment of avocado germplasm determined from specific length amplified fragment sequencing and transcriptomes: population structure, genetic diversity, identification, and application of race-specific markers. Genes (Basel) 10:1–13. https://doi.org/10.3390/genes10030215

Goodall GE, Little TM, Rock RC et al (1970) Useful life of avocado trees in commercial orchards in California. Yearbook 54:33–36

Granato ISC, Galli G, de Oliveira Couto EG et al (2018) snpReady: a tool to assist breeders in genomic analysis. Mol Breeding 38:102. https://doi.org/10.1007/s11032-018-0844-8

Gross-German E, Viruel MA (2013) Molecular characterization of avocado germplasm with a new set of SSR and EST-SSR markers: genetic diversity, population structure, and identification of race-specific markers in a group of cultivated genotypes. Tree Genet Genomes 9:539–555. https://doi.org/10.1007/s11295-012-0577-5

Gruber B, Unmack PJ, Berry OF, Georges A (2018) dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. Mol Ecol Resour 18:691–699. https://doi.org/10.1111/1755-0998.12745

Guevara-Escudero M, Osorio AN, Cortés AJ (2021) Integrative pre-breeding for biotic resistance in forest trees. Plants 10(10). https://doi.org/10.3390/plants10102022

Guindon S, Dufayard JF, Lefort V et al (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321. https://doi.org/10.1093/SYSBIO/SYQ010

Healy PF (1978) Excavations at Rio Claro, Northeast Honduras: preliminary report. J Field Archaeol 5:15–28. https://doi.org/10.2307/529768

Heiser CB (1979) Origins of some cultivated New World plants. Annu Rev Ecol Syst 10:309–326

Hickey JM, Chiurugwi T, Mackay I, Powell W (2017) Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nat Genet 49(9):1297–1303. https://doi.org/10.1038/ng.3920

Hijmans R, Karney C, Williams E, Vennes C (2021) Geosphere: spherical trigonometry. https://cran.r-project.org/web/packages/geosphere/index.html

Illumina (2022) bcl2fastq Conversion Software. https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html

Ingvarsson PK, Dahlberg H (2019) The effects of clonal forestry on genetic diversity in wild and domesticated stands of forest trees. Scand J for Res 34:370–379. https://doi.org/10.1080/02827581.2018.1469665

IPGRI (1995) Descriptors for avocado (Persea spp). International Plant Genetic Resources, Rome, Italy

Jombart T, Ahmed I (2011) adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. Bioinformatics 27:3070–3071. https://doi.org/10.1093/bioinformatics/btr521

Kamvar ZN, Tabima JF, Grünwald NJ (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ 2:e281. https://doi.org/10.7717/peerj.281

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, New York

Kearse M, Moir R, Wilson A et al (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647. https://doi.org/10.1093/BIOINFORMATICS/BTS199

Kelleher CT, Wilkin J, Zhuang J et al (2012) SNP discovery, gene diversity, and linkage disequilibrium in wild populations of Populus tremuloides. Tree Genet Genomes 8:821–829. https://doi.org/10.1007/s11295-012-0467-x

Knight RC, Amos J, Hatton RG, Witt AW (1927) The vegetative propagation of fruit tree rootstocks. Rep East Mail Res Station A 11:11–30

Kremer-Köhne S (1999) Evaluation of new Hass-like avocado cultivars at Merensky Technological Services. S Afr Avocado Growers' Assoc Yearb 22:120–122

Krome WH (1956) Avocado growing in Dade County. Rev CEIBA 4:339–350

Krueger F (2012) Trim-galore. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore

Kumar R, Vassilvitskii S (2010) Generalized distances between rankings. Proceedings of the 19th international conference on World wide web (IW3C2). pp 571–580. https://doi.org/10.1145/1772690.1772749

Larranaga N, van Zonneveld M, Hormaza JI (2021) Holocene land and sea-trade routes explain complex patterns of pre-Columbian crop dispersion. New Phytol 229:1768–1781. https://doi.org/10.1111/nph.16936

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754. https://doi.org/10.1093/BIOINFORMATICS/BTP324

Lloyd S (1982) Least squares quantization in PCM. IEEE Trans Inf Theory 28:129–137. https://doi.org/10.1109/TIT.1982.1056489

López-Galé Y, Murcia-Riaño N, Romero-Barrera Y et al (2022) Morphological characterization of seed-donor Creole avocado trees from three areas in Colombia. Rev Chapingo Ser Hortic 28:93–108. https://doi.org/10.5154/R.RCHSH.2021.06.010

López-Guzmán GG, Palomino-Hermosillo YA, Balois-Morales R et al (2021) Genetic diversity of native avocado in Nayarit, Mexico, determined by ISSRs. Cienc Tecnol Agropecuaria 22:1686. https://doi.org/10.21930/RCTA.VOL22_NUM1_ART:1686

López-Hernández F, Cortés AJ (2019) Last-generation genome–environment associations reveal the genetic basis of heat tolerance in common bean (Phaseolus vulgaris L.). Front Genet 10:954. https://doi.org/10.3389/fgene.2019.00954

MacQueen J (1967) Some methods for classification and analysis of multivariate observations In: Le Cam LM, Neyman J (eds) Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp 281-297

Marcus J (2003) Recent advances in maya archaeology. J Archaeol Res 11:71–148. https://doi.org/10.1023/A:1022919613720

McCouch S (2004) Diversifying selection in plant breeding. PLoS Biol 2:. https://doi.org/10.1371/journal.pbio.0020347

Mickelbart MV, Arpaia ML (2002) Rootstock influences changes in ion concentrations, growth, and photosynthesis of 'Hass' avocado trees in response to salinity. J Am Soc Hortic Sci Jashs 127:649–655. https://doi.org/10.21273/JASHS.127.4.649

Miller ME, Liberatore KL, Kianian SF (2017) Optimization and comparative analysis of plant organellar DNA enrichment methods suitable for next-generation sequencing. J Vis Exp (125):55528. https://doi.org/10.3791/55528

Miller AJ, Gross BL (2011) From forest to field: perennial fruit crop domestication. Am J Bot 98:1389–1414. https://doi.org/10.3732/ajb.1000522

Mussmann SM, Douglas MR, Chafin TK, Douglas ME (2019) BA3-SNPs: Contemporary migration reconfigured in BayesAss for next-generation sequence data. Methods Ecol Evol 10:1808–1813. https://doi.org/10.1111/2041-210X.13252

Nath O, Fletcher SJ, Hayward A et al (2022) A haplotype resolved chromosomal level avocado genome allows analysis of novel avocado genes. Hortic Res 9:uhac157. https://doi.org/10.1093/hr/uhac157

Oliveira RR, Viana AJC, Reátegui ACE, Vincentz MGA (2015) An efficient method for simultaneous extraction of high-quality RNA and DNA from various plant tissues. Genet Mol Res 14:18828–18838. https://doi.org/10.4238/2015.DECEMBER.28.32

Osorio-Guarín JA, Berdugo-Cely JA, Coronado-Silva RA et al (2020) Genome-wide association study reveals novel candidate genes associated with productivity and disease resistance to Moniliophthora spp. in cacao (Theobroma cacao L.). G3 Genes|Genomes|Genet 10:1713–1725. https://doi.org/10.1534/G3.120.401153

Oyuela-Caycedo A (2008) Late pre-Hispanic chiefdoms of northern Colombia and the formation of anthropogenic landscapes. In: Silverman H, Isbell WH (eds) The Handbook of South American Archaeology .Springer, New York, NY, pp 405–428 https://doi.org/10.1007/978-0-387-74907-5_22

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2(12):e190. https://doi.org/10.1371/journal.pgen.0020190

Peng B, Guan K, Tang J et al (2020) Towards a multiscale crop modelling framework for climate change adaptation assessment. Nat Plants 6:338–348. https://doi.org/10.1038/s41477-020-0625-3

Peterson AT, Soberón J, Pearson RG et al (2011) Ecological niches and geographic distributions (MPB-49). Princeton University Press

Piperno DR (2011) The origins of plant cultivation and domestication in the New World tropics: patterns, process, and new developments. Curr Anthropol 52:S453–S470. https://doi.org/10.1086/659998

Pironon S, Borrell JS, Ondo I et al (2020) Toward unifying global hotspots of wild and domesticated biodiversity. Plants 9:1128. https://doi.org/10.3390/PLANTS9091128

Pope KO, Pohl MED, Jones JC et al (2001) Origin and environmental setting of ancient agriculture in the lowlands of Mesoamerica. Science (1979) 292:1370–1373. https://doi.org/10.1126/SCIENCE.292.5520.1370

Pozorski SG (1979) Prehistoric diet and subsistence of the Moche Valley, Peru. World Archaeol 11:163–184. https://doi.org/10.1080/00438243.1979.9979759

R project (2022) The R project for statistical computing. https://www.r-project.org/. Accessed 7 Feb 2021

Rambaut A, Drummond AJ, Xie D et al (2018) Posterior summarization in Bayesian phylogenetics using tracer 1.7. Syst Biol 67:901–904. https://doi.org/10.1093/sysbio/syy032

Ramirez-Villegas J, Khoury CK, Achicanoy HA et al (2022) State of ex situ conservation of landrace groups of 25 major crops. Nat Plants 8(5):491–499. https://doi.org/10.1038/s41477-022-01144-8

Rendón-Anaya M, Ibarra-Laclette E, Méndez-Bravo A 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. https://doi.org/10.1073/PNAS.1822129116

Reyes-Gómez H, Genaro Martínez-González E, Aguilar-Ávila J et al (2023) Gobernanza de la cadena global de valor del aguacate en México. Cienc Tecnol Agropecuaria 24:3120. https://doi.org/10.21930/rcta.vol24_num2_art:3120

Reyes-Herrera PH, Muñoz-Baena L, Velásquez-Zapata V et al (2020) Inheritance of rootstock effects in avocado (Persea americana Mill.) cv. Hass. Front Plant Sci 11:1957. https://doi.org/10.3389/FPLS.2020.555071/BIBTEX

Rios Castaño D, Tafur Reyes R (2003) Variedades de aguacate para el trópico: Caso Colombia. Proceedings V World Avocado Congress. pp 143-147

Ríos-Castaño D, Corrales-Medina DM, Daza-Gómez GJ, Aristizábal-Gallo A (2005) Aguacate: Variedades y patrones importantes para Colombia. Feriva, Profrutales, Palmira

Rodríguez-Henao E, Caicedo-Arana Á, Enriquez-Valencia AL, Muñoz-Florez JE (2017) Evaluation of tolerance to Phytophthora cinnamomi Rands in avocado (Persea americana Miller.) germplasm. Acta Agron 66:128–134

Rubinstein M, Eshed R, Rozen A et al (2019) Genetic diversity of avocado (Persea americana Mill.) germplasm using pooled sequencing. BMC Genomics 20:379. https://doi.org/10.1186/s12864-019-5672-7

Sánchez-González EI, Gutiérrez-Díez A, Mayek-Pérez N (2020) Outcrossing rate and genetic variability in Mexican race avocado. J Am Soc Hortic Sci 145:53–59. https://doi.org/10.21273/JASHS04785-19

Sekula M, Datta S, Datta S (2017) optCluster: an R package for determining the optimal clustering algorithm. Bioinformation 13:101. https://doi.org/10.6026/97320630013101

Skidmore J (2005) Earliest complex culture in the Americas. Mesoweb Reports. https://www.mesoweb.com/reports/caral2.html. Accessed 28 Oct 2022

Smith CE (1966) Archeological evidence for selection in avocado. Econ Bot 20(2):169–175. https://doi.org/10.1007/BF02904012

Smith CE (1969) Additional notes on pre-Conquest avocados in Mexico. Econ Bot 23(2):135–140. https://doi.org/10.1007/BF02860618

Solares E, Morales-Cruz A, Balderas RF et al (2023) Insights into the domestication of avocado and potential genetic contributors to heterodichogamy. G3 Genes|Genomes|Genet 13:jkac323. https://doi.org/10.1093/g3journal/jkac323

Sommaruga R, Eldridge HM (2021) Avocado production: water footprint and socio-economic implications. EuroChoices 20:48–53. https://doi.org/10.1111/1746-692X.12289

Navarro-Villa P (2022) Value of avocado exports from Colombia from 2014 to 2020 https://www.statista.com/statistics/1069951/colombia-avocado-exports-value/. Accessed 14 Jul 2022

Stölting KN, Nipper R, Lindtke D et al (2013) Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. Mol Ecol 22:842–855. https://doi.org/10.1111/mec.12011

Talavera A, Soorni A, Bombarely A et al (2019) Genome-Wide SNP discovery and genomic characterization in avocado (Persea americana Mill.). Sci Rep 9(1):1–13. https://doi.org/10.1038/s41598-019-56526-4

Thorp TG, Aspinall D, Sedgley M (2015) Influence of shoot age on floral development and early fruit set in avocado (Persea americana Mill.) cv. Hass. J Hortic Sci 68:645–651. https://doi.org/10.1080/00221589.1993.11516396

Tracy CA, Widom H (1994) Level-spacing distributions and the Airy kernel. Commun Math Phys 159:151–174. https://doi.org/10.1007/BF02100489

van der Werff H (2002) A synopsis of Persea (Lauraceae) in Central America. Novon 12:575–586. https://doi.org/10.2307/3393142

Wang L, Zhang W, Li Q (2020) AssocTests: an R package for genetic association studies. J Stat Softw 94:1–26. https://doi.org/10.18637/jss.v094.i05

Webster AD (1995) Temperate fruit tree rootstock propagation. N Z J Crop Hortic Sci 23:355–372. https://doi.org/10.1080/01140671.1995.9513912

Weigend M (2002) Observations on the biogeography of the Amotape-Huancabamba zone in northern Peru. Bot Rev 68(1):38–54. http://www.jstor.org/stable/4354410

Wienk R, Mostert-O'Neill M, Abeysekara N et al (2022) Genetic diversity, population structure, and clonal verification in South

African avocado cultivars using single nucleotide polymorphism (SNP) markers. Tree Genet Genomes 18:41. https://doi.org/10.1007/s11295-022-01573-8

Wiersum KF (1997) From natural forest to tree crops, co-domestication of forests and tree species: an overview. Neth J Agric Sci 45:425–438. https://doi.org/10.18174/NJAS.V45I4.503

Wilson GA, Rannala B (2003) Bayesian inference of recent migration rates using multilocus genotypes. Genetics 163:1177–1191. https://doi.org/10.1093/genetics/163.3.1177

Wolf JBW, Ellegren H (2017) Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet 18:87–100. https://doi.org/10.1038/nrg.2016.133

Wolters B (1999) Dispersion and ethnobotany of the cacao tree and other amerindian crop plants. Angew Bot 73:128–137

Zeng W, Zhou B, Lei P et al (2015) A molecular method to identify species of fine roots and to predict the proportion of a species in mixed samples in subtropical forests. Front Plant Sci 6:1–10. https://doi.org/10.3389/FPLS.2015.00313/XML/NLM

Zuazo VHD, Lipan L, Rodríguez BC et al (2021) Impact of deficit irrigation on fruit yield and lipid profile of terraced avocado orchards. Agron Sustain Dev 41:1–16. https://doi.org/10.1007/S13593-021-00731-X/FIGURES/4