Muñoz-Peréz et al. Population genomics of Northern South America coconut

Genome-wide diversity of northern South America cultivated Coconut (*Cocus nucifera* L.) uncovers diversification times and targets of domestication of coconut globally

Jorge Mario Muñoz-Pérez[*], Gloria Patricia Cañas[†], Lorena López[†,ϒ] and Tatiana Arias[*,1]

[*] Laboratorio de Biología Comparativa. Corporación para Investigaciones Biológicas, Cra. 72 A No. 78 B 141, Medellín, Colombia.

[†] Unidad de Control Biológico. Corporación para Investigaciones Biológicas, Cra. 72 A No. 78 B 141, Medellín, Colombia.

[ϒ]Grupo de Investigación en Biociencias (GIB). Institución Universitaria Colegio Mayor de Antioquia, Medellín, Colombia.

All data have been deposited in Bioproject (PRJNA579494) and SRA (SRR10345275-SRR10345401).

Population genomics of Northern South America coconut

5-8 Key words: Arecaceae, coconut, diversification times, northern South America, population genomics, selection.

[1] Corresponding Author:

T.A. current address: The Colombian Orchid Society (SCO), Cra. 52 No 73-29, Medellín, Colombia. E-mail: tatiana.tatianaarias@gmail.com, Phone: +57 300 845 6250.

Total word count (excluding summary, references and legends):		No. of figures:	5
Summary:	196	No. of Tables:	1
Introduction:	946	No of Supporting Information files:	7 (4 figures, 3 tables)
Materials and Methods:	2583		
Results:	1101		

Discussion:	1128		
Acknowledgments:	74		

22

Summary

- Independent origins of cultivated Coconut (*Cocos nucifera L.*) have been identified in the Old-World Tropics corresponding to the Pacific and Indo-Atlantic oceanic basins. We use unknown coconut cultivated varieties in northern South America to conduct genetic diversity and population structure analysis for the region using single nucleotide polymorphisms (SNPs) markers along the genome and phenotypic data.
- Analyses revealed two highly genetically differentiated subpopulations corresponding to Pacific and Atlantic oceanic basins. While, this pattern was initially revealed at a global scale, it has now been confirmed at a local scale using genomics in northern South America. We use genotypes identified here to estimate diversification times for main cultivated coconut varieties at a global scale: An early divergent Atlantic tall cultivar, splits from sisters Pacific Tall and Pacific Dwarf cultivars approximately 5400 years ago, while Pacific Tall and Pacific Dwarf cultivars split from a shared common ancestor approximately 1600 years ago.
- Five loci have been identified as candidate domestication targets, including genes related to vital functions such reproduction, respiration, and defense.
- These results propose a dated evolutionary hypothesis for the independent origins of cultivated coconut globally were diversification between coconut populations has happened very recently and within few generations.

Introduction

Coconut (Cocos nucifera L.; Arecaceae) is one of the most economically important palms worldwide due to the versatility of its products, which have been used by humanity since ancient times as a source of food, drink, fuel, medicine, and construction material (Aljohi et al., 2016; Food and Agriculture Organization of the United Nations, 2015; Prades, et al., 2012). Coconut is cultivated extensively in all the tropical countries of the world and is considered one of the most important plant species to guarantee the survival of humankind (Gunn et al., 2011). Many studies have been conducted to characterize the genetic diversity of coconut collections and to understand coconut cultivation history. However, northern South America is an under represented region in these studies. Using a combination of population genomics, and diversification analysis, we evaluated the relationships and origins of three non-admixed genetically and morphologically distinctive coconut populations: Atlantic Tall, Pacific Tall and Pacific Dwarf.

Coconut is a diploid with 32 chromosomes (2n = 32) and a genome size of about 2.4 Gb (Xiao et al., 2017; Yang et al., 2018), and is the sole species of the genus *Cocos*. Two clearly structured subpopulations of coconut have been identified at the genetic level (Gunn et al., 2011), suggesting that coconut was brought under cultivation in two separate locations globally, the Pacific basin and the Indo-Atlantic Ocean basin. There is also a record of prehistoric trade routes and of the colonization of the Americas (Gunn et al., 2011). Its selection by cultivation, hybridization and introgression in different tropical countries has led to the evolution, dissemination and classification of two well-recognized palm genotypes (Geethanjali et al., 2018). They are identified by a combination of their geographical location, type of reproduction and growth habits. At present, this tropical tree crop is distributed across 93 tropical countries (Tang et al., 2006), including Central and South America, East and West Africa, Southeast Asia, and the Pacific Islands, and is grown over 12 million hectares of land (Batugal, 2005).

Worldwide coconut diversity studies have been carried out using RFLP (Lebrun et al., 2005; Meerow et al., 2012), microsatellite (Gunn et al. 2011; Lebrun et al., 2005; Meerow et al., 2012; Geethanjali et al., 2018), and AFLP (Perera et al., 1998; Lebrun et al., 2005 Meerow et al., 2012). All these studies detected high level of genetic differentiation between Pacific and Indo-Atlantic Oceanic basin samples. The Pacific group includes four admixed sub-groups with blurred boundaries: Southeast Asia, Melanesia, Micronesia, Polynesia; and a clearly distinct Panama and Peru group. Dwarf coconut palms also belong to the Pacific and form a genetically uniform group, irrespective of their geographical origin. The Indo-Atlantic group includes cultivated coconut palms from Africa and the Atlantic coast of Latin America. They originated from the Indian subcontinent and have extensively breed with coconuts from Southeast Asia, resulting from Austronesian migrations to Madagascar (Lebrun et al., 2005). The clear genetic differentiation between these lineages suggest that coconuts have a long-standing evolutionary presence in both ocean basins and were brought into cultivation independently in each of these regions. This finding has several important implications for coconut domestication including two geographical origins of coconut cultivation.

Coconuts lack clear domestication traits making it difficult to trace the origins of cultivation strictly by morphology. Two distinctively different forms of the coconut fruit, known as *niu kafa* and *niu vai*-- Samoan names for traditional Polynesian varieties-- are known. The *niu kafa* form is triangular and oblong with a large fibrous husk. The *niu vai* form is rounded and contains abundant sweet coconut “water” when unripe. Also, there are "tall" type palms which are generally cross-pollinated, having have long stems and slow fruit production, and can be of the *niu kafa* or *niu vai* type (Gunn et al., 2011). The "dwarf" type palms are recognized for being self-pollinated, evidencing important domestication features such as short stems, high productivity, and low genetic variability. These two types of palms gave rise to a vast number of coconut populations of pantropical distribution that are poorly characterized at the genomic level, and that are generally characterized and identified based on variable morphological

agronomic traits, which are not apparent in juvenile phases (Ribeiro et al., 2013; Guevara & Jáuregui, 2012; Teulat et al., 2000).

High-throughput sequencing (HTS) has enabled the discovery of SNPs throughout the genome, greatly increasing power for detecting neutral and adaptive patterns of variation (Taranto et al., 2016; Torkamaneh et al 2016). Genotyping by sequencing (GBS) is considered an efficient and economical method to quickly discover SNPs among several individuals simultaneously (Pavan et al., 2017; Taranto et al., 2016) allowing geneticists and plant breeders to use these resources for crop improvement. Such an approach has benefited many crops, like watermelon (Nimmakayala et al., 2014), cowpea (Xiong et al., 2016), rice (Xu et al., 2012), and spinach (Shi et al., 2017).

Here we explored the genetic variability of coconut populations distributed in coastal areas of northern South America using combination of phenotypic data and a population genomics approach. Genotyping by Sequencing libraries were used to generate a genome-wide SNP panel, study the genetic diversity, population structure and diversification times among cultivated coconut palms growing in Colombian coastal areas, and discovering alleles under selection. We aim to a) determine the genetic identity and diversity of cultivated coconut in Colombia, b) estimate dates for the independent origins of cultivated Coconut from the Pacific and Indo-Atlantic oceanic basins, c) discover outlier markers showing significant divergence from neutrality resulting from selection in the course of coconut domestication. Our work informs management practices for coconut germplasm resources in Colombia, provides a better understanding of independent origins of cultivated coconut, and add to understand how coconut palms evolve under domestication.

Materials and methods

Plant materials and experimental design: A total of 127 coconut individuals were selected from main coconut producing departments on the Atlantic (two sampled sites: 30 individuals from Córdoba and 30 from Antioquia) and Pacific (three sampled sites: 30 individuals from Nariño, 30 from Cauca and 7 from Chocó) Colombian coasts. Selection of individuals was made trying to cover all different cultivated coconut varieties present in each zone. There was not available information about specific cultivars and varieties that were grown in each zone. All selected individuals were geo-referenced and marked in the field (Table S1, S3).

The Minimum Descriptor System for coconut palms was used to select a suite of qualitative and quantitative variables that could be used to describe coconuts (IPGRI, 1998). Analyzed variables included: palm shape, trunk circumference (cm), plant height (m), petiole length (cm), rachis length (cm), number of leaflets, fruit shape, nut shape, number of coconuts per palm, fruit polar circumference (cm), fruit equatorial circumference (cm), nut polar circumference (cm), nut equatorial circumference (cm), fruit weight (gr), nut weight (gr), copra weight (gr), mesocarp weight (gr), and volume of coconut water (ml) (Table S3). Sampled coconuts were clustered using two and three visualization clusters through a Principal Component Analysis (PCA) in R.

DNA extraction, library preparation and sequencing: DNA extraction was performed from fresh leaf tissue collected in the field. A combination of CTAB (Hexadecyl trimethyl ammoniunm bromide) extraction method (Murray & Thompson, 1980) with Epoch Life Science® (Epoch Life Science Inc. Missouri, Texas, USA) purification columns was used to extract and purify DNA. DNA quality was determined by spectrophotometry using Nanodrop 2000 (Thermo Fischer Scientific®, Waltham, Massachusetts, USA) and DNA concentration through fluorometry using Qubit 3.0 (Life Technologies® Carlsbad, California, USA).

Library construction and sequencing was outsourced with LGC Genomics (Berlin, Germany). Initially a pilot experiment using 12 samples (3 individuals x 4 populations/2 pacific coast and 2

Atlantic coast) was performed to find out which set of restriction enzymes work better to recover more Single Nucleotide Polymorphisms (SNPs) in coconut. Between 100 and 200 ng of genomic DNA was digested using 2 units of the *MslI* restriction enzyme (New England Biolabs ®, Ipswich, Massachusetts, USA.) using 1 unit of NEB4 buffer in 20 µl of volume and for 2 hours at 37 ° C. The restriction enzyme was inactivated by incubation at 80 ° C for 20 min. The same procedure was used for double digestion with the *PstI-MspI* enzymes (New England Biolabs ®, Ipswich, Massachusetts, USA).

For the ligation reaction, 10 µl of each restriction digest were transferred to a new 96 well PCR plate, mixed on ice first with 1.5 µl of one of 96 inline-barcoded forward blunt Adaptors (pre-hybridized, concentration 5 pmol/µl), followed by addition of 20µl Ligation master mix (containing: 15 µl NEB Quick ligation buffer, 0.4 µl NEB Quick Ligase, 7.5 pM pre- hybridized common reverse blunt Adaptor). Ligation reactions were incubated for 1h at room temperature, followed by heat inactivation for 10 min at 65°C. For library purification, all reactions were diluted with 30 µl TE 10/50 (10mM Tris/HCl, 50mM EDTA, pH: 8.0) and mixed with 50 µl Agencourt XP beads, incubated for 10 min at room temperature and placed for 5 min on a magnet to collect the beads. The supernatant was discarded and the beads were washed two times with 200 µl 80% ethanol. Beads were air dried for 10 min and libraries were eluted in 20 µl Tris Buffer (5 mM Tris/HCl pH: 9.0).

For library amplification, 10 µl of each of the 96 Libraries were separately amplified in 20 µl PCR reactions using MyTaq (Bioline) and standard Illumina TrueSeq amplification primers. Cycle number was limited to 14 Cycles. Pooling and cleanup of GBS libraries was done using 5 µl from each of the 96 amplified libraries. PCR primer and small amplicons were removed by Agencourt XP bead purification using 1 volume of beads. The PCR enzyme was removed by an additional purification on Qiagen MinElute Columns. The pooled Library was eluted in a final volume of 20µl Tris Buffer (5 mM Tris/HCl pH: 9).

Normalization was done using Trimmer Kit (Evrogen). One µg pooled GBS library in 12 µl was mixed with 4 µl 4x hybridization buffer, denatured for 3 min at 98°C and incubated for 5 hours at 68°C to allow re-association of DNA fragments. Twenty µl of 2 x DSN master buffers was added, and samples were incubated for 10 min at 68°C. One unit of DSN enzyme (1U/µl) was added and the reaction was incubated for another 30 min. Reaction was terminated by the addition of 20µl DSN Stop Solution, purified on a Qiagen MinElute Column and eluted in 10µl Tris Buffer (5 mM Tris/HCl pH: 9). The normalized library pools were re-amplified in 100µl PCR reactions using MyTaq (Bioline). For each pool, a different i5-Adaptor primer was used to include i5-Indices into the libraries, allowing parallel sequencing of multiple libraries on the Illumina NextSeq 500 sequencer. Cycle number was limited to 14 cycles. The GBS libraries were size selected on Blue Pippin, followed by a second size selection on a LMP-Agarose gel, removing fragments smaller than 300bp and those larger than 400bp. Sequencing was done on an Illumina NextSeq 500 using V2 Chemistry (300 cycles).

Variant calling: Read pre-processing and de-multiplexing of all library groups was done using the Illumina bcl2fastq 2.17.1.14 software (Illumina, Inc), 1 or 2 mismatches or Ns were allowed in the barcode read when the barcode distances between all libraries on the lane allowed for it. De-multiplexing of library groups into samples was completed according to their inline barcodes and verification of restriction site. No mismatches or Ns were allowed in the inline barcodes, but Ns were allowed in the restriction site. Clipping of sequencing adapter remnants from all reads was completed and reads with final length < 20 bases were discarded. Restriction enzyme site filtering of read 5' ends was performed and reads with 5' ends not matching the restriction enzyme site were discarded. Quality trimming of adapter clipped Illumina reads was done by removing reads containing Ns, and trimming reads at 3'-end to get a minimum average Phred quality score of 20 over a window of ten bases. Reads with final length < 20 bases were discarded. A subsampling (evenly across the complete FASTQ files) of quality-trimmed reads was performed to 1.5 million read pairs per sample. FastQC reports (Andrews, 2008) were prepared for all FASTQ files and read counts were recorded.

Two software were implemented to call variants for the pilot study with the goal of finding the best performing program: DiscoSnpRad (Gauthier et al., 2017), and Stacks (Catchen et al., 2013). The best performing restriction enzyme and best bioinformatics pipelines to call variants were subsequently used for this study (Table S2). All pipelines used here can be found at https://github.com/TheAriasLab/POPULATIONS_GENOMICS.

To call all variants Stacks (reference genome based approaches) and DiscoSnpRad (*de novo* approaches) were used. For Stacks, Bowtie2 (Langmead & Salzberg, 2012) was used to index the coconut draft reference genome (Xiao et al., 2017) and to align reads to the reference genome index. Samtools (Li et al., 2009) was used to compress sam files and to convert them to their binary version (bam), necessary for Stacks. The Stacks pipeline was implemented as follows: The "populations" script was used to filter and identify SNPs according to *minimum percentage of individuals in a population in which a SNP must be present to be considered* = 0.8, *minimum allele frequency of a SNP to be considered*= 0.05 (Fig. S1), *maximum observed heterozygosity of a SNP to be considered*= 0.70. Five datasets were obtained ranging from p [1-5] indicating the minimum populations in which one SNP must be present to be included.

For DiscoSnpRad, the script run_discoSnpRad.sh was ran with the following parameters: remove low complexity bubbles, extend each polymorphism with left and right unitigs, *k-mer size* = 31, *branching strategy*, *allows non-symmetrical branching bubbles* = 1, *minimum coverage* = 12, *maximum errors per read* = 8, *maximum deletion size* = 10, *maximum SNPs per bubble* = 5. Other parameters were left as default. DiscoSnpRad SNPs were filtered according to the "populations" script that filtered Stacks data.

The Stacks p5 data set containing 27,600 SNPs for 125 individuals, after removing accessions 3527 and 3418 (Table S1 y S3) due 100% missing data rate, was used as base for posteriors analyses. Stacks took longer to run than DiscosnpRad and used more RAM but recovered more SNPs with less missing data (56-8.57%; Fig. S1). All further analyses were done with the dataset Stacks p5 called full data set.

Populations statistics, population structure and clustering analysis: Hierfstats 0.04-30 (Goudet, 2005) from the R package was used to calculate pairwise F_{ST}, inbreeding coefficients F_{IS} and heterozygosity H_o. Confidence intervals of 95% in each population F_{IS} and pairwise F_{ST} were estimated with 1,000 bootstraps. Adegenet 2.1.1 (Jombart, 2008) was used initially to visualize genetic differences between populations using Principal Component Analysis (PCA) and Principal Component Discriminant Analysis (DPCA). Poppr (Kamvar et al., 2015) was used to calculate Euclidean distances between individuals and build a Minimum Spanning Network (MSN) to visualize genetic grouping.

A Bayesian method using Structure 2.3.4 (Hubisz et al., 2009) was used to identify genetic structure of populations with an admixture model, a burn-in of 25,000, and a number of MCMC replicates after burn-in of 250,000. Runs were performed from k = 1 to k = 6, assuming coconut populations from Colombia could show population substructure based on a global divide between Atlantic and Pacific populations (Gunn et al., 2011) and further substructures within each coast, each run was repeated 10 times for a total of 60 runs per test. Evanno's method (Evanno et al., 2005) was implemented in Structure Harvester (Earl, 2012) and the delta K was determined as a measure that best describes the data. Permutation of the runs was done in Clumpp 1.1.2 (Jakobsson & Rosenberg, 2007) and visualization of the data in Distruct 1.1 (Rosenberg, 2004). To account for the influence of missing data we ran FastStructure (Raj, Stephens, & Pritchard, 2014) for a reduced missing dataset containing 69 individuals with missing data < 0.05.

Phylogenetic analysis and diversification times: An un-rooted phylogenetic tree was built using SNPhylo (20140701 version) (Lee et al., 2014). Low quality data was removed and a concatenated SNPs matrix was used to build a Maximum Likelihood (ML) tree with 100 non-parametric bootstraps. A full data set was filtered using: A *percentage of low coverage samp*le= 5, a *minimum depth of coverage* = 5, a *Minimum Allele Frequency* (*MAF*) *threshold* = 0.1, a *missing rate* = 0.1, and a *Linkage Disequilibrium rate* (*LD*) = 0.1. Using these parameters, the final data set contained 125 accessions and 2,446 SNPs.

The program SNAPP (Bryant et al., 2012) was used to calculate divergence times between subgroups identified by Structure and phenotypic PCA analysis. The Stacks p5 dataset containing 27,600 SNPs was filtered in vcftools (Danecek et al., 2011), and locus with any missing data were removed resulting in a new data matrix of 18,385 SNPs, next the Phrynomics R package (Banbury & Leache, 2014) was used to remove invariant and non-binary sites that could be properly used in SNAPP. Due to high computational demand only five non-admixed individuals and a 1,000 randomly selected SNPs for each genetic group detected here were included in this analysis.

An initial tree without calibration points was calculated using mutation rates fixed at 1. A *gamma distribution with alpha* = 2, *beta* = 200 (shape scale mode), *kappa* = 1 were used as priors; the lambda parameter was left to be estimated by the MCMC with a gamma distribution with *alpha* = 2, *beta* = 200, and *offset* = 0. The coalescence rate was sample from the chain. The resulting tree was scaled according to the calibration distribution starting points and used as starting tree for the divergence time estimations runs.

The input file for SNAPP was prepared with the script snapp_prep.rb (Stange et al., 2018). Two calibration points were implemented to date the SNAPP tree: a fossil calibration and an indirect calibration point. Gunn et al. (2011) suggest that Indo-Atlantic coconuts were already under cultivation in the southern India subcontinent around 2,500-3,000 years ago base on archaeobotanical (Wheeler et al., 1946), linguistics (Fuller, 2007) and written evidence (Menon and Pandalai, 1958). An indirect calibration for the split of the Indo-Atlantic and Pacific coconut cultivated varieties was used based on a hypothesis putted forward by Harries et al. (2004). They suggested that coconut was cultivated in the Indo-Malaysian peninsula sometime before catastrophic floods caused by ice sheet collapse in North American glacier lakes 8-14 thousand years before present (Blanchon & Shaw, 1995) resulting in the submergence of a land mass equivalent in size to the Indian sub-continent. This was the center of development of paddy rice cultivation by agriculturally-based civilizations that migrated north with their crops, including coconut, after the flooding (Oppenheimer, 1998). This proposed center of origin is also

consistent with that proposed in the 1930s by Vavilov (Chester, 1951). We calibrate this split at 7000 years using a normal distribution with *offset* = 10.5, *mean* = 0, *standard deviation* = 1.3.

A second calibration point using fossils of coconut fruits dated with C_{14} at 1400 years was implemented to calibrate the divergence between Tall and Dwarf Pacific coconuts (Lepofsky et al., 1992). For this calibration, a lognormal distribution for MRCA (most recent common ancestor), with *offset* = 1.4, *mean* = 0.4 (real space) and *standard deviation* = 1.25 was implemented.

All xml files were analyzed using BEAST 2.5.2 (Bouckaert & Vaughan et al., 2019). To ensure congruence we ran the analyses three times for 10 million generations each, sampling parameters every 5000 generations, assessing convergence and sufficient chain mixing (Effective Sample Sizes > 200) using Tracer 1.5 (Rambaut et al., 2018). After removal of 10 % of each run as burn-in, the remaining trees were combined using LogCombiner, summarized as maximum clade credibility (MCC) trees in TreeAnnotator, and visualized using FigTree (Rambaut, 2016).

Outlier detection and analysis: Bayesian approaches were implemented to identify markers showing significant divergence from neutral loci resulting from selection in the course of coconut domestication. The software BayeScan v2.1 (Foll & Gaggiottti, 2008) was used detect outlier loci using differences in allele frequencies between populations. Differences in allele frequency between a common gene pool and each subpopulation are measured by their F_{ST} values. Only non-admixed individuals were included using each group predicted by Structure. BayeScan was run using default settings: *sample size* = 5,000, *thinning* = 10, *20 pilot runs of length* = 5,000 and an additional *burn-in* = 50,000, resulting in a total of 100,000 iterations following the initial burn-in of 50,000. We filtered loci by a FDR of 0.05 and repeated the runs three times to ensure consistency.

Outlier locus identification was made by mapping the location of each SNP and their flanking regions (200bp) in the reference genome (Xiao et al., 2017). Sequences were translated into

proteins and the largest Open Reading Frames (ORFs) were selected for each locus using ExPASy (Gasteiger et al., 2003) to use in HmmerWeb version 2.34.0. A protein sequence vs. a protein sequence database search (phmmer) was performed for each locus against SwissProt, Reference Proteomes and UniProtKB. Additionally, a BLAST search was performed using nucleotide (nblast), megablast and discontinuos megablast (dmegablast).

Results

Genotyping of coconut accessions: A total of 127 coconut individuals were obtained from main coconut producing departments in the Atlantic and Pacific Colombian coasts. We obtained 367,830,670 raw reads for the 127 individuals analyzed in this study (average 3,198,528, standard deviation 1,196,244; Table S1; NCBI Bioproject PRJNA579494; NCBI SRA Accessions SRR10345275-SRR10345401). The FastQC results indicated that reads were of high quality across the entire 150bp length. Because GBS datasets often have high levels of missing data, filtering parameters can have a major impact on the overall size of the dataset. After filtering loci, the dataset resulted in 125 individuals and 27,600 unlinked SNPs markers that had a minor allele frequency > 0.01%, mean missing data 8.57% and standard deviation 3.5%.

Genetic diversity and differentiation: Summary statistics for each sampling site (Antioquia, Córdoba, Nariño, Cauca, Chocó) and coastal region (Atlantic and Pacific) consisted of global F_{ST} calculations between sampling sites and regions. Observed heterozygosity, expected heterozygosity and inbreeding coefficient averaged across all loci were also calculated. Global F_{ST} suggest larger genetic divergence between Pacific and Atlantic coasts (0.38274 $p<0.001$) rather than between sampling site comparisons (0.29249, $p<0.001$). Expected and observed heterozygosity (H_e, H_o) and inbreeding coefficients (F_{IS}) within Pacific and Atlantic regions are comparable. Average H_e, H_o per site were lower for the Pacific than the Atlantic; while F_{IS} was higher in the Pacific (Fig. 1).

Here, we found a pattern of genetic differentiation between the two coasts Atlantic and Pacific. However, F_{ST} within samples from the same coast are smaller than F_{ST} of samples from different coasts, for example: Nariño and Cauca F_{ST} (0.0153 - 0.0167), and Nariño and Chocó F_{ST} (0.0821 - 0.0844). These F_{ST} do not overlap indicating differentiation within the Pacific. These are an order of magnitude smaller than F_{ST} of Nariño with Antioquia (0.4001 - 0.4073) and Córdoba (0.352 - 0.3577). This level of differentiation suggests long-term evolutionary divergence between two subpopulations, with independent origins of cultivated coconuts from within each

lineage. Lower levels of population divergence were detected in the full data set for comparisons within Atlantic and Pacific populations. The highest genetic divergence was found between Antioquia and Cauca's populations (F_{ST} = 0.400-0.407). Meanwhile the closest genetic divergence was found between Nariño and Cauca's populations (F_{ST} = 0.0152-0.0167) (Table 1). Global F_{ST} suggest larger genetic divergence between Pacific and Atlantic coasts (0.38274 $p<0.001$) rather than the five-group comparison (0.29249, $p<0.001$).

We detected F_{IS} ranges from -0.5 to values close to 1 (Fig. 1e, f), Atlantic coconut populations are predominantly exogamous due to cross-pollination and those of the Pacific are predominantly inbred due to self-pollination. Global heterozygosity is less than 0.2, which partly reflects the inbreeding habits of most of the Pacific palms and some of the Atlantic (Fig. 1a-d). Inbreeding coefficients for Antioquia and Cordoba palms had similar positive values close to zero. In Atlantic populations, there is a slight excess of homozygotes and they are slightly in HW disequilibrium. Pacific populations (Cauca, Nariño and Chocó) had inbreeding coefficients close to 0.5, suggesting excess close to 50% in homozygotes according to what is expected under HW equilibrium (Fig. 1 c, d). Similarly, heterozygotes observed grouped in either Atlantic or Pacific regions. Atlantic populations had higher heterozygosity than those of the Pacific (Fig. 1 a-d).

Population structure: To understand population structure of coconut cultivars in northern South America we used several approaches, PCA and DPAC, MSN and Bayesian clustering methods. The results of PCA using Hierfstats (Fig. 2) clearly showed that accessions tend to group by either Pacific or Atlantic coasts. The first two principal components explained 49.47% of the total variation. DPAC was done with the first 39 eigenvalues representing 81.33% of the variation (Fig. 2b). This analysis splits genotypic data into four groups divided according to the two core groups (Pacific and Atlantic), with three sub-groups within the Pacific accessions, and a single group from the Atlantic accessions. Some Atlantic coconut accessions show a degree of admixture with accessions from the Pacific but the contrary is rare. PCA and DPAC of these coconut accessions clustered similarly in un-rooted phylogenetic trees (Fig. 3c).

Bayesian clustering algorithm implemented in Structure (Hubisz et al., 2009) was first used to estimate ancestry proportions for each coconut accession. Evanno's test showed that four populations (K = 4) represented the best number of clusters for 125 coconut accessions and 27,600 SNPs and an average of missing data of 8.57% (Fig. 3a). A K = 4 was also confirmed by the MSN (Fig. 2a) in a reduced dataset accounting for a missing data rate > 5% (27,600 SNPs and 58 of 125 total accessions). The reduced dataset was also analyzed in FastStructure (Raj et al., 2014). As shown in Fig. 3 at K = 4, the Atlantic group was clearly separated from the Pacific group. The population structure result at this optimal K was consistent with the un-rooted phylogenetic tree, DPCA and PCA results (Figs. 2b, d and 3c); all suggested two primary clusters and a subtle substructure within. A large portion of accessions showed genetic admixture but a group of SNPs are exclusive from Atlantic individuals. At K = 4 three new subpopulations emerged mainly in accessions from the Pacific.

Population divergence times: Population divergence times were estimate among a tree built using SNAPP for individuals of three non-admixed genotypes found with Structure (Atlantic tall, Pacific tall and Pacific dwarf). Even though three different non-admixed genotypes were found for the Pacific coast of Colombia in the Structure analysis (K=4), one of these genotypes was only represented one individual in our dataset. This genotype was excluded from the analysis. Structure groups corresponding to Pacific accessions form a clade with a mean divergence time of 1683 YA (95% High posterior density HPD) 1403 – 2357 YA; while there is an outer clade corresponding to Atlantic accessions, with a mean divergence time of 5417 YA (95% HPD 3669 – 7810 YA) .

Outlier detection and selection evidence: We detected a total of three outliers in all runs using *BayeScan.* All loci had negative alpha values suggesting purifying or balancing selection. Top hit *BLAST* results correspond to loci No. 26018 of *Elaeis guineensis* (max. score of 279) annotated a *TPase-like pseudogene*. Loci No 10902 of *Phoenix dactylifera* clone dpB2Y annotated as a sex-determination region and Loci No 22204 of *Phoenix dactylifera* annotated as uncharacterized (Fig. 5b). Top hits using *HMMER* came from loci No 10902 (e-value 8.9e-45) of *Artemisia*

annua annotated as *Ribonuclease H-like* domain containing a protein involved in DNA integration processes. Loci No 26018 of *Cajanus cajan* and *Dendrobium catenatum* correspond to two uncharacterized proteins (Suppl. Table 2).

Discussion

Reduced representation libraries from high-throughput sequencing (HTS) allow analysis of SNPs throughout the *Cocus nucifera* genome, providing both neutral and putatively selected markers. Genotypes were successfully obtained for 125 samples from five different geographical regions in northern South America. This is the first genome wide approach used in coconut to discover SNPs, filling an important gap in an already published first genome draft (Xiao et al., 2017). In this study, we provide a large genome variation data set for cultivated coconut grown in Colombian coastal regions that have not been previously characterized at the molecular level. Additionally, because our results were congruent with previously published results about the independent origins of cultivated coconut, thousands of SNPs in representative coconut strains provided an unprecedented opportunity to finely resolve the domestication history of cultivated coconut. Population structure and phylogenetic analyses not only support the hypothesis that there is a strong genetic break between Atlantic and Pacific populations in northern South American but also provide diversification times for the split of these different genotypes. This pattern corroborates previous hypothesis for worldwide genetic diversity of coconut (Gunn et al., 2011).

Genetic diversity and population structure analysis for Colombia cultivated coconuts: According to mean F_{ST} values obtained here there is a high level of genetic differentiation between coconut populations of northern South America. However, considering genetic diversity of coconut varieties initially introduced in Colombia, it can be seen that there is a trend towards lower levels of genetic diversity in comparison to other world regions (Baudouin & Lebrun, 2009), where the status of coconut origin center is attributed, especially in the Southeast Asia. This implies smaller genetic distances between coconut accessions of Colombia, nonetheless, the difference between the Colombian Pacific and Atlantic populations are remarkable suggesting different origins in the cultivated coconuts in each coast (Table 1).

Here, we found a pattern of genetic differentiation between the two coasts Atlantic and Pacific. F_{ST} within samples from the same coast are smaller than F_{ST} of samples from different coasts, for example all the Atlantic – Pacific F_{ST} are on order of magnitude below than the Pacific – Pacific and Atlantic – Atlantic F_{ST}. This level of differentiation suggests long-term evolutionary divergence between two subpopulations, with independent origins of cultivated coconuts from within each lineage (Table 1). Highest genetic diversity was found in tall cultivars from the Atlantic, while lowest diversity was found in populations from the Pacific where dwarf palms were common (Fig. 1). This is similar to reports by Meerow (2003). Average values of genetic diversity for dwarf coconut populations corroborate reports by previous studies that used microsatellites, for example, an average value of 0.348 observed by Perera et al. (1998), and a mean value of 0.222 described by Meerow et al. (2003).

Our Structure (K=4) data suggest extensively admixed genotypes for Colombia. However, we identified four non-admixed genotypes including most of the total genetic variation in the region with minimum redundancy, except for one accession from the Pacific from which we only have one individual. A high proportion of hybrids with various degrees of admixture from these genotypes were present in our dataset. A group of SNPs are exclusive to the Atlantic populations while SNPs from the Pacific populations were found in hybrids of the Atlantic (Fig. 2 and 3), indicating gene flow is unidirectional. We could be overestimating K, due to the wide spread pattern of admixture; plots with K=2 and K=3 are also included in supplemental material (Fig. S3).

Bayesian analyses reveal two highly genetically differentiated subpopulations that correspond to the Pacific and Atlantic oceanic basins. This pattern was initially revealed at a global scale (Gunn et al., 2011) and is now confirmed at a local scale using genomics in northern South America Pacific coconuts showing genetic substructure and higher inbreeding coefficients (Fig. 1-3).

Diversification times for main cultivated coconut varieties at a global scale: Here we present for first time diversification times using indirect evidence and fossil calibrated phylogenies for three non-admixed genetic populations detected for northern South America (tall Atlantic, tall Pacific and dwarf Pacific). First, we used an indirect calibration for the split of the tall Atlantic coconut populations from the Pacific coconut populations. This was based on the assumption that coconut was apt for human cultivation and that wild coconut populations were already mixed with cultivated ones because of human movements. This occurred sometime after catastrophic floods caused by the ice sheet collapse in North American glacier lakes about 8000 – 14,000 YA (Blanchon & Shaw, 1995). These events resulted in the submergence of a land mass which had been the center for the development of paddy rice by agriculturally-based civilizations in Indo-Malaysia (Oppenheimer, 1998). Period after which these human populations have to migrate north with their crops, including coconut. Also, a carbon dated fossil calibration reported for semi-domesticated coconut from the Pacific was used. A coconut fruit fossil with intermediate morphological characters between domesticated and cultivated coconuts was used to calibrate the separation of two pacific coconuts populations (tall pacific and dwarf pacific) reported here (Lepofsky et al., 1992).

Diversification between tall Atlantic coconut populations and all Pacific coconut populations (tall and dwarf) happened during the Holocene about 5416 YA [95%HPD: 3669 - 7809]. While the diversification between tall Pacific already cultivated populations and dwarf Pacific domesticated coconut populations happened during the Holocene about 1682 YA [95%HPD: 1402 – 2356] (Fig. 4). Even though wild coconut fossils have been found and dated to ca. 60 MYA (Gomez-Navarro et al., 2009), based on our calibration we infer that diversification between cultivated and domesticated coconut populations has happened very recently and in few generations (Fig. 4). Divergence between Atlantic and Pacific populations occurred approximately in 50 generations, assuming a coconut palm can last up to 100 years (Harries, 1978). More surprisingly domestication events in the Pacific between tall cultivated coconuts to dwarf highly productive coconuts seem to have happened in less than 10 generations (Fig. 4).

Putative selection in protein coding genes: Results indicate that selection may also be driving differentiation between four populations of *C. nucifera* identified by Structure (Fig. 3). SNPs mapped to the *C. nucifera* partially assembled genome showed evidence of selection at three putatively selected loci (22204, 10902, 26018). These loci were related to sex determination, DNA binding (Zn finger enzyme) and a *TPase* pseudo gene respectively (Fig. 5). Outliers detected using HmmerWeb had functions generally related to DNA integration, nucleic acid binding and uncharacterized proteins (Suppl. Table 2). Analyses of outlier loci reveal a signal of isolation between Atlantic and Pacific populations, suggesting that selection and local adaptation play a key role in driving genetic change among these regions. Outlier tests indicate that allele frequencies at a subset of the SNPs appear to be driven by selection, in particular for genes with vital functions such reproduction, respiration and defense (Fig. 5). Candidate genes may have been artificially selected during domestication of cultivated varieties.

Conclusion: Genotyping by sequencing data has proved useful and reliable for the identification of high quality SNPs. We investigated genetic diversity and structure of coconut populations of northern South America. The combination of Bayesian and Hierarchical clustering tools proved to be effective in elucidating population genetic structure of coconut genotypes since the two methods corroborate each other well. Our results propose a dated evolutionary hypothesis for the independent origins of cultivated coconut globally. Even though wild coconut fossils have been found and dated to ca. 60 MYA, our calibration suggest that diversification between cultivated and domesticated coconut populations has happened very recently and in few generations.

Information generated in this study will contribute to the knowledge of coconut genotypes and genetic diversity present in Colombia. Red ring disease in coconut and other cultivated and native palms in the Neotropics have reached epidemic levels. Knowledge of the northern South America genetic structure of the coconut, including regions with augmented levels of genetic diversity, may ultimately prove useful in targeting source populations for disease resistance and other crop improvement traits. This work is a first step towards future genome- wide association

mapping studies and the identification of SNP markers able to enhance the precision breeding for horticultural traits in cultivated coconut palms.

Acknowledgments

We thank Juan Carlos Bedoya from El Colegio Mayor de Antioquia, David Granada and Sara Ramirez from La Corporación para Investigaciones Biológicas for help with fieldwork. Diego Mauricio Riaño Pachon, Craig Barrett, Nhora Helena Ospina-Calderon, Alejandro Zuluaga and Ivan Soto-Calderón for comments on the paper. Juan David Pineda Cardenas for advice about computational resources used through EAFIT. This work was funded through the Colombian Research Grants Administration System (COLCIENCIAS) to GPC and TA 221371353189.

Literature cited

Aljohi, H. A., Liu, W., Lin, Q., Zhao, Y., Zeng, J., Alamer, A., ... & Yu, J. (2016). Complete sequence and analysis of coconut palm (Cocos nucifera) mitochondrial genome. PloS one, 11(10), e0163990.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Banbury, B. L., & Leache, A. (2014). phrynomics: An SNP toolkit.

Batugal, P. (2005). *Coconut genetic resources*. Bioversity International.

Baudouin, L., & Lebrun, P. (2009). Coconut (Cocos nucifera L.) DNA studies support the hypothesis of an ancient Austronesian migration from Southeast Asia to America. *Genetic Resources and Crop Evolution*, *56*(2), 257–262. https://doi.org/10.1007/s10722-008-9362-6

Blanchon, P., & Shaw, J. (1995). Reef drowning during the last deglaciation: Evidence for catastrophic sea-level rise and ice-sheet collapse. 8.

Bouckaert, R., & Vaughan, T. G. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, *29*(8), 1917–1932.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular ecology*, *22*(11), 3124-3140.

Chester K. The origin, variation, immunity and breeding of cultivated plants. Selected writings of N. I. Valvilov; trans Chester KS. New York, NY: The Ronald Press Company; 1951. 364

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & McVean, G. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156-2158.

Earl, D. A. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*(2), 359–361.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, *14*(8), 2611–2620.

Fuller, D. Q. (2007). Contrasting patterns in crop domestication and domestication rates: recent archaeobotanical insights from the Old World. *Annals of Botany*, *100*(5), 903-924.

Foll, M., & Gaggiotti, O. (2008). A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, *180*(2), 977–993. https://doi.org/10.1534/genetics.108.092221

Food and Agriculture Organization of the United Nations. (2015). *FAO statistical pocketbook 2015: World food and agriculture.*

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic acids research, 31(13), 3784-3788.

Gauthier, J., Mouden, C., Suchan, T., Alvarez, N., Arrigo, N., Riou, C., ... & Peterlongo, P. (2017). DiscoSnp-RAD: de novo detection of small variants for population genomics. bioRxiv, 216747.

Geethanjali, S., Anitha Rukmani, J., Rajakumar, D., Kadirvel, P., & Viswanathan, P. L. (2018). Genetic diversity, population structure and association analysis in coconut (Cocos

nucifera L.) germplasm using SSR markers. *Plant Genetic Resources: Characterization and Utilization*, *16*(02), 156–168. https://doi.org/10.1017/S1479262117000119

Goudet, J. (2005). Hierfstat, a package for r to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, *5*(1), 184–186. https://doi.org/10.1111/j.1471-8286.2004.00828.x

Guevara, L., & Jáuregui, D. (2012). Aspectos morfológicos florales de 'Alto Criollo' y un nuevo morfotipo de *Cocos nucifera* L (Arecaceae: Arecoideae) en Venezuela. Eenstia 22 (1) 2012: 23-36

Gunn, B. F., Baudouin, L., & Olsen, K. M. (2011). Independent Origins of Cultivated Coconut (Cocos nucifera L.) in the Old World Tropics. *PLoS ONE*, *6*(6), e21143. https://doi.org/10.1371/journal.pone.0021143

Harries, H. C. (1978). The evolution, dissemination and classification of Cocos nucifera L. *The Botanical Review*, *44*(3), 265–319. https://doi.org/10.1007/BF02957852

Hubisz, M. J., Falush, D., Stephens, M., & Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, *9*(5), 1322–1332.

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population

structure. *Bioinformatics*, *23*(14), 1801–1806. https://doi.org/10.1093/bioinformatics/btm233

Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405.

Kamvar, Z. N., Brooks, J. C., & Grünwald, N. J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in genetics*, *6*, 208.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357.

Lebrun, P., Berger, A., Hodgkin, T., & Baudouin, L. (2005). Biochemical and molecular methods for characterizing coconut diversity.

Lee, T. H., Guo, H., Wang, X., Kim, C., & Paterson, A. H. (2014). SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC genomics*, *15*(1), 162.

Lepofsky, D., Harries, H. C., & Kellum, M. (1992). Early coconuts on Mo'orea Island, French Polynesia. *The Journal of the Polynesian Society*, *101*(3), 299–308.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Meerow, A. W., Wisser, R. J., Brown, S. J., Kuhn, D. N., Schnell, R. J., & Broschat, T. K. (2003). Analysis of genetic diversity and population structure within Florida coconut (Cocos nucifera L.) germplasm using microsatellite DNA, with special emphasis on the

Fiji Dwarf cultivar. *Theoretical and Applied Genetics*, *106*(4), 715–726. https://doi.org/10.1007/s00122-002-1121-z

Meerow, A. W., Francisco-Ortega, J., Ayala-Silva, T., Stevenson, D. W., & Nakamura, K. (2012). Population genetics of Zamia in Puerto Rico, a study with ten SSR loci. Memoirs of the New York Botanical Garden, 106, 204-223.

Menon, K. P. V., & Pandalai, K. M. (1958). The coconut palm. A monograph. *The coconut palm. A monograph.*

Murray, M. G., & Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic acids research*, *8*(19), 4321-4326.

Nimmakayala, P., Levi, A., Abburi, L., Abburi, V. L., Tomason, Y. R., Saminathan, T., … Wehner, T. C. (2014). Single nucleotide polymorphisms generated by genotyping by sequencing to characterize genome-wide diversity, linkage disequilibrium, and selective sweeps in cultivated watermelon. *BMC Genomics*, *15*(1), 767.

Oppenheimer, M. (1998). Global warming and the stability of the West Antarctic Ice Sheet. *Nature*, *393*(6683), 325–332. https://doi.org/10.1038/30661

Pavan, S., Marcotrigiano, A. R., Ciani, E., Mazzeo, R., Zonno, V., Ruggieri, V., … Ricciardi, L. (2017). Genotyping-by-sequencing of a melon (*Cucumis melo* L.) germplasm collection from a secondary center of diversity highlights patterns of genetic variation and genomic features of different gene pools. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-016-3429-0

Perera, L., Russell, J. R., Provan, J., McNicol, J. W., & Powell, W. (1998). Evaluating genetic relationships between indigenous coconut (*Cocos nucifera* L.) accessions from Sri Lanka by means of AFLP profiling. *Theoretical and Applied Genetics*, *96*(3–4), 545–550. https://doi.org/10.1007/s001220050772

Prades, A., Dornier, M., Diop, N., & Pain, J.-P. (2012). Coconut water uses, composition and properties: A review. *Fruits*, *67*(2), 87–107. https://doi.org/10.1051/fruits/2012002

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Rambaut, A. (2016). FigTree v1. 4.3 software. Institute of Evolutionary Biology, University of Edinburgh.

Rambaut, Andrew, Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904.

Ribeiro, F. E., Baudouin, L., Lebrun, P., Chaves, L. J., Brondani, C., Costa, E. F. N., & Vencovsky, R. (2013). Genetic diversity in Brazilian tall coconut populations by microsatellite markers. *Crop Breeding and Applied Biotechnology*, *13*(4), 356–362. https://doi.org/10.1590/S1984-70332013000400006

Rosenberg, N. A. (2004). DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes*, *4*(1), 137–138.

Shi, A., Qin, J., Mou, B., Correll, J., Weng, Y., Brenner, D., … Dong, L. (2017). Genetic diversity and population structure analysis of spinach by single-nucleotide polymorphisms identified through genotyping-by-sequencing. *PloS One*, *12*(11), e0188745.

Stange, M., Sánchez-Villagra, M. R., Salzburger, W., & Matschiner, M. (2018). Bayesian Divergence-Time Estimation with Genome-Wide Single-Nucleotide Polymorphism Data of Sea Catfishes (Ariidae) Supports Miocene Closure of the Panamanian Isthmus. *Systematic Biology*, *67*(4), 681–699. https://doi.org/10.1093/sysbio/syy006

Tang, B., Tang, M., Chen, C., Qiu, P., Liu, Q., Wang, M., & Li, C. (2006). Characteristics of soil fauna community in the Dongjiao coconut plantation ecosystem in Hainan, China. *Acta Ecologica Sinica*, *26*(1), 26–32.

Taranto, F., D'Agostino, N., Greco, B., Cardi, T., & Tripodi, P. (2016). Genome-wide SNP discovery and population structure analysis in pepper (Capsicum annuum) using genotyping by sequencing. *BMC genomics*, *17*(1), 943.

Teulat, B., Aldam, C., Trehin, R., Lebrun, P., Barker, J. H. A., Arnold, G. M., … Rognon, F (2000). An analysis of genetic diversity in coconut (Cocos nucifera) populations from across the geographic range using sequence-tagged microsatellites (SSRs) and AFLPs.

Theoretical and Applied Genetics, *100*(5), 764–771. https://doi.org/10.1007/s001220051350

Torkamaneh, D., Laroche, J., and Belzile, F. (2016). Genome-Wide SNP Calling from genotyping by sequencing (GBS) data: a comparison of seven pipelines and two sequencing technologies. *PLoS ONE* 11:e0161333. doi: 10.1371/journal.pone.0161333

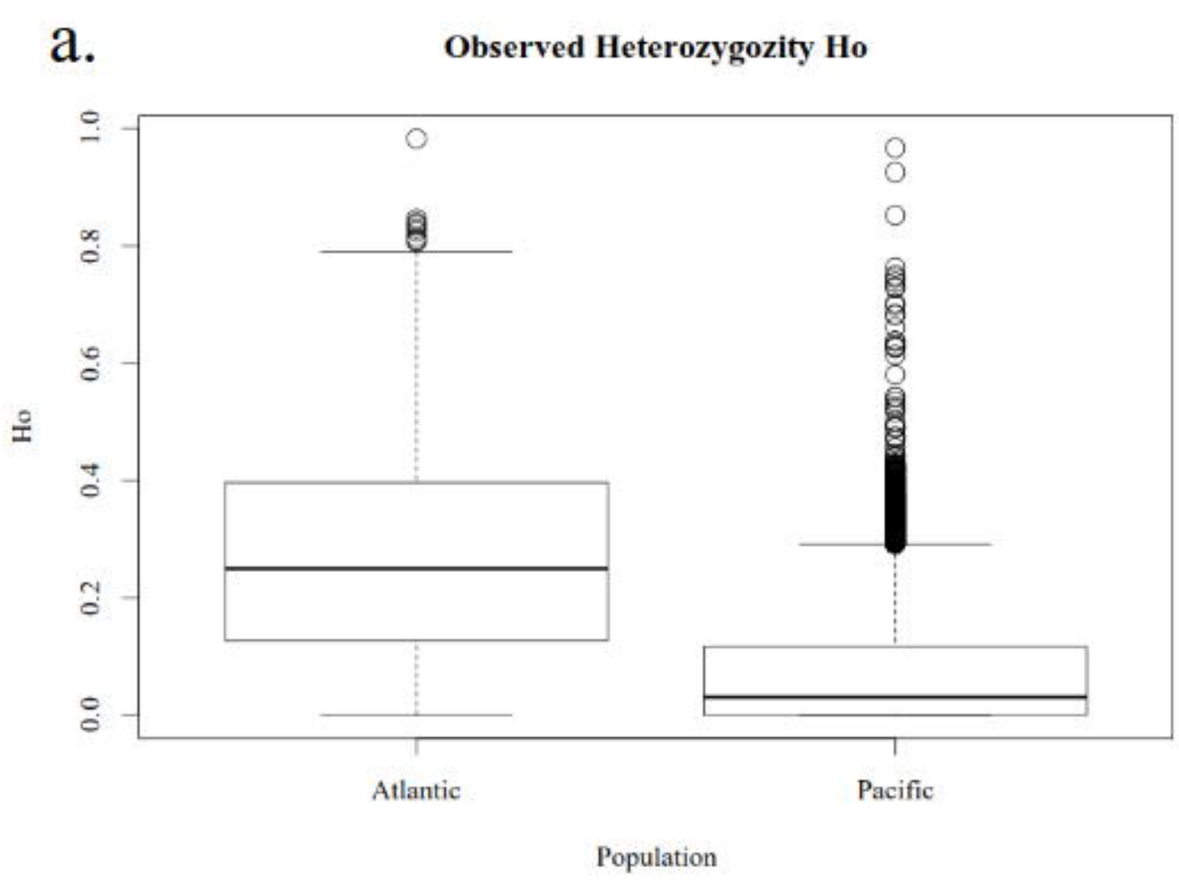
Wheeler, R. E. M., Ghosh, A., & Deva, K. (1946). Arikamedu: an Indo-Roman trading station on the east coast of India. *Ancient India*, *2*, 17-124.

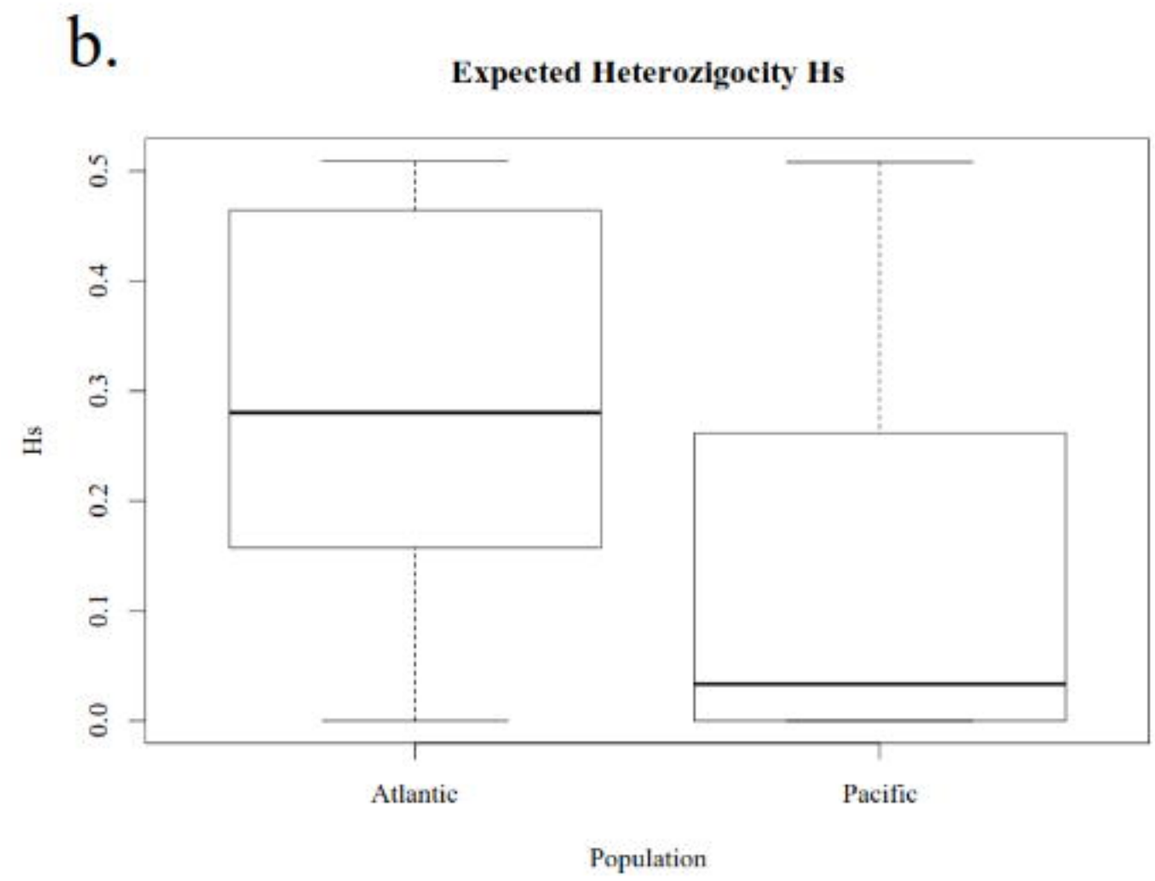
Xiao, Y., Xu, P., Fan, H., Baudouin, L., Xia, W., Bocs, S., … Zhou, L. (2017). The genome draft of coconut (Cocos nucifera). *GigaScience*, *6*(11), 1–11.

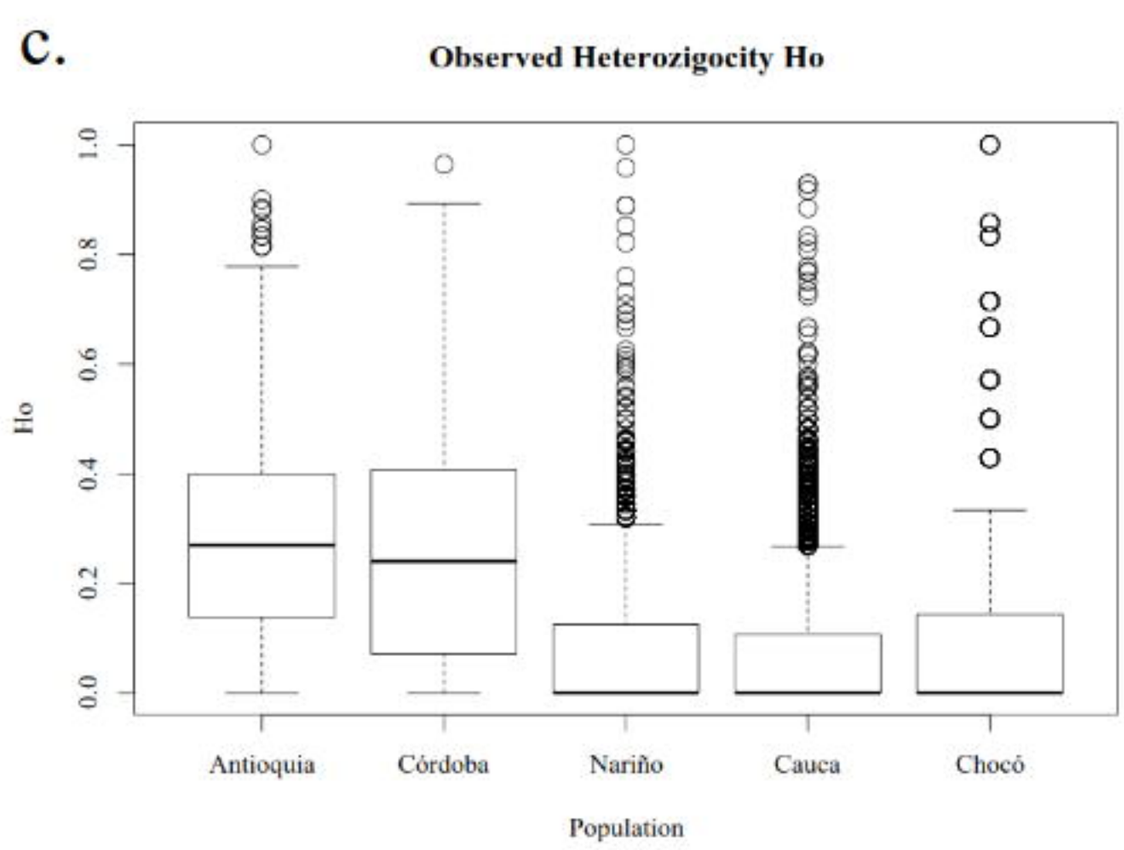
Xiong, H., Shi, A., Mou, B., Qin, J., Motes, D., Lu, W., … Wu, D. (2016). Genetic diversity and population structure of cowpea (Vigna unguiculata L. Walp). *PloS One*, *11*(8), e0160941.

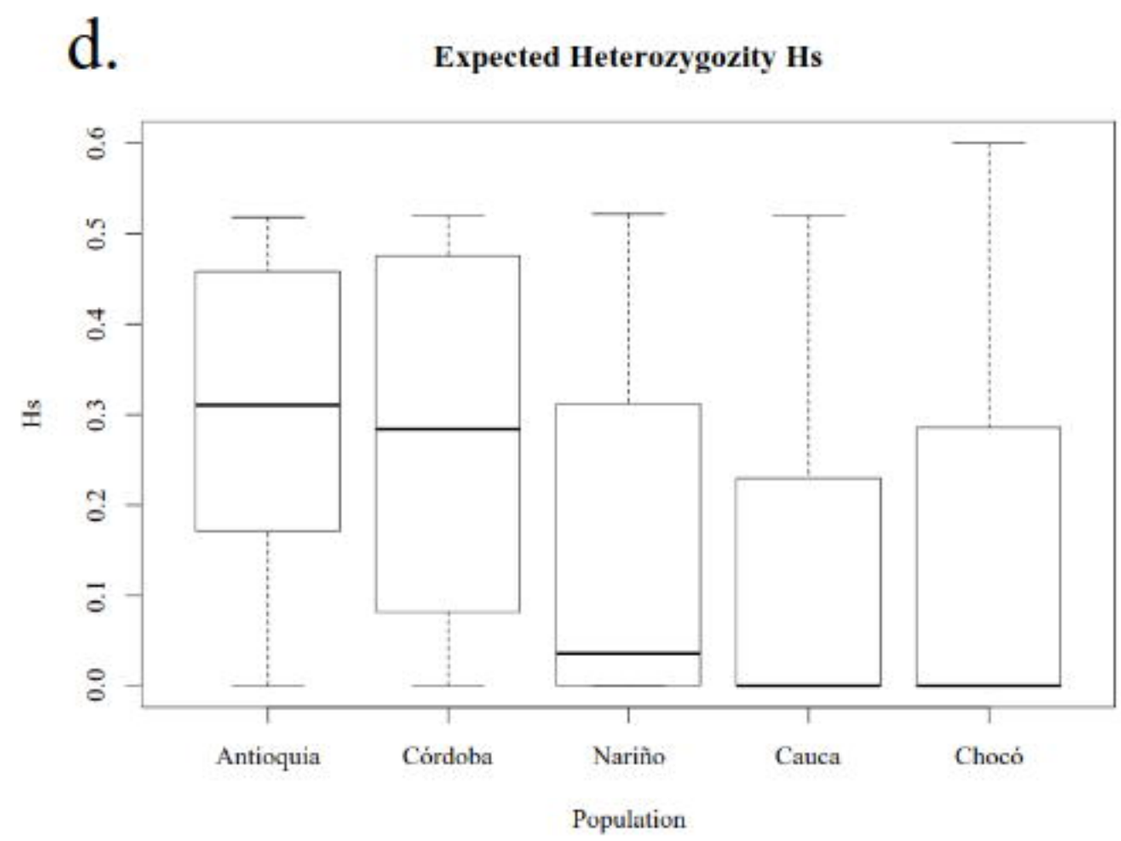
Xu, X., Liu, X., Ge, S., Jensen, J. D., Hu, F., Li, X., … Huang, L. (2012). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature Biotechnology*, *30*(1), 105.

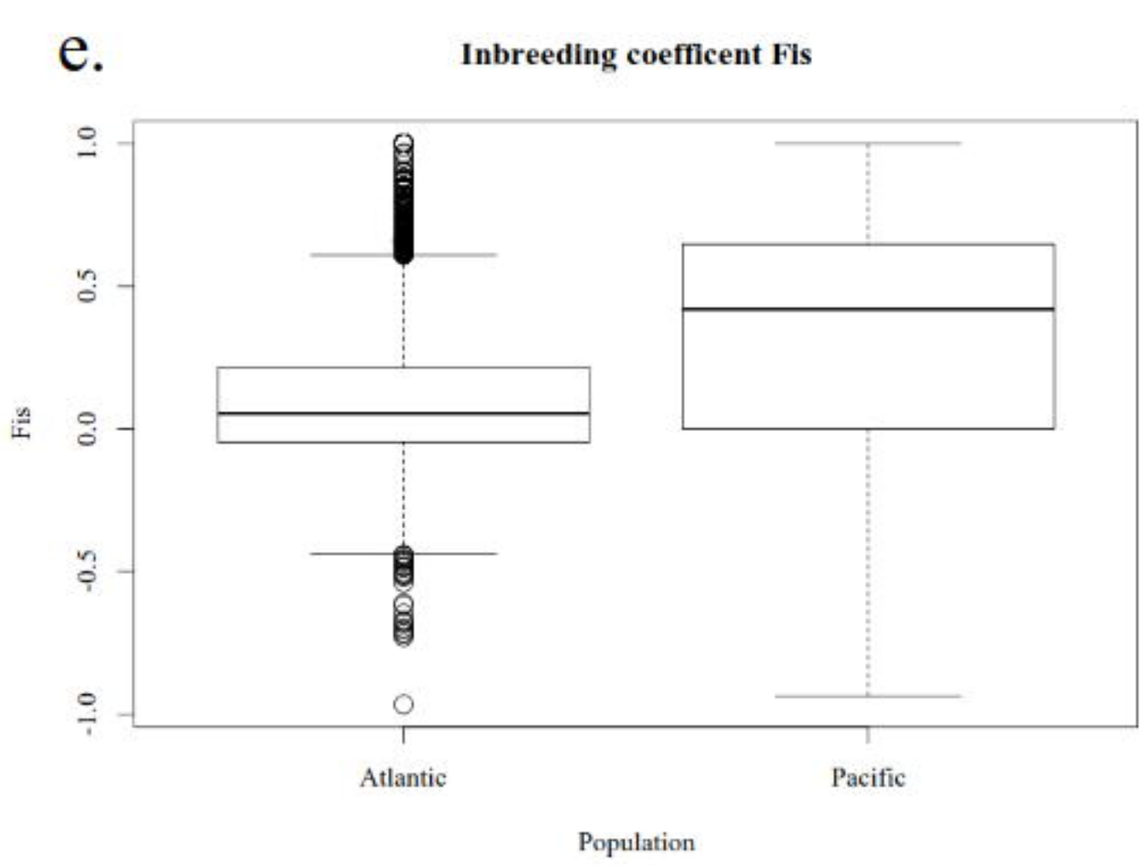
Yang, Y., Iqbal, A., & Qadri, R. (2018). Breeding of Coconut (Cocos Nucifera L.): The Tree of Life. In *Advances in Plant Breeding Strategies: Fruits* (pp. 673–725). Springer.

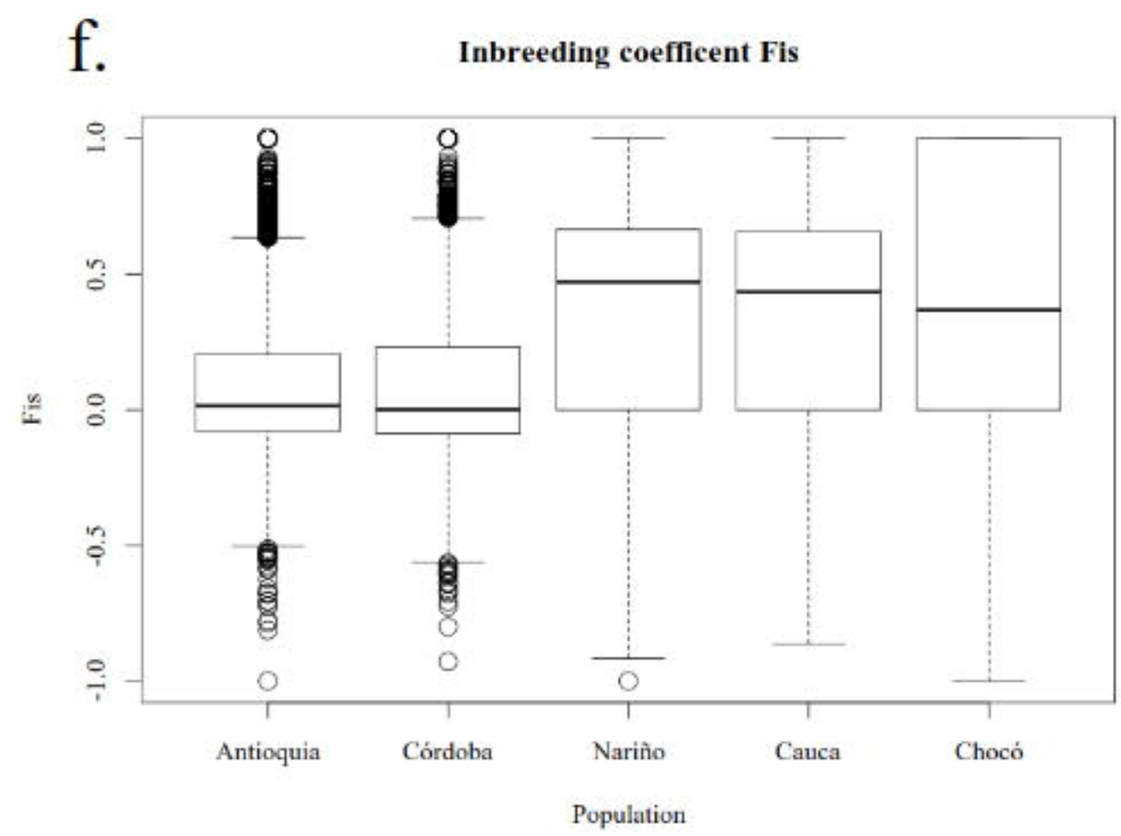

a.
Observed Heterozygozity Ho
Ho
0.0
0.2
0.4
0.6
0.8
1.0
Atlantic
Pacific
Population

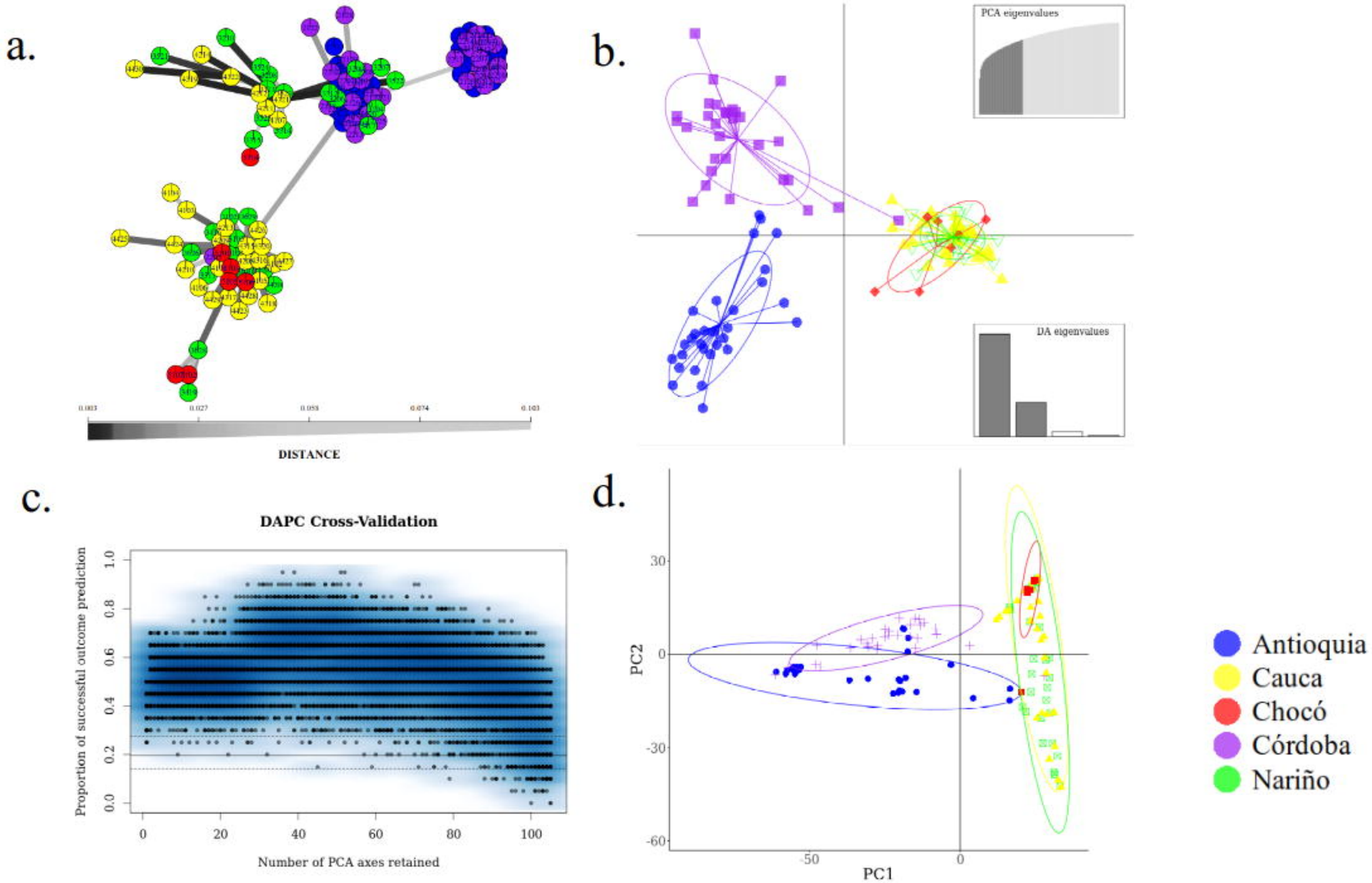

b.
Expected Heterozigocity Hs
Hs
0.0
0.1
0.2
0.3
0.4
0.5
Atlantic
Pacific
Population


c.
Observed Heterozigocity Ho
Ho
0.0
0.2
0.4
0.6
0.8
1.0
Antioquia
Córdoba
Nariño
Cauca
Chocó
Population

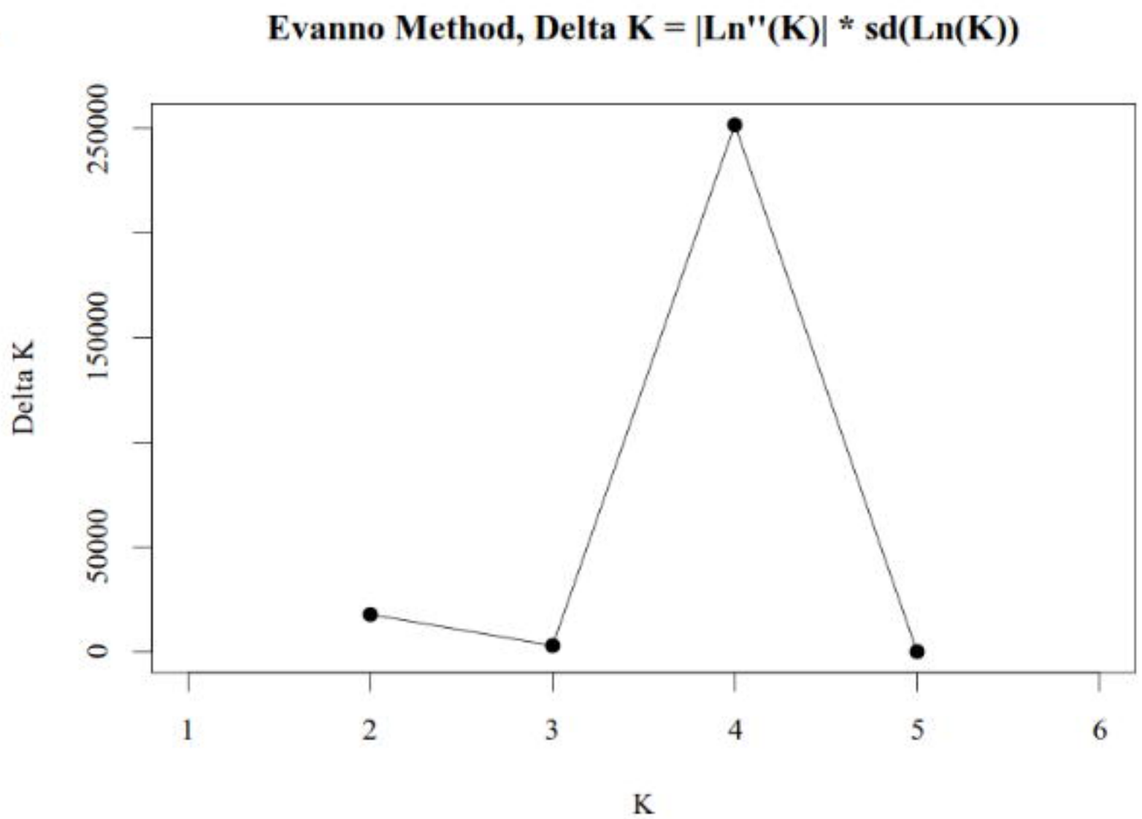

d.
Expected Heterozygozity Hs
Hs
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Antioquia
Córdoba
Nariño
Cauca
Chocó
Population


e.
Inbreeding coefficent Fis
Fis
-1.0
-0.5
0.0
0.5
1.0
Atlantic
Pacific
Population

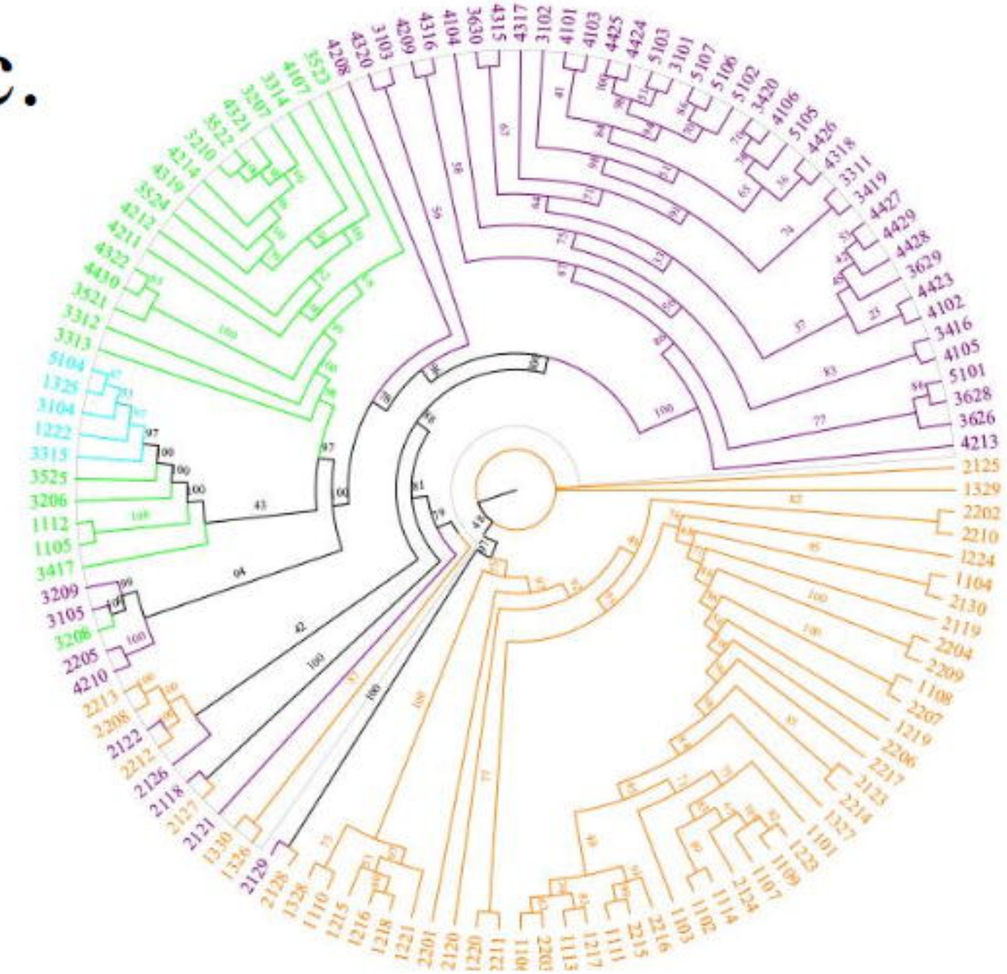

f.
Inbreeding coefficent Fis
Fis
-1.0
-0.5
0.0
0.5
1.0
Antioquia
Córdoba
Nariño
Cauca
Chocó
Population


a.
DISTANCE
b.
PCA eigenvalues
DA eigenvalues
c.
DAPC Cross-Validation
Proportion of successful outcome prediction
Number of PCA axes retained
d.
PC1
PC2
Antioquia
Cauca
Chocó
Córdoba
Nariño

a.

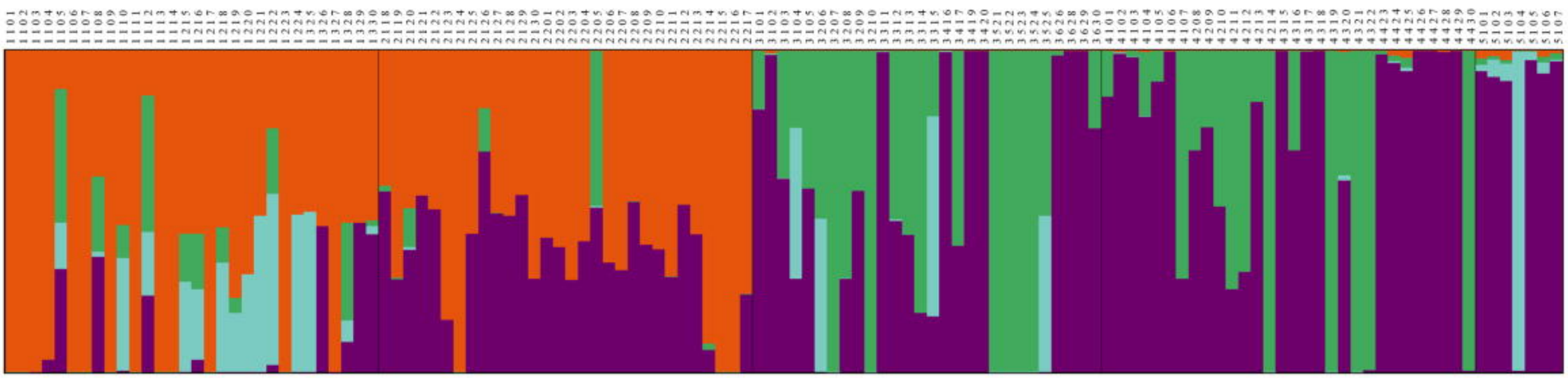


c.

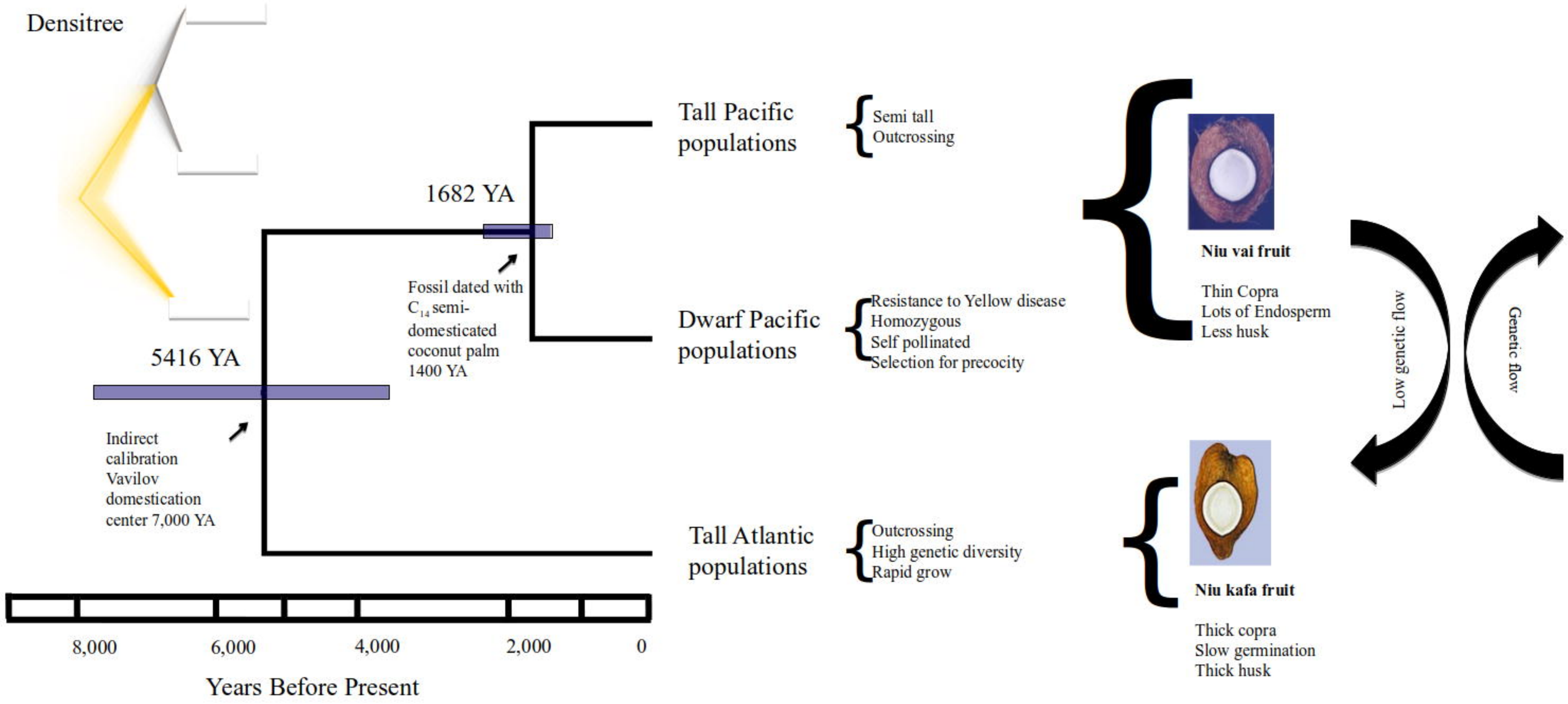


b.

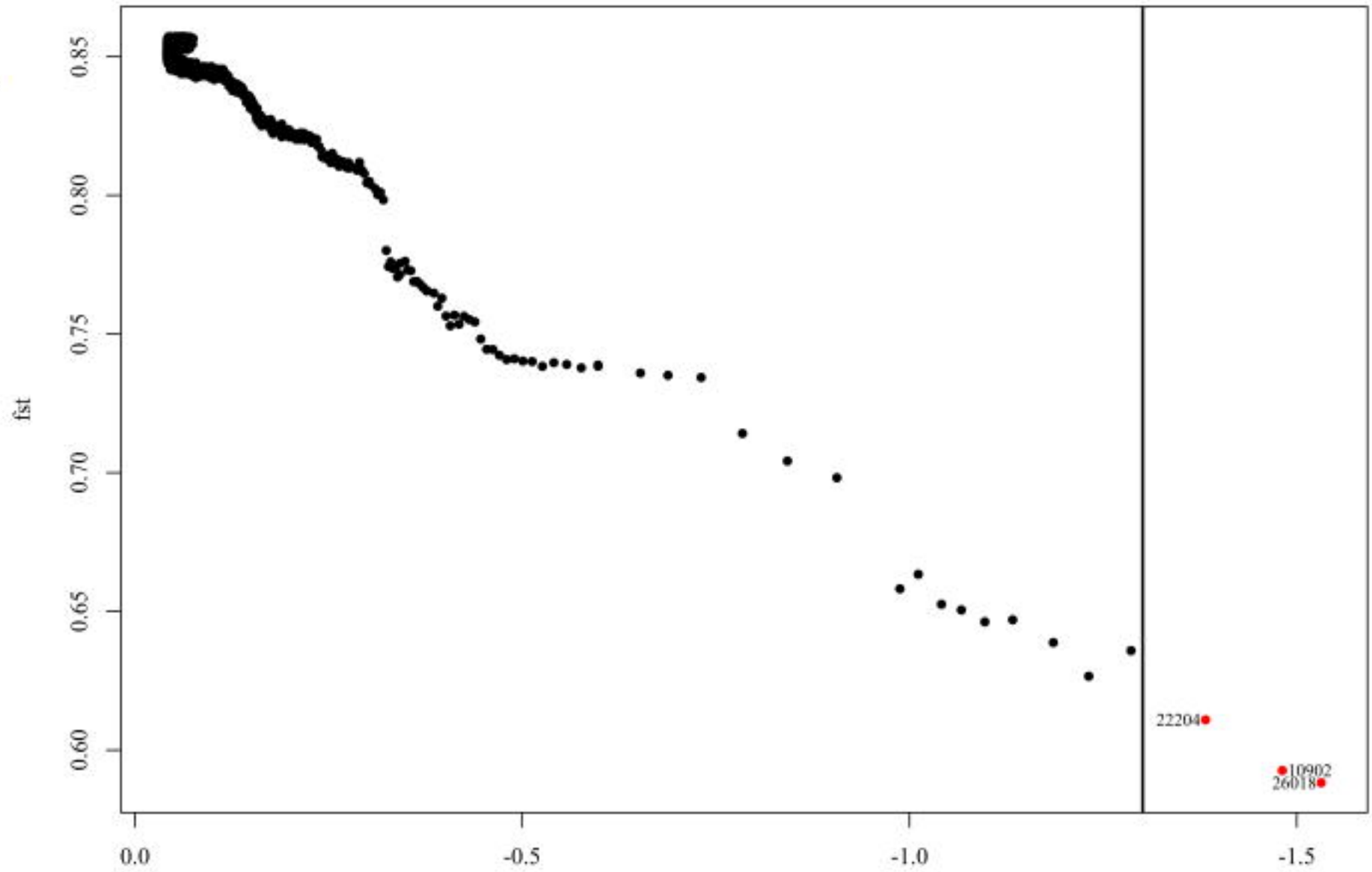
K=4



Antioquia Córdoba Nariño Cauca Chocó


Densitree
1682 YA
5416 YA
Fossil dated with C_{14} semi-domesticated coconut palm 1400 YA
Indirect calibration Vavilov domestication center 7,000 YA
Tall Pacific populations
Semi tall
Outcrossing
Dwarf Pacific populations
Resistance to Yellow disease
Homozygous
Self pollinated
Selection for precocity
Tall Atlantic populations
Outcrossing
High genetic diversity
Rapid grow
Niu vai fruit
Thin Copra
Lots of Endosperm
Less husk
Niu kafa fruit
Thick copra
Slow germination
Thick husk
Low genetic flow
Genetic flow
8,000
6,000
4,000
2,000
0
Years Before Present

a.



b.

Locus	Scaffold	Blast	Description	Accession	Organism	Blast Score	e-value (Blast)	q-value (Bayescan)
10902	scaffold1575	dmegablast	*Phoenix dactylifera* clone dpB2Y sex-determination region sequence	MH681002.1	*P. dactylifera*	218	2.00E-52	0.033007
10902	scaffold1575	blastn	*Phoenix dactylifera* clone dpB3Y sex-determination region sequence	MH681003.1	*P. dactylifera*	216	8.00E-52	0.033007
10902	scaffold1575	blastn	*Cocos nucifera* putative endoribonuclease dicer 3a-like protein 3a mRNA, partial cds	KJ851191.1	*C. nucifera*	30.1	0.56	0.033007
22204	scaffold36998	dmegablast	PREDICTED: *Ictalurus punctatus* PLAG1 zinc finger (plag1), Mrna	XM_017470994.1	*I. punctatus*	42.8	9.9	0.041475
22204	scaffold36998	blastn	PREDICTED: *Phoenix dactylifera* uncharacterized LOC103703419 (LOC103703419), mRNA	XM_026803080.1	*P. dactylifera*	50.9	0.067	0.041475
22204	scaffold36998	blastn	*Cocos nucifera* phospholipase D alpha 1-like protein mRNA, complete cds	KJ396258.1	*C. nucifera*	28.3	0.95	0.041475
26018	scaffold36998	megablast	*Elaeis guineensis* partial *TPase* pseudogene, clone EgES9	AJ507421.1	*E. guineensis*	250	3.00E-62	0.029406
26018	scaffold36998	dmegablast	*Elaeis guineensis* partial *TPase* pseudogene, clone EgES9	AJ507421.1	*E. guineensis*	279	8.00E-71	0.029406
26018	scaffold36998	blastn	*Elaeis guineensis* partial *TPase* pseudogene, clone EgES9	AJ507421.1	*E. guineensis*	279	8.00E-71	0.029406
26018	scaffold36998	blastn	*Cocos nucifera* chloroplast, complete genome	KX028884.1	*C.nucifera*	30.1	5.30E-01	0.029406