Genome-wide footprints in the carob tree (Ceratonia siliqua) unveil a new domestication pattern of fruit trees in the Mediterranean

Alex Baumel[1], Gonzalo Nieto Feliner[2], Frederic Medail[1], Stefano La Malfa[3], Mario Diguardo[3], Magda Boudagher-Kharrat[4], Fatma Mirleau[5], Valentine Frelon[1], Lahcen Ouahmane[6], Katia Diadema[7], Hervé Sanguin[6], and Juan Viruel[8]

[1]Aix Marseille University
[2]2. Real Jardín Botánico
[3]Università degli Studi di Catania
[4]Laboratoire Caractérisation Génétique des Plantes, Faculté des sciences, Université Saint-Joseph, B.P. 11-514 Riad El Solh, Beyrouth 1107 2050, Liban
[5]Aix Marseille Univ, Univ Avignon, CNRS, IRD, IMBE, Marseille, France
[6]Affiliation not available
[7]Conservatoire Botanique National Méditerranéen de Porquerolles
[8]Royal Botanic Gardens Kew

February 22, 2024

Abstract

Intense research efforts on phylogeography over the last two decades uncovered major biogeographical trends and renewed our understandings of plant domestication in the Mediterranean. We aim to investigate the evolutionary history and the origin of domestication of the carob tree that has been cultivated for millennia for food and fodder. We used >1000 microsatellite genotypes to identify carob evolutionary units (CEUs) based on genetic diversity structure and geography. We investigated genome-wide diversity and evolutionary patterns of the CEUs with 3557 SNPs generated by restriction-site associated DNA sequencing (RADseq). The 56 populations sampled across the Mediterranean basin, classified as natural, semi-natural or cultivated, were examined. Although, RADseq data are consistent with previous studies identifying a strong West-to-East genetic structure and considerable admixture in some geographic parts, we reconstructed a new phylogeographic scenario with two migration routes occurring from a single refugium likely located in South-Western Morocco. Our results do not favour the regionally bound or single origin of domestication. Indeed, our findings support a cultivation model of locally selected wild genotypes, albeit punctuated by long-distance westward dispersals of domesticated varieties by humans, concomitant with major cultural waves by Romans and Arabs in the regions of dispersal. Ex-situ efforts to preserve carob genetic resources should prioritize accessions from both western and eastern populations, with emphasis on the most differentiated CEUs situated in South-Western Morocco, South Spain and Eastern Mediterranean. Our study underscores the relevance of natural and seminatural habitats of Mediterranean forests and their refugia in the conservation efforts of tree crops.

Introduction

Fruit trees played a major role in the development of Mediterranean civilizations during the last millennia. Their evolutionary histories represent examples of plant evolution under three important drivers, geological, climatic and human, which have been defined as the Mediterranean triptych (Thompson 2020). Several tree species survived in refugia during the Pleistocene climatic changes, and suffered repeated range expansions and contractions, which shaped their genetic diversity and structure. Human activities constitute the most

recent of the three Mediterranean triptych drivers, but they had great consequences on shaping global biodiversity (Boivin et al., 2016). Humans have been modifying Mediterranean ecosystems for thousands of years, profoundly altering the forests (Quézel and Médail, 2003). As a result, it is difficult to document the evolutionary history of fruit trees, which may have cultivated, feral or wild populations in the same region. However, recent phylogeographic studies have revealed that the imprints of ancestral populations preceding agriculture are still present in the genetic diversity structure of Mediterranean cultivated tree species (Gros-Balthazard et al., 2017; Besnard et al., 2018). Identifying the oldest components of the genetic legacy is essential to conserve genetic resources in the Mediterranean region, and it will also improve our understanding of the domestication process. As a general pattern, domestication in the Mediterranean started in the East and was followed by human-mediated westward dispersal of crops across the basin (Zeder, 2008; Zohary and Hopf, 2012). However, recent studies suggest that domestication in the Mediterranean was a protracted process involving local resources from several diversity centers during which genetic admixture, within or between species, played a crucial role (Fuller et al., 2011; Purugganan, 2019; Thompson, 2020).

The carob, a common tree in traditional Mediterranean orchards, has been traditionally valued, and still is, for its ability to produce food and fodder on marginal lands, especially during unfavourable years. Domestication of the carob tree is known to have aimed at increasing the pulp in the fruit (Zohary, 2002), but new uses have recently emerged, such as ecological restoration of degraded land, production of bioethanol or the use of a galactomannan obtained from the seeds as food stabilizer. A recent review outlined the potential of carob for developing health-beneficial food products (Brassesco et al., 2021). Because the propagation of carob cultivars is done by grafting, it is assumed that the origin of its cultivation is linked to the development of grafting methods c. 3,000 years ago (Zohary, 2002; Meyer et al., 2012). As for several crops, the Near East and the Eastern Mediterranean were initially proposed as the center of domestication for the carob tree (Zohary, 2002; Ramon-Laca and Mabberley, 2004). However, a recent multidisciplinary phylogeographic study based on wild and cultivated carob trees revealed the existence of four main genetic groups across the Mediterranean with a strong west-east structure (Viruel et al., 2020), which has also been documented for other Mediterranean plants (Désamoré et al. 2011; Nieto Feliner 2014; Chen et al. 2014; Migliore et al. 2018; Garcia-Verdugo et al. 2021). Using coalescent simulations based on microsatellite data, the estimated divergence times between these genetic groups pre-date the Neolithic origin of agriculture (Viruel et al., 2020). This contrast with two genetic studies focused on carob cultivars (Caruso et al. 2008; La Malfa et al. 2014), which reported a lack of geographical structure and strong genetic admixture. However, a recent study of the world's largest carob cultivar germplasm collection based on microsatellite and plastid markers (Di Guardo et al., 2019) detected a genetic cluster in South Spain sharing ancestry with genotypes from Morocco and separated from cultivars of West Spain, Italy and the eastern Mediterranean. As postulated in Viruel et al. (2020), integrating the results from these studies on wild and cultivated carob trees supports a regional use and domestication of local carob in several parts of the Mediterranean. Di Guardo et al. (2019) also emphasized that mixed ancestry found in current cultivars was the result of the diffusion of selected productive, female or hermaphrodite genotypes via grafting. Cultivated and wild carob trees are often spatially close to each other and seeds are efficiently dispersed by cattle. Therefore, recurrent cultivated-wild genetic admixture could have determined diffuse domestication effects with potential impact throughout the whole carob geographical range including the wild trees. Nevertheless, the effects of domestication were not homogeneous across the Mediterranean basin. In Andalusia and Morocco, carob orchards were less intensive than in the rest of the distribution range (Di Guardo et al., 2019). In these two areas, the carob pods from cultivars have a low pulp content, similar to those of the wild type. By contrast, in the eastern and central Mediterranean areas, especially in Sicily, Crete and Cyprus, carob cultivation is more intensive and supports several traditional uses, suggesting an ancient history of selection and domestication. Recently, in agreement to this pattern, Baumel et al. (2018) showed, through a study of floristic diversity on a Mediterranean scale, that carob habitats were more heterogeneous in the west than in the east of the basin. To confirm the historical scenarios that have shaped the current diversity of both wild and cultivated carob, genomic-based approaches are required.

In this study, we aim to clarify carob evolutionary history and to assess the contribution of agriculture to the genome-wide diversity of carob. Our appraisal is based on two facts: carob cultivation and selection

efforts were not homogeneous throughout the Mediterranean and carob populations are currently observed following a gradient of ecological conditions from natural habitats to cultivated lands. We hypothesize a stronger impact of domestication on genetic diversity in central and eastern Mediterranean populations. We also postulated a pattern of gene flow from east to west due to the spread by Greeks and Arabs of already domesticated carob trees (Ramon-Laca and Mabberley, 2004).

Our first objective was to identify geographical boundaries among carob population units with respect to genetic diversity structure. Building from the SSR polymorphism and SNP data obtained by Viruel et al. (2018, 2020), we aimed at delimiting geographically homogenous genetic groups of carob populations (hereafter called CEUs for Carob Evolutionary Units). Then, using these CEUs, we investigate carob genome-wide diversity and differentiation with data developed for the present study from a reduced-representation genomic approach, restriction associated DNA sequencing (RADseq), which was successful to decipher evolutionary history in several tree species (Hodel et al., 2017; Borrell et al. 2018; Warschefsky and von Wettberg, 2019; Hipp et al., 2020). Our second objective was to assess the potential impact of agriculture on the genome-wide diversity of the carob tree. We performed a comparative diversity analysis and searches for candidate loci under opposite statuses (natural versus cultivated) of carob populations. Our third objective was to reconstruct the history of CEUs including population splits and gene flow.

Materials and Methods

Plant material

For this study, we used the materials collected from 78 populations of*Ceratonia siliqua* across the Mediterranean basin as described in Viruel et al. (2020). Based on field observations, the habitat of each population was recorded as natural, semi-natural or cultivated. Habitats where human impact is low and there is no evidence of recent land use were recorded as 'natural'. 'Semi-natural habitats' lack cultivation but show the clear presence of human activities such as pasture, old farming or sometimes scattered almond, carob or olive trees from which fruits are occasionally harvested. Habitats containing carob trees in specialized or consociated orchards (with cereals or other fruit trees) –even if abandoned–, were recorded as 'cultivated'.

Preliminary delimitation of *Ceratonia siliqua* units using microsatellite data

We used microsatellite data for 17 loci and 56 localities across the Mediterranean basin as described in Viruel et al. (2020). We selected localities with at least ten carob genotypes per population (totalling 1020 trees). We also scored 15 single nucleotide polymorphisms present in the flanking regions of the 17 SSR loci following Viruel et al. (2018). We calculated genetic differentiation between localities using D index (Winter 2012, Mmod R package) and converted it into Euclidean genetic distances based on the coordinates of a NMDS (Non-metric Multidimensional Scaling; vegan R package). The Euclidean genetic distances and Euclidean geographical distances were then processed by ClustGeo (Chavent et al. 2018; clustgeo R package), which relies on Ward's method to minimize intra-group variance for both geographical and genetic distances. K and ? are the main parameters. K is the number of clusters and ? is a mixing parameter determining the weight of the spatial constraint. Several values of ? were tested to optimize the spatial contiguity of populations without deteriorating genetic differentiation structure. Several values of K were also tested. This clustering method provided genetically homogeneous groups of spatially adjacent carob populations, named "CEUs" (Carob Evolutionary Units), which were used in subsequent analyses. A neighbor-joining tree based on pairwise genetic differentiation (Gst, Mmod package) among the seven CEUs was built to display the overall differentiation structure.

RADseq methodology and sequencing

CEUs were used to select 376 samples representative of the genetic diversity and structure of *C. siliqua* for RADseq. Eight samples of the only other species in the genus, *C. oreothauma,* were added. Genomic library preparation and sequencing were conducted by Microsynth ecogenics GmbH (Blagach, Switzerland). DNA samples (200-400 ng input) were digested with the restriction enzymes EcoRI/MSeI following heat inactivation according to the manufacturer's protocol (New England Biolabs, NEB). Fragments between

500 and 600 bp were selected by automated gel cut, Illumina Y-shaped adaptors were ligated, and ligation products were bead purified. Each library was then individually barcoded by PCR using a dual-indexing strategy. Individually barcoded libraries were pooled and subsequently purified before sequencing on an Illumina NextSeq platform (300 million of 75 bp reads per run).

Bioinformatic pipeline to extract and filter SNPs from RADseq data

The bioinformatic approach used in this study is summarized in Fig. 1. FASTQC reports (multiQC, Ewels et al., 2016) were used to exclude 26 *C. siliqua* and 4 *C. oreothauma* samples due to low sequencing coverage. The remaining 354 samples (350 *C. siliqua*and 4 *C. oreothauma*) had an average of 3 million raw reads, ranging between 0.7 and 15 million. Assembly (Fig. 1) was performed using ipyrad (Eaton and Overcast 2020) in a high-performance computing cluster (HPC pytheas). Thereafter, a "locus" is a RADseq marker of 65 bp, that resulted from the ipyrad workflow, and a "SNP" is a polymorphic position of a specific locus, considering that a locus can contain several SNPs. We conducted an assembly limited to *C. siliqua* following two steps. First, we selected 36 samples representative of the diversity in *C. siliqua* with high sequencing coverage to conduct four *de novo* assemblies with varying thresholds of clustering reads (*clust_threshold*) between 0.9 and 0.96, the minimum number of samples per locus (*min_sample_locus*) fixed to 30, the minimum depth (*mindepth_statistical*) for base calling fixed to 8, the minimum size of reads fixed to 50 and the first 5 bp of both ends of locus trimmed. The maximum number of indels and the maximum percentage of SNPs per locus were set to 1 and 10%, respectively. Default values were used for all other parameters. Considering the number of loci, heterozygosity, and error rates, we estimated an optimal clustering threshold of 0.94 (see Tab. S1 supplementary material). A reference file was generated by extracting the first sequence of each locus with pyrad2fasta script (available on https://github.com/pimbongaerts). Second, this file was used for a subsequent reference assembly for the 350 samples with the same parameters as previously except for the minimum number of samples per loci, which was fixed to 45. The dataset constructed using this reference assembly contained more than 64% missing data. The vcf file obtained from ipyrad was then processed to run a principal component analysis (PCA) using the glPca function of adegenet R package, and a neighbor-joining tree, based on pairwise genetic differentiation (Gst, Mmod R package) among the seven CEUs, to check that the structure from RAD markers was congruent with previous analysis based on microsatellite markers.

We used *Matrix condenser* software (de Medeiros and Farrell, 2018; de Medeiros 2019) to visualize the samples causing this high missing data rate. We filtered samples to optimize locus coverage and to keep the sampling equilibrium among CEUs, and reconducted the reference assembly with ipyrad on a new set of 190 samples. The data was then reduced to one SNP by locus, based on two criteria: SNP having the maximal minimum allelic frequency per locus and only keeping SNPs with allelic frequency above 1.05 % (i.e., rarest allele present in at least two individuals). To examine the carob genetic diversity based on neutral processes, such as genetic drift and gene flow, we reduced the effect of outlier loci (i.e. having unexpectedly high differentiation among populations, which could have a great effect on the variance among populations). We used Outflank software (Whitlock & Lotterhos, 2015), which produces a lower false-positive rate compared to other methods (e.g. Silliman 2019), considering CEUs as populations. The false discovery rate (qvalue) was fixed to 0.05. BLAST searches using the nucleotide NCBI database limited to flowering plant sequences were conducted for outliers, which were mostly assigned to plastid (pDNA) or mitochondrial (mtDNA) genomes. Nuclear sequences were then discarded whereas pDNA and mtDNA sequences were used as a reference in a new ipyrad assembly, with adapted parameters for haploid data. The sequences of these new sets of markers were concatenated to construct separate alignments of mitochondrial and plastid haplotypes for the 190 carob trees.

Hereafter, different analyses were conducted on three datasets: the genome-wide nuclear data (GWN), was used for the analysis of genetic differentiation, population divergence history and gene flow, whereas the mitochondrial and plastid-based haplotypes datasets were used in phylogenetic reconstructions (MH and PH datasets respectively).

Raw and filtered vcf files, as well as custom R scripts are available in ####: ####.

Population genomic analysis

The GWN dataset was converted from genind to genlight, vcf and treemix data formats using dartR (Gruber et al., 2018) and radiator (Gosselin 2020) R packages. Population structure analysis was performed using snmf (R LEA package, Frichot and Francois 2015) which estimates admixture coefficients from the genotypic matrix assuming K ancestral populations. Snmf runs fast even on large datasets and without loss of accuracy compared to other Bayesian modelling software such as ADMIXTURE (Alexander et al. 2009; Frichot et al. 2014). Snmf was run for K = 2 to 15, 100 repetitions, regularization parameter of 250 and 25% of the genotypes were masked to compute the cross-entropy criterion. Barplots showing ancestry coefficients were obtained with the compoplot function in adegenet R package (Jombart and Ahmed 2011), with genotypes sorted according to CEUs. To visualize the genetic diversity structure, we performed a PCA using the glPca function of adegenet R package. Pairwise differentiation among CEUs (Gst) were also computed with the Mmod package. The R scripts of these analyses are available in #### web site: ####.

A new reference assembly was performed to add *C. oreothauma*samples to the GWN dataset and obtain one including the two species. As previously, to obtain unlinked SNPs, we retained one SNP per loci, choosing the one with the highest frequency. Using PAUP*4.0a (Swofford, 2018), we built a coalescent-based tree with the SDV quartet method (Chifman & Kubatko 2014). Ten million randomly selected quartets were analyzed and node support was assessed by 1,000 bootstrap replicates. We used the 'distribute' option for heterozygous sites.

The GWN dataset without *C. oreothauma* genotypes was used to perform a Treemix analysis (Pickrell and Pritchard 2012) to infer population splits and mixtures across the evolutionary history of the carob tree. Treemix builds a backbone tree based on population allelic frequency without gene flow and then adds reticulate branches, representing gene flow, aiming at improving the fit of the data. For this analysis, we used the CEUs as populations. One of the CEUs was fixed as an outgroup according to the results obtained in the phylogenetic analysis (see SDV quartet method in Results). As previously, only unlinked SNPs were used.

Assemblies based on MH and PH sequences failed to include *C. oreothauma* . The MH and PH matrices were analyzed with IQTREE with the default parameters. The F81+F+I and TN+F+I+G4 models were retained for the MH and PH matrices, respectively.

Footprints of domestication were investigated by estimating genetic diversity and/or the presence of candidate loci under selection (outliers) for the CEUs. A stronger effect of domestication is expected in cultivated habitats compared to natural habitats and also in the eastern CEUs compared to the western ones. Therefore, the analyses were organized according to combinations of these two factors. Genetic diversity indexes (H_{obs}, H_{exp}, f and rbardD) were estimated using *hierfstats* and *poppr* R packages (Goudet 2005; Kamvar et al. 2014). Outliers were searched with the Outflank method with a FDR threshold of 0.05 as described above when building the GWN dataset considering combinations of habitats and CEUs as populations.

Results

Preliminary delimitation of Carob Evolutionary Units (CEUs) using microsatellite data

Based on genotypes from 17 microsatellite loci (including SSR and SNPs) and geographical coordinates, we grouped 1,020 carob trees into CEUs (Fig. S1 in supplementary material). The Ward cluster method, which considers geographic and genetic distances, found seven CEUs grouped in two clusters. The western cluster included populations from South Morocco (SM), South-West Spain (SWS) and South-East Spain (SES). The eastern cluster included the remaining four CEUs: North Morocco (NM), Central Mediterranean (CM), North-East Mediterranean (NEM) and South-East Mediterranean (SEM). Based on microsatellite markers, the overall genetic differentiation among these seven CEUs is 16% (Gst). Pairwise Gst values visualized by a neighbor-joining tree (Fig. S1C in supplementary material) confirmed the Ward clustering scheme, but placed SEM and SM in intermediate positions along the NJ tree instead of grouping these two CEUs together.

Assembly of RADseq markers and identification of outliers

The *de novo* assembly conducted on 36 samples with a clustering threshold of 0.94 and a minimum number of samples by loci of 30, produced a mean number of 48,828 loci by sample (Tab. S1) reduced to 13,371 loci retained after ipyrad filtering. The sequences of these loci were used as a reference for the assembly of 350 genotypes, which resulted in 10,012 loci with an overall missing data rate of 64%. The structure of genetic diversity revealed by this RADseq data set is congruent with the microsatellite dataset (Fig. S1) and with a west-east pattern of differentiation (Fig. S2A in supplementary material), even picturing a more pronounced longitudinal structure. However, SEM and NM exchange their positions in the RADseq dataset: NM appears in an intermediate position between west and east CEUs close to CM whereas SEM forms a group with NEM (Fig. S2B).

The detection of the genotypes and loci having the highest missing rate by Matrix condenser, and their subsequent removal, produced a matrix with 190 samples and 12,767 loci, with 14% missing rate. After filtering to keep one SNP by locus, the dataset included 190 samples for 3,613 loci (or SNPs), with 9.5% overall missing data and 3.5% median missing rate by genotype. The Outflank method detected 56 outlier markers within the dataset of 190 samples by 3,613 SNPs (Fig. S3 in supplementary material). The global differentiation computed for these outliers was six times higher than for other SNPs (Gst 0.74 vs. 0.12). Nucleotide BLAST searches revealed that 27 outliers were assigned to plastid genome, 14 to the mitochondrial genome, 11 to nuclear genomes and 4 were not assigned. Excluding the outliers detected by Outflank, the GWN dataset included 3,557 SNPs. The nuclear matrix produced for phylogenetic inference after filtering out samples with missing data rates above 20%, contained 171 *C. siliqua* and 4 *C. oreothauma* genotypes for 3,284 unlinked SNPs and 8% missing data rate. The MH matrix contained 190 *C. siliqua* genotypes for 946 nucleotides, 17 variable sites and 60% missing data. The PH matrix contained 190 *C. siliqua* genotypes for 1,897 nucleotides, 31 variables sites and 47 % missing data.

Phylogenetic and genetic structure of carob trees across the Mediterranean based on RADseq data

Phylogenetic analysis of MH and PH concatenated in two alignments revealed the same pattern with two clearly differentiated haplogroups with branch support over 95% (Fig. S4 in supplementary material). In the Western Mediterranean CEUs (SES, SWS, NM and SM) a single haplogroup represented 85 % and 94.5 % of plastid and mtDNA haplotypes respectively. These western mtDNA and pDNA haplogroups were observed in only 11 % of the samples from central and eastern CEUs. A central-eastern haplogroup was found across CM, NEM and SEM CEUs, and was very rare in Western CEUs. Although strictly matching otherwise, mtDNA and pDNA haplogroups were not congruent in the northern part of SM (population from Imouzzer Ida Ou Tanane area, in Morocco). In this area, most samples (8/13) belong to the western haplogroup for mtDNA, but to the central-eastern one for pDNA, the others have matching haplogroups belonging to the central-eastern one (4/13) or the western one (1/13).

The strong West-East differentiation observed for the 350 genotypes dataset (Fig. S2 in supplementary material) and for MH and PH datasets (Fig. S4) was confirmed in the analysis conducted using GWN data. The SVDquartets reconstruction of 171 samples of *C. siliqua* and four samples of *C. oreothauma* as an outgroup, confirmed the monophyly of *C. siliqua* and resolved a strongly supported first split of the SM clade from the remaining *C. siliqua* lineages (Fig. 2A). A subsequent split separated the remaining Western CEUs (SWS, SES and NM) from Central and Eastern CEUs (CM, NEM, SEM). This phylogenetic topology based on GWN data is congruent with MH and PH data in supporting southwest Morroco as the ancestral area in the carob evolutionary history.

SVDquartets phylogenetic topology is highly congruent with the K=4 genetic structure obtained with SSR data (Fig. 2A). The four genetic groups match the following clades: SM, SWS + SES, NEM + SEM, CM. The relative position of NM in the SVDquartets tree, although sister to the other three western CEUs, is congruent with its mixed assignation to different genetic clusters for SSR and RADseq data (Fig. 2A,D).

The genetic clustering estimated with snmf based on RADseq data was repeated from K=2 to K=7 ancestral populations. The cross-entropy criterion suggested two optimal solutions for 5 and 7 groups (Fig. S5B in

supplementary material), of which K=7 had the highest probability. The five groups estimated with snmf (Fig. S5A) are congruent with SSR data and the SVDquartets reconstruction for the Western CEUs: three genetic groups have a good match with SM, SWS + SES and NEM + SEM, respectively, whereas CM have ancestry mainly from the two remaining genetic groups and NM shows admixture from the five genetic groups. The most likely genetic clustering, K=7 solution (Fig. 2A,D), resolved the differentiation of four ancestral populations roughly corresponding to south Morocco (SM), south Spain (SWS and SES) and the eastern CEUs (NEM and SEM). However, departures from a good match between microsatellite-based CEUs and RADseq genetic groups are substantial too. NEM exhibits admixture from SEM. CM includes individuals with admixture from three ancestral groups, two of them (in orange and yellow) almost restricted to this CEU and the last one, coloured in turquoise, present in all CEUs although with higher proprotions in NM, SWS and CM. NM was resolved as admixed with the contribution of all ancestral populations. The PCA results are consistent with the main genetic groups found with snmf, showing a clear delimitation of SM, SWS + SES and NEM+SEM groups, but partial overlap between the NM and CM groups (Figure 2B, C). RADseq genetic differentiation among the seven CEUs (Tab. S2. in supplementary material) was moderate with an overall Gst of 11%; all values above the mean involved a west-east differentiation, the highest differentiation being SES between and SEM (Gst = 19%).

Impact of dispersals on the genetic diversity and structure of

Ceratonia siliqua

For the Treemix analysis, we rooted *C. siliqua* phylogenetic tree with SM following SVDquartets reconstructions (Fig. 2). Tree topology reconstructed by Treemix (Fig. 3) is highly similar to that of SVDquartets, showing the same strong west-east differentiation. The position of NM was resolved as intermediate between Western (SM, SWS, SES) and Central and Eastern CEUs (CM, NEM, SEM), which agrees with the highly admixed pattern in NM inferred by snmf (Fig. 2 A) and with its low to moderate differentiation with respect to other CEUs (Tab. S2 in supplementary material). The Treemix model without migrations explained 96 % of the covariance (Fig. 3A), and the addition of four migrations resulted in 99.8 % of the total covariance explained (Fig. 3B). Three of the four migrations identified SEM as the source of introgression, into SM, SWS and NM, whereas the fourth migration connected the central-eastern CEUs to SES, again indicating a westward migration.

Impact of carob cultivation effort on carob genetic diversity

Overall genetic diversity values are reported in Table 1. The Nei diversity (Hexp) values decrease from west CEUs and natural or seminatural habitats to east and/or cultivated habitats, regardless of the markers. Interestingly, this trend is negatively correlated with the association index (rbardD), which reports the highest lack of mixing in cultivated CM carobs. The fixation index (f) suggests a lack of heterozygosity in SM, NM, and NEM (and SEM only for SSR), significant excess in CM, but was non significantly different from zero in SWS and SES (and SEM too for RADseq. Based on the differentiation between genetic groups formed by the combination of CEUs and habitats (Table 1), OUTFLANK did not detect candidate loci with a false discovery rate (qvalue) below 0.05, which indicates that the hypothesis of neutrality cannot been rejected (Fig. S6 in supplementary material).

Discussion

Intense research on plant phylogenetics and phylogeography over the last two decades have allowed the discovery of several major biogeographical trends in the Mediterranean basin (Garcia-Verdugo et al., 2021) and renewed our understandings of plant domestication (Purugganan, 2019). Following an initial focus on biogeographic refugia, recent studies have revealed the genetic imprints of past expansions and migration processes, some involving the entire Mediterranean basin (see reviews in Medail and Diadema 2009; Nieto Feliner 2011; Nieto Feliner, 2014; Migliore et al., 2018; Vargas et al., 2018; Thompson, 2020; Garcia-Verdugo et al., 2021). Our study provides a better understanding of the phylogeography of Mediterranean plants by revealing a new historical scenario: the main gene pools of carob (i.e. CEUS) originated from a biogeographic refugium probably located in southwest Morocco. Our results also highlight that carob domestication has

mainly relied on the use of locally selected and disseminated varieties, albeit punctuated by long-distance westward dispersal events by humans, which match major cultural waves by Romans and Arabs.

Evolutionary history of the carob tree

Our phylogeographic investigation allows rejecting a long-standing hypothesis that proposes an introduced origin of the carob tree in most of the Mediterranean. An Eastern Mediterranean or Southern Arabian origin followed by a human-mediated expansion were proposed by several authors partly based on linguistic evidence from vernacular names,*Ceratonia siliqua* and *C. oreothauma* occurrences in western Asia and carob agricultural practices (reviewed in Ramon-Laca & Mabberley, 2004). However, genetic data from SSR and plastid markers based on a thorough population sampling across the Mediterranean (Viruel et al., 2020) found a better explanation to account for all the data according to which current *C. siliqua* populations originated from two disjunct refugia after the Last Interglacial ca. 116 Ka ago. SSR data revealed introgression in the Central Mediterranean and Northern Morocco, but the strong west vs. central-east pattern based on plastid data revealed a low human influence on the main current patterns of genetic diversity and structure of the carob tree across the Mediterranean. The comprehensive review of carob fossil data done by Viruel et al., (2020) did not provide support for an eastern origin of*C. siliqua* either. Instead, the fossil record shows a mostly continuous presence of *Ceratonia* around the palaeo-Mediterranean Sea since the Oligocene with a progressive decline starting c. 20 Ma.

Compared to SSR data, RADseq allows bridging phylogenetic and population genetic inferences (Parchman et al., 2018). Here, the inclusion of*Ceratonia oreothauma* , the sister species of *C. siliqua* , despite their divergence around 6.4 Ma (Viruel et al., 2020), corroborates our previous conclusion on the importance of western Mediterranean in the history of the carob. Our SVDquartets phylogenetic reconstruction provides further resolution pointing to southwest Morocco as closest to the ancestral population of *C. siliqua* (Figure 2). This origin was suggested based on a slightly higher genetic diversity revealed for both nuclear (SSR) and plastid data in Viruel et al. (2020). However, in our previous study, coalescent-based models tested by an approximate Bayesian computation approach supported a two refugia hypothesis to explain the west-east split in the genetic diversity structure of the carob tree. Here, our phylogenetic and population genomic inferences support a different scenario. According to SVDquartets phylogenetic reconstruction based on nuclear genome-wide diversity, the carob tree followed two routes of migration from south Morocco; one northward that reached western north Africa and south Spain (NM, SES and SWS CEUs) and another towards the east that gave rise to the central-eastern CEUs. Both mitochondrial and plastid data extracted from RADseq data also support the existence of the ancestral pool in southwest Morocco. Specifically, eastern mtDNA and pDNA haplotypes are present in the northern part of SM (Imouzzer Ida Ou Tanane area) thus suggesting that this is the most credible source for the eastern populations. By contrast to our previous study, our new scenario explains the west-east split in the carob genetic diversity by two migration routes from an ancestral population situated in south Morocco.

As shown in our previous study, species distribution modelling (SDM) indicates that both the Last Interglacial and the Last Glacial Maximum were periods of contraction for this species during the Pleistocene (Viruel et al., 2020). Moreover, SDM predicted that some areas in the North African and South European Atlantic coasts could have been continuously suitable since the last 130 ka. Southwest Morocco has been identified as a biogeographic refugium and even as a diversification cradle for several taxonomic groups (e.g. Medail and Quezel, 1999; Medail et al. 2001; Ortiz et al., 2009; Martinez-Freiria et al., 2017; Bobo-Pinilla et al., 2018; Villa-Machio et al., 2018; Klesser et al., 2021). Although Mediterranean phylogeographic studies focused mostly on glacial refugia, three recent studies have highlighted South and West Morocco as a refugium for plant populations during the LIG (Villa-Machio et al., 2018: Bobo-Pinilla et al., 2018: Viruel et al., 2020) where an overall pronounced climate continentality could have been buffered by the ocean vicinity.

Footprints of domestication on the current genetic structure of the carob tree across the Mediterranean

Although disentangling the history of cultivated plants is complex, our phylogeographic investigation in the carob tree sheds light on the history of agriculture. Our previous study based on SSR data (Viruel et al.,

2020) suggested that local domestication events from wild populations were the most likely scenario. The RADseq data here presented, depicting a strong east-west genome-wide differentiation could not explain domestication solely based on translocations and/or human-based dispersals from east to west. Agriculture practices in the carob tree are based on propagation by grafting (Zohary, 2002) although seeds could have also been transported. In either case, domestication based only on westward propagations of cultivars from the east would have maintained the maternal (eastern) haplotypes in the Western Mediterranean. Instead, our results conclude that the dispersal of selected varieties (vegetatively propagated), between remote geographical areas, was not the main force of domestication in carob tree.

The use of genomic data at the infraspecific level has permitted identifying footprints of domestication in crop models where PCR-based molecular markers had previously failed. In the case of date palm, genomic data revealed that human-mediated dispersal imprints were superimposed on a previous phylogeographical structure (Gros-Balthazar et al. 2017; Flowers et al. 2019). In the carob tree, despite a moderate differentiation (Gst = 11%), genome-wide diversity is structured into three main genetic sources: SM, SWS+ SES and NEM + SEM (Figure 2, and Fig. S5 in supplementary material). Although this pattern does not suggest translocation of eastern domesticated varieties into South Morocco or the Iberian Peninsula, it does fit with the patterns found in geographically intermediate groups (NM, CM and NEM). These are less differentiated, which is explained by high rates of admixture (Fig. 2). To untangle the role of human-based dispersals in these strong genetic admixtures, we used allelic-frequency based models aiming at estimating the intensity and origin of dispersal events throughout the evolutionary tree of the carob tree (Fig. 3): results of Treemix recover westwards migrations that were mostly originated from SEM, or from central-eastern CEUs (SEM, NEM and CM). These translocations match with the beginning of carob agriculture in the East, its dispersal by Greeks, Romans and after by Arabs in historical times (Ramon Laca and Mabberley, 2001; Viruel et al., 2020). They may have contributed to genetic admixed pool used locally for cultivation as observed in North Morocco (NM).

The second footprint of domestication was observed in CM, which is the area among those considered in our study in which cultivated varieties (either local or imported selections) are most diffused (Di Guardo et al., 2019). This CEU is characterized by a slightly lower genetic diversity and a small excess of heterozygosity whereas all other CEUs showed a deficit. We detected a genetic group of individuals without admixture in CM, corresponding to the monumental carobs of the Ragusa district (Sicily, South Italy). Without being clones, these individuals, harvested without interruption for centuries, are genetically very close to each other and form a lineage within CM. The genetic patterns of these ancient CM individuals have not been observed in other CEUs, supporting again the idea that diffusion of selected genotypes at the local scale local, rather than long-distance dispersal, played a major role in the domestication of carob. Despite this pattern, we did not detect any candidate loci under selection due to domestication pressures, which could be explained by the limitations of our method and sampling or by a low effect of domestication on the carob genome. Compared to other perennial crop species for which candidate and adaptive loci have been found by whole genome sequencing as well as RADseq (Cornejo et al., 2018; Alves-Pereira et al., 2020; Groppi et al., 2021), a relatively lower impact of selection is likely in carob. Domestication leading to fine-tuning of gene expression patterns rather than genome-wide evolution, as observed in olive (Gros-Balthazard et al., 2019), maybe almost undetectable by a reduced-representation genomics approach such as RADseq.

Conservation of genetic diversity within Carob Evolutionary Units (CEUs)

Knowing the structure of genome-wide diversity is essential for preserving the genetic resources of cultivated species and for future breeding (Purugganan 2019). We used an integrative approach combining geographic and genetic differentiation to characterize evolutionary units for *Ceratonia siliqua* across the Mediterranean. In a survey including 1020 samples, seven non-overlapping CEUs were identified as the best solution to minimize intra-group variance and obtain homogenous groups non overlapped geographically. Four genetic clusters, identified within carob tree populations based on a thorough sampling across the Mediterranean using nuclear SSR and SNP data, are contained within the seven CEUs (Figure 2): South Morocco (SM), Iberian Peninsula (SES, SWS), Central Mediterranean (CM, NM) and Eastern Mediterranean (NEM, SEM).

These four genetic clusters exhibit moderate introgression in the West and East CEUs, but high patterns of admixture in the Central Mediterranean (CM, NM), more intense in NM. RADseq data further resolved these genetic structuring across the Mediterranean by identifying seven genetic clusters (Fig. 2 A,D), which, in some cases fully matched with a CEU (e.g. SM, SES) or two CEUs (SWS, SES) whereas, in other cases, a mixture of more than one genetic cluster was observed (e.g. CM). These data permit a better interpretation of the genetic diversity patterns between CEUs and are thus important for future designs of ex situ conservation. Our results suggest that moderate genetic diversity is uniformly distributed across CEUs (Table 1). Only a slightly higher genetic diversity was estimated in Western CEUs (SWS, SES, SM) based on SSR loci. Although Central CEUs (CM, NM) are highly admixed, these factors did not entail an increase in genetic diversity compared to non-admixed clusters. Conservation of genetic resources for the carob tree should recover genetic diversity found across the Mediterranean by preserving materials from western and eastern CEUs prioritizing the most differentiated CEUs SM, SES + SWS, and SEM. CM, which contains three genetic groups and for which carob cultivars have been well characterized, specifically in Italy, should benefit from more investigations on carob evolution under domestication.

Aknowledgments

This study is part of the DYNAMIC project supported by the French national agency of research (ANR-14-CE02-0016) and benefited from equipment and services from the molecular biology facility (SCBM) at IMBE (Marseille, France). All bioinformatics and simulations were done on the High-Performance Computing Cluster from the Pytheas informatic facility (OSU Institut Pytheas Aix Marseille Univ, INSU-CNRS UMS 3470) J.V. benefited from a Postdoc Fellowship funded by DYNAMIC and a Marie Skłodowska-Curie Individual Fellowship (704464 - YAMNOMICS - MSCA-IF-EF-ST). The authors thank for their help to complete our sampling: Annette Patzelt (Oman Botanic Garden), Minas Papadopulos (Department of Forests of the Republic of Cyprus), Zahra Djabeur (Oran University), Nabil Benghanem (Tizi-Ouzou University), Gianluigi Bacchetta (Cagliari University), Sonja Yakovlev (Paris-Sud University), Errol Vela (CIRAD), Maria Panitsa (Patras University), and the services of Junta de Andalucia.

Author contributions

H. S., A.B., F.M., S.L.M., M.B.K., L.O., G.N.F. and J.V. conceived, planed the study and collected samples. F.L.M. performed the DNA extraction and quality assessment. J.V, A.B. and V.F. performed curation and analysis of microsatellite data. A.B. performed RADseq data curation and SNPs filtering. A.B. and J.V provided the analysis, tables and figures. A.B., J.V., G.N.F. and F.M. interpreted the results. A.B. and J.V. drafted the manuscript. J.V., G.N.F., S.L.M. and M.D.G. edited the manuscript. J.V., G.N.F. and A.B. wrote the final manuscript. H.S. was in charge with funding acquisition and project administration. All authors read and approved the final version.

Data Availability Stament

Full information on populations sampling and microsatellite data are available in Viruel et al. (2020) and deposited in DRYAD (*https://doi.org/10.5061/dryad.k7m020r*). Raw RADseq reads are deposited at NCBI under Bioproject accession (######). Assemblies from ipyRAD, data files and R scripts of analyses are available at Zenodo (##########).

References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome research, 19(9), 1655-1664. https://doi.org/10.1101/gr.094052.109

Alves-Pereira, A., Clement, C. R., Picanço-Rodrigues, D., Veasey, E. A., Dequigiovanni, G., Ramos, S. L. F., Baldin Pinheiro, J., Pereira de Souza, A. & Zucchi, M. I. (2020). A population genomics appraisal suggests independent dispersals for bitter and sweet manioc in Brazilian Amazonia. Evolutionary applications, 13, 342-361.

Baumel, A., Mirleau, P., Viruel, J., Bou Dagher Kharrat, M., La Malfa, S., Ouahmane, L., ... & Medail,

F. (2018). Assessment of plant species diversity associated with the carob tree (Ceratonia siliqua, Fabaceae) at the Mediterranean scale. Plant Ecology and Evolution, 151, 185–193. https ://doi.org/10.5091/plecevo.2018.1423

Borrell, J.S., Wang, N., Nichols, R.A., & Buggs, R.J. (2018). Genetic diversity maintained among fragmented populations of a tree undergoing range contraction. Heredity, 121: 304-318

Besnard, G., Terral, J. F., & Cornille, A. (2018). On the origins and domestication of the olive: a review and perspectives. Annals of botany, 121(3), 385-403. https://doi.org/10.1093/aob/mcx145

Bobo-Pinilla, J., Penas de Giles, J., Lopez-Gonzalez, N., Mediavilla, S., & Martinez-Ortega, M. M. (2018). Phylogeography of an endangered disjunct herb: long-distance dispersal, refugia and colonization routes. AoB Plants, 10, ply047

Boivin, N. L., Zeder, M. A., Fuller, D. Q., Crowther, A., Larson, G., Erlandson, J. M., Denham, T. , & Petraglia, M. D. (2016). Ecological consequences of human niche construction: Examining long-term anthropogenic shaping of global species distributions. Proceedings of the National Academy of Sciences, 113(23), 6388-6396.

Brassesco, M. E., Brandao, T. R., Silva, C. L., & Pintado, M. (2021). Carob bean (Ceratonia siliqua L.): A new perspective for functional food. Trends in Food Science & Technology.

Caruso, M., La Malfa, S., Pavliček, T., Frutos Tomñs, D., Gentile, A., & Tribulato, E. (2008). Characterisation and assessment of genetic diversity in cultivated and wild carob (Ceratonia siliqua L.) genotypes using AFLP markers. The Journal of Horticultural Science and Biotechnology, 83(2), 177-182.

Chavent, M., Kuentz-Simonet, V., Labenne, A., & Saracco, J. (2018). ClustGeo: an R package for hierarchical clustering with spatial constraints. Computational Statistics, 33(4), 1799-1822.

Chen, C., Qi, Z.C., Xu, X.H., Comes, H.P., Koch, M.A., Jin, X.J., Fu, C.X. & Qiu, Y.X. (2014). Understanding the formation of Mediterranean–African–Asian disjunctions: evidence for Miocene climate-driven vicariance and recent long-distance dispersal in the Tertiary relict Smilax aspera (Smilacaceae). New Phytologist, 204, 243–255.

Chifman, J., & Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. Bioinformatics, 30(23), 3317-3324.

Cornejo, O. E., Yee, M. C., Dominguez, V., Andrews, M., Sockell, A., Strandberg, E., Livingstone D. III, Stack C., Romero A., Umaharan P., Royaert S., Tawari N.R., Ng P., Gutierrez O., Phillips W., Mockaitis K., Bustamante C.D.& Motamayor, J. C. (2018). Population genomic analyses of the chocolate tree, Theobroma cacao L., provide insights into its domestication process. Commun Biol 1, 167 (2018). https://doi.org/10.1038/s42003-018-0168-6

de Medeiros, B.A.S., & Farrell, B.D. (2018). Whole-genome amplification in double-digest RADseq results in adequate libraries but fewer sequenced loci. PeerJ 6:e5089*https://doi.org/10.7717/peerj.5089*

de Medeiros, B.A.S., (2019). Matrix Condenser v.1.0. Available at:*https://github.com/brunoasm/matrix_condenser/*

Désamoré, A., Laenen, B., Devos, N., Popp, M., González-Mancebo, J. M., Carine, M. A., & Vanderpoorten, A. (2011). Out of Africa: north-westwards Pleistocene expansions of the heather Erica arborea. Journal of Biogeography, 38(1), 164-176.

Di Guardo, M., Scollo, F., Ninot, A., Rovira, M., Hermoso, J. F., Distefano, G., La Malfa S. & Batlle, I. (2019). Genetic structure analysis and selection of a core collection for carob tree germplasm conservation and management. Tree Genetics & Genomes, 15(3), 1-14.

Eaton, D. A., & Overcast, I. (2020). ipyrad: Interactive assembly and analysis of RADseq datasets. Bioinformatics, 36(8), 2592-2594.

Ewels, P., Magnusson, M., Lundin, S., & Kaller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047-3048.

Flowers, J.M., Hazzouri, K.M., Gros-Balthazard, M., Mo, Z., Koutrumpa, K., Perrakis, A., Ferrand, S., Khierralah, H.S.M., Fuller, D.Q., Aberlenc, F., Fournaraki, C., Purugganan, M.D., 2019. Cross-species hybridization and the origin of North African date palms. Proc. Natl. Acad. Sci. U. S. A. 116, 1651–1658. https://doi.org/10.1073/pnas.1817453116

Frichot, E., & Francois, O. (2015). LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution, 6(8), 925-929.

Fuller, D. Q., Willcox, G., & Allaby, R. G. (2011). Cultivation and domestication had multiple origins: arguments against the core area hypothesis for the origins of agriculture in the Near East. World Archaeology, 43(4), 628-652.

Garcia-Verdugo, C., Mairal, M., Tamaki, I., & Msanda, F. (2021). Phylogeography at the crossroad: Pleistocene range expansion throughout the Mediterranean and back-colonization from the Canary Islands in the legume Bituminaria bituminosa. Journal of Biogeography ....

Gosselin, T. (2020). radiator: RADseq Data Exploration, Manipulation and Visualization using R. R package version 1.1.9 https://thierrygosselin.github.io/radiator/. doi : 10.5281/zenodo.3687060

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes. 5: 184-186

Groppi, A, Liu, S, Cornille, A, Decroocq, S, Bui, Q T, Tricon, D, & Decroocq, V (2021) Population genomics of apricots unravels domestication history and adaptive events. Nature communications, 12(1), 1-16. https://doi.org/10.1038/s41467-021-24283-6

Gros-Balthazard, M., Galimberti, M., Kousathanas, A., Newton, C., Ivorra, S., Paradis, L., Vigouroux, Y., Carter, R., Tengberg, M., Battesti, V., Santoni, S., Falquet, L., Pintaud, J.-C.C., Terral, J.-F.F., Wegmann, D., 2017. The Discovery of Wild Date Palms in Oman Reveals a Complex Domestication History Involving Centers in the Middle East and Africa. Curr. Biol. 27, 2211–2218. https://doi.org/10.1016/j.cub.2017.06.045

Gros-Balthazard, M., Besnard, G., Sarah, G., Holtz, Y., Leclercq, J., Santoni, S., Wegmann D., Glemin S., Khadari, B. (2019). Evolutionary transcriptomics reveals the origins of olives and the genomic changes associated with their domestication. The Plant Journal, 100(1), 143-157. https://doi.org/10.1111/tpj.14435

Gruber, B., Unmack, P. J., Berry, O. F., & Georges, A. (2018). dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. Molecular Ecology Resources, 18(3), 691-699.

Hodel, R.G.J., Chen S, Payton A.C., McDaniel S.F., Soltis P., & Soltis D.E. (2017). Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: comparing microsatellites and RAD-Seq and investigating loci filtering. Sci Rep. 7:17598. https://doi.org/10.1038/s41598-017-16810-7.

Hipp, A. L., Manos, P. S., Hahn, M., Avishai, M., Bodenes, C., Cavender-Bares, J., ... & Valencia-Avalos, S. (2020). Genomic landscape of the global oak phylogeny. New Phytologist, 226, 1198-1212.

Jombart T, Devillard S and Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics11:94. doi:10.1186/1471-2156-11-94

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics, 27(21), 3070-3071

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

Klesser, R., Husemann, M., Schmitt, T., Sousa, P., Moussi, A., & Habel, J. C. (2021). Molecular biogeography of the Mediterranean Buthus species complex (Scorpiones: Buthidae) at its southern Palaearctic margin. Biological Journal of the Linnean Society, 133, 166-178.

La Malfa, S., Curro, S., Douglas, A. B., Brugaletta, M., Caruso, M., & Gentile, A. (2014). Genetic diversity revealed by EST-SSR markers in carob tree (Ceratonia siliqua L.). Biochemical Systematics and Ecology, 55, 205-211.

Martinez-Freiria, F., Crochet, P. A., Fahd, S., Geniez, P., Brito, J. C., & Velo-Anton, G. (2017). Integrative phylogeographical and ecological analysis reveals multiple Pleistocene refugia for Mediterranean Daboia vipers in north-west Africa. Biological Journal of the Linnean Society, 122, 366-384.

Medail, F., Quezel, P., Besnard, G., & Khadari, B. (2001). Systematics, ecology and phylogeographic significance of Olea europaea L. ssp. maroccana (Greuter & Burdet) P. Vargas et al., a relictual olive tree in south-west Morocco. Botanical Journal of the Linnean Society, 137(3), 249-266.

Medail, F., & Diadema, K. (2009). Glacial refugia influence plant diversity patterns in the Mediterranean Basin. Journal of biogeography, 36(7), 1333-1345

Meyer R.S., Duval A.E., & Jensen H.R. (2012). Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. New Phytologist, 196, 29–48.

Migliore J., Baumel A., Leriche A., Juin M., & Medail F. (2018). Surviving glaciations in the Mediterranean region: an alternative to the long-term refugia hypothesis. Botanical Journal of the Linnean Society, 187, 537–549.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol., 37:1530-1534.*https://doi.org/10.1093/molbev/msaa015*

Nieto Feliner, G. (2014). Patterns and processes in plant phylogeography in the Mediterranean Basin. A review. Perspectives in Plant Ecology, Evolution and Systematics, 16, 265–278.

Nieto Feliner, G. (2011). Southern European glacial refugia: a tale of tales. Taxon, 60, 365–372.

Ortiz MA, Tremetsberger K, Stuessy T, Terrab A, Garcia-Castano JL, Talavera S. 2009. Phylogeographic patterns in Hypochaeris sect. Hypochaeris (Asteraceae, Lactuceae) of the western Mediterranean. Journal of Biogeography 36:1384–1397

Ortiz, E.M. 2019. vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. DOI:10.5281/zenodo.2540861

Parchman, T. L., Jahner, J. P., Uckele, K. A., Galland, L. M., & Eckert, A. J. (2018). RADseq approaches and applications for forest tree genetics. Tree Genetics & Genomes, 14(3), 1-25.

Pickrell, J., & Pritchard, J. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. Nature Precedings, 1-1.

Purugganan, M. D. (2019). Evolutionary insights into the nature of plant domestication. Current Biology, 29(14), R705-R714

Quezel, P., & Medail, F. (2003). Ecologie et biogeographie des forets du bassin Mediterraneen. Paris, France: Elsevier Editions.

Rambaut, A., & Drummond, A. J. (2012). FigTree version 1.4. 0.

Ramon-Laca, L. & Mabberley, D.J. (2004). The ecological status of the carob-tree (Ceratonia siliqua, Leguminosae) in the Mediterranean. Botanical Journal of the Linnean Society, 144, 431–436.

Silliman, K. (2019). Population structure, genetic connectivity, and adaptation in the Olympia oyster (Ostrea lurida) along the west coast of North America. Evolutionary applications, 12(5), 923-939.

Swofford, D. L. (2018). PAUP*(* Phylogenetic Analysis Using PAUP). Version 4a161.

Thompson, J. D. (2020). Plant Evolution in the Mediterranean: Insights for Conservation. Oxford University Press, USA.

Vargas, P., Fernandez-Mazuecos, M., & Heleno, R. (2018). Phylogenetic evidence for a Miocene origin of Mediterranean lineages: species diversity, reproductive traits and geographical isolation. Plant Biology, 20, 157-165.

Villa-Machio, I., Fernandez de Castro, A. G., Fuertes-Aguilar, J., & Nieto Feliner, G. (2018). Out of North Africa by different routes: phylogeography and species distribution model of the western Mediterranean *Lavatera maritima* (Malvaceae). *Botanical Journal of the Linnean Society* , *187* , 441-455.

Viruel J, Le Galliot N, Pironon S, Nieto Feliner G, Suc JP, Lakhal-Mirleau F, Juin M, Selva M, Bou Dagher Kharrat M, Ouahmane L, Malfa S, Diadema K, Sanguin H, Medail F, Baumel A (2020) A strong east–west Mediterranean divergence supports a new phylogeographic history of the carob tree (Ceratonia siliqua, Leguminosae) and multiple domestications from native populations. Journal of Biogeography 47, 460-471

Viruel J, Haguenauer A, Juin M, Mirleau F, Bouteiller D, Boudagher-Kharrat M, Ouahmane L, La Malfa S, Medail F, Sanguin H, Nieto Feliner, G, & Baumel A (2018). Advances in genotyping microsatellite markers through sequencing and consequences of scoring methods for Ceratonia siliqua (Leguminosae). Applications in Plant Sciences, 6, e01201.

Warschefsky EJ, von Wettberg EJ (2019). Population genomic analysis of mango (Mangifera indica) suggests a complex history of domestication. New Phytologist, 222, 2023-2037. https://doi.org/10.1111/nph.15731

Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F ST. The American Naturalist, 186(S1), S24-S36.

Winter, D. J. (2012). MMOD: an R library for the calculation of population differentiation statistics. Molecular ecology resources, 12(6), 1158-1160.

Zeder, M. A. (2008). Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proceedings of the national Academy of Sciences, 105(33), 11597-11604.

Zohary D. (2002). Domestication of the carob (Ceratonia siliqua L.). Israel Journal of Plant Sciences, 50, 141-15.

Zohary, D., & Hopf, M. (2012). Domestication of plants in the Old World: The origin and spread of cultivated plants in West Asia, Europe and the Nile Valley. Oxford, UK: Oxford University Press.

Supporting information

Tab. S1: Statistics from four assemblies conducted on RADseq data (36 samples) elaborated with ipyrad with varying the clustering threshold from 0.9 to 0.96 % of similarity.

Tab. S2: pairwise Gst differentiations among CEUs based on RADseq data. Values above the overall Gst (11%) in bold.

Fig. S1 : Design of Carob Evolutionary Units (CEUs) using ClustGeo, which considers euclidean genetic and geographical distances. 17 SSRs and 15 SNP markers from microsatellite loci were used. The Ward dendrogram of 56 carob populations with a partition in K=7 clusters (A) was obtained with a normalized proportion α of explained inertia of 0.2 for the geographic distance and 0.8 for the genetic distance (B). C) Neighbor Joining tree based on pairwise genetic differentiation (Gst SSR markers) among the seven clusters. See main text for acronyms of CEUs.

Fig. S2 : Genome wide diversity structure of 350 carob trees based on 10,012 RADseq loci with an overall missing data rate of 64%. (A) PCA scatter plot of 350 carob RADseq genotypes. (B) Neighbor joining tree of pairwise Gst differentiations among seven Carob Evolutionary Units (CEUs).

Fig. S3: FST per loci distribution (1 SNP by locus) with 56 loci identified as outliers by OUTFLANK due to their unexpectedly high Fst differentiation (FDR <0.05). The blue line is the inferred neutral distribution.

Fig. S4: Map of mtDNA and pDNA haplogroups from 14 and 21 RADseq loci respectively obtained for 190 carob trees. West and East haplogroups match strictly for both organelle data sets except for South Morocco (SM).

Fig. S5 : Population genetic structure of the carob according to RADseq. A) Genetic admixture plots for 190 carob trees from k=2 to K=7 ancestral populations obtained with the snmf method (LEA package) performed on 3,557 unlinked SNPs. The West and East lineages refer to organellar haplogroups (Fig. S4). B) Cross-entropy criterion suggesting two optimal solutions with K= 5 or 7.

Fig. S6: FST per loci distribution (1 SNP by locus). OUTFLANK method did not detected any outlier (FDR <0.05). The blue line is the inferred neutral distribution.

Figure legends

FIGURE 1: Bioinformatic pipeline to extract and filter SNPs from RADseq data for the carob tree.

FIGURE 2 : Population genetic structure of the carob tree. A) SVDquartets tree of seven genetically and geographically homogeneous groups (CEUs) based on RADseq markers. Genetic admixture plots are based on four ancestral populations for SSR markers (1020 genotypes, 17 loci) and on 7 ancestral populations for RADseq markers (190 genotypes, 3557 neutral and unlinked SNPs). B) & C) PCA scatterplots of RADseq genotypes -the first three components = 15.2% of variance). D) Map of genetic admixture based on RADseq markers and 7 ancestral populations.

FIGURE 3: Evolutionary history of the carob tree reconstructed with Treemix. Maximum likelihood trees obtained without (A) and with gene flow (B) events explaining 96% and 99% of the variance, respectively. The color of the arrows indicates the migration weight which is the fraction of ancestry derived from the migration edge.

TABLE 1: Estimates of genetic diversity based on microsatellites (17 SSR loci) and genome-wide markers (3557 SNPs) for seven *Ceratonia siliqua* units (CEUs). Within each CEU, samples were split into groups according to their origin (cultivated, seminatural or natural habitats).

	SSR					RAD				
CEUs	**N**	**Hobs**	**Hexp**	**f**	**RbarD**	**N**	**Hobs**	**Hexp**	**f**	**RbarD**
SM cultivated	57	0.49	0.54	0.09*	15*	4	0.21	0.20	-0.07*	26*
SM seminatural	51	0.50	0.58	0.12*	23*	11	0.22	0.24	0.07^{ns}	7*
SM natural	61	0.53	0.57	0.06*	15*	15	0.23	0.25	0.07*	7*
SM CEU	**169**	**0.51**	**0.58**	**0.11***	**16***	**30**	**0.22**	**0.25**	**0.11***	**5***
SWS cultivated	20	0.51	0.49	-0.06*	68*	7	0.22	0.21	-0.06*	27*
SWS seminatural	48	0.53	0.55	0.04^{ns}	8 ns	7	0.23	0.22	-0.02*	3*
SWS natural	36	0.55	0.52	0.04^{ns}	8 ns	7	0.21	0.22	0.01^{ns}	8*
SWS CEU	**104**	**0.53**	**0.54**	**0.02^{ns}**	**17***	**21**	**0.22**	**0.23**	**0.04^{ns}**	**5***
SES cultivated	61	0.51	0.53	0.04^{ns}	35*	10	0.23	0.22	-0.06*	14*
SES natural	40	0.53	0.54	0.02^{ns}	21*	8	0.22	0.21	-0.04*	3*
SES CEU	**101**	**0.52**	**0.54**	**0.04^{ns}**	**24***	**18**	**0.23**	**0.22**	**-0.01^{ns}**	**8***
NM cultivated	70	0.47	0.50	0.05^{ns}	26*	19	0.19	0.22	0.12*	200^{ns}
NM seminatural	101	0.47	0.51	0.06*	30*	4	0.22	0.19	-0.14*	82*
NM CEU	**171**	**0.47**	**0.51**	**0.06***	**25***	**23**	**0.19**	**0.22**	**0.11***	**167^{ns}**

	SSR					RAD				
CM cultivated	10	0.40	0.44	0.05*	28*	21	0.22	0.18	-0.06*	155*
CM seminatural	184	0.42	0.46	0.07*	15*	21	0.24	0.23	-0.03*	25*
CM CEU	**194**	**0.42**	**0.46**	**0.08***	**17***	**42**	**0.23**	**0.22**	**-0.02^{ns}**	**55***
NEM cultivated	42	0.46	0.47	0.01 ns	76*	14	0.21	0.22	0.07^{ns}	47^{ns}
NEM seminatural	131	0.47	0.51	0.05*	17*	20	0.22	0.23	0.03^{ns}	20*
NEM CEU	**173**	**0.45**	**0.50**	**0.80***	**20***	**34**	**0.21**	**0.23**	**0.07***	**33***
SEM seminatural	**106**	**0.45**	**0.51**	**0.10***	**22***	**22**	**0.20**	**0.22**	**0.07^{ns}**	**11***

N= number of genotypes analysed, Hobs= observed heterozygosity, Hexp= expected heterozygosity or Nei genetic diversity, f=Wright's inbreeding coefficient, rbarD = standardized form of the index of association accounting for multilocus linkage (x1000). * indicates f or rbarD siginificantly different from zero based upon 500 and 100 boostrap iterations respectively. SM = south Morocco, SWS= south west Spain, SES = south east Spain, NM = north Morocco, CM= central Mediterranean, NEM= north east Mediterranean, SEM= south east Mediterranean.

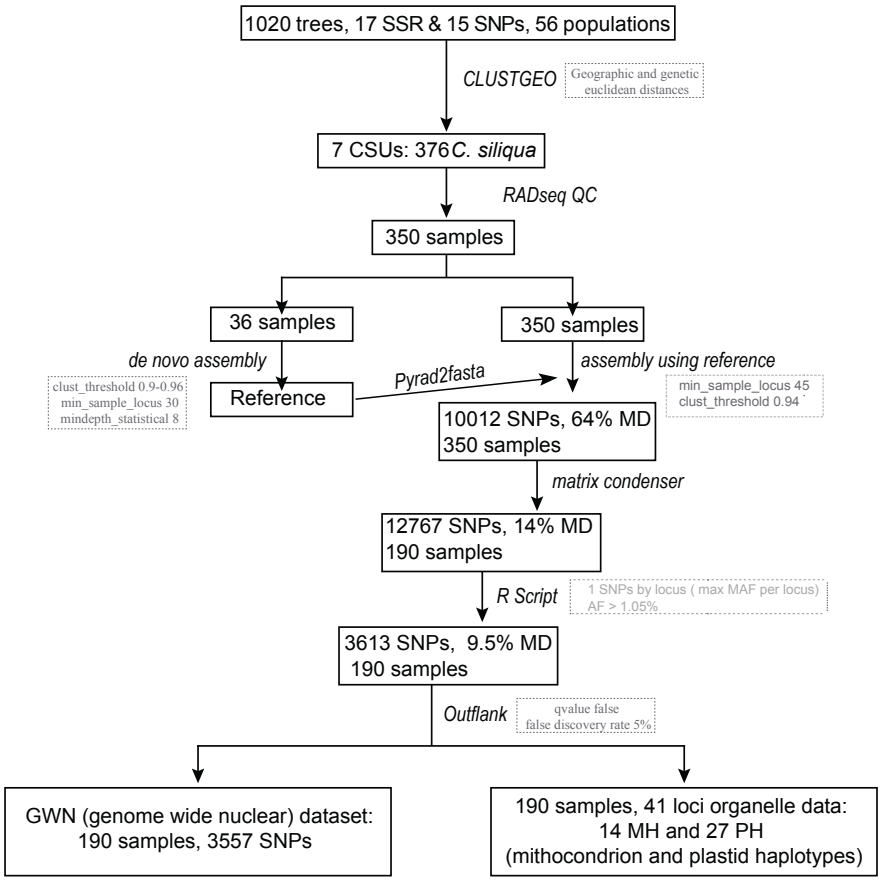


FIGURE 1: Bioinformatic pipeline to extract and filter SNPs from RADseq data for the carob tree.

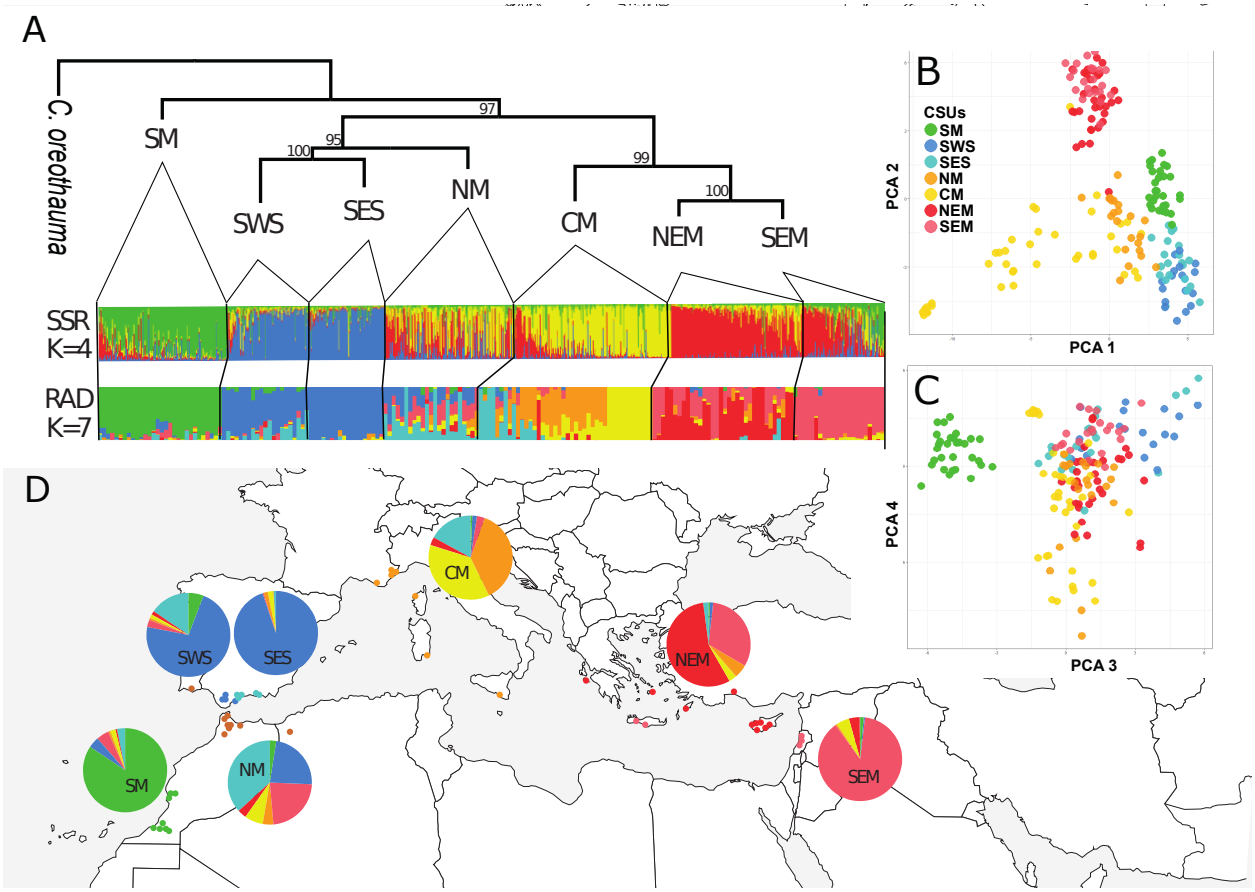


FIGURE 2 : Population genetic structure of the carob tree. A) SVDquartets tree of seven genetically and geographically homogeneous groups (CEUs) based on RADseq markers. Genetic admixture plots are based on four ancestral populations for SSR markers (1020 genotypes, 17 loci) and on 7 ancestral populations for RADseq markers (190 genotypes, 3557 neutral and unlinked SNPs). B) & C) PCA scatterplots of RADseq genotypes -the first three components = 15.2% of variance). D) Map of genetic admixture based on RADseq markers and 7 ancestral populations.

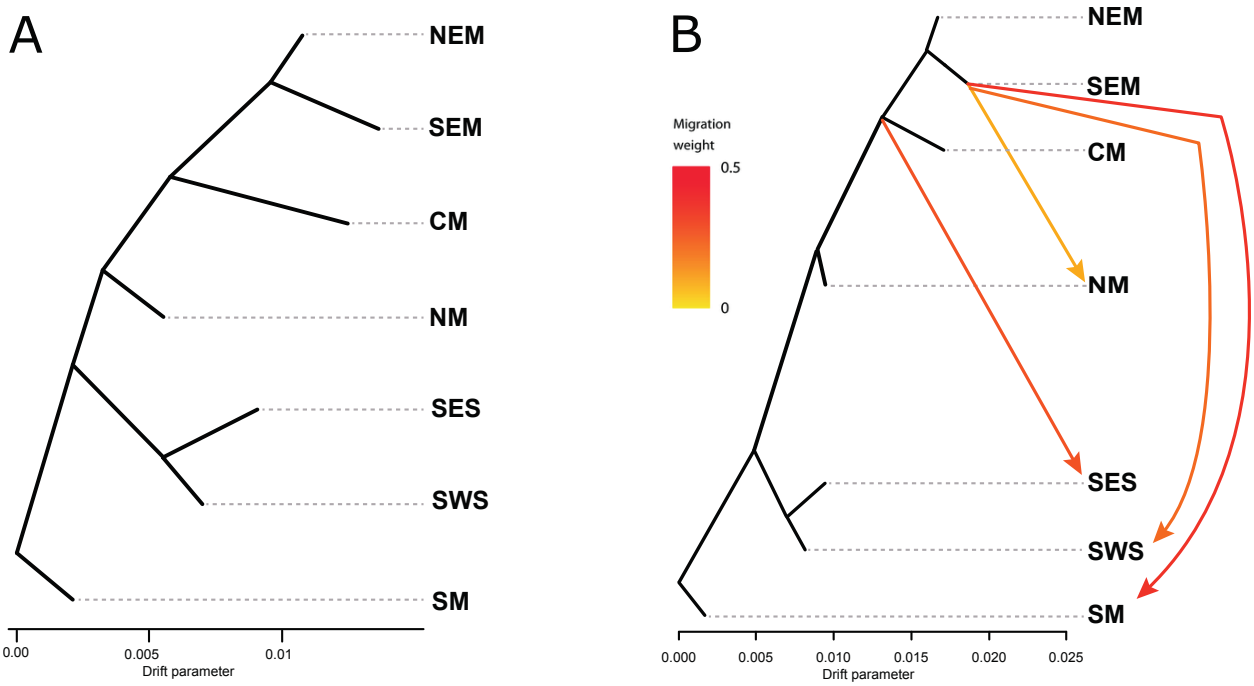
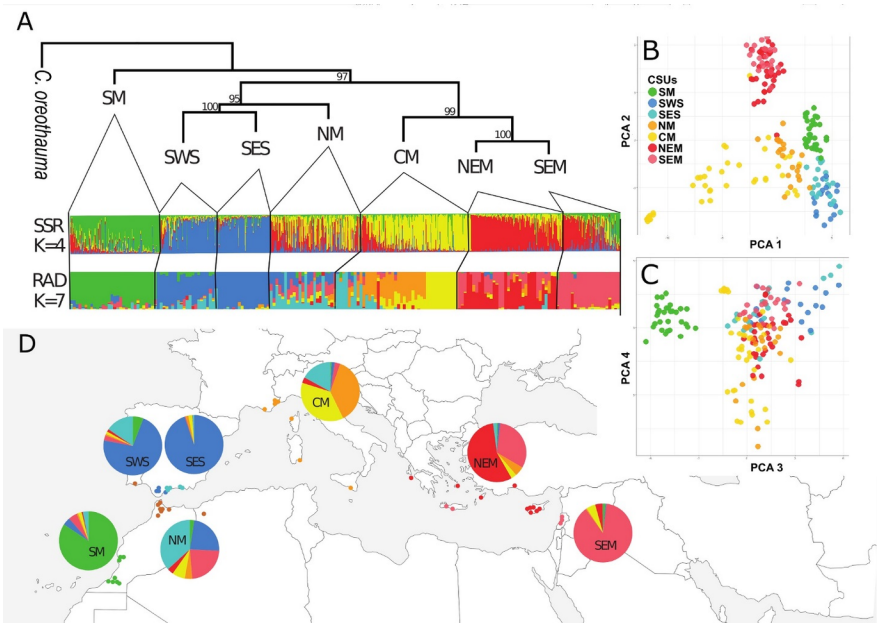


FIGURE 3: Evolutionary history of the carob tree reconstructed with Treemix. Maximum likelihood trees obtained without (A) and with gene flow (B) events explaining 96% and 99% of the variance, respectively. The color of the arrows indicates the migration weight which is the fraction of ancestry derived from the migration edge.

Hosted file

`image1.emf` available at https://authorea.com/users/434651/articles/540477-genome-wide-footprints-in-the-carob-tree-ceratonia-siliqua-unveil-a-new-domestication-pattern-of-fruit-trees-in-the-mediterranean



Hosted file

`image3.emf` available at https://authorea.com/users/434651/articles/540477-genome-wide-footprints-in-the-carob-tree-ceratonia-siliqua-unveil-a-new-domestication-pattern-of-fruit-trees-in-the-mediterranean