bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Interspecific gene flow and vicariance in Northern African rear-edge white oak populations revealed by joint analysis of microsatellites and flanking sequences.

Olivier Lepais[1*], Abdeldjalil Aissi[2], Errol Véla[3], Yassine Beghami[2].

[1]BIOGECO, INRAE, Univ. Bordeaux, 33610 Cestas, France.

[2]LAPAPEZA, Université Batna 1 Hadj Lakhdar, ISVSA, Batna, Algeria.

[3]AMAP, Université de Montpellier/CIRAD/CNRS/INRA/IRD, Montpellier, France.

[*]corresponding author: olivier.lepais@inrae.fr.

Abstract

Rear-edge populations represent reservoirs of potentially unique genetic diversity but are particularly vulnerable to global changes. While continental-scale phylogeographic studies usually do not cover these populations, more focused local scale study of rear-edge populations should help better understand both past evolutionary history and its consequences for the persistence and conservation of these potentially unique populations. We studied molecular variation at 36 sequenced nuclear microsatellites in 11 rear-edge *Quercus faginea* and *Q. canariensis* populations across Algeria to shed light on taxonomic relationship, population past evolutionary history and recent demographic trajectory. We used descriptive approach and simulation-based inference to assess the information content and complementarity of linked microsatellite and flanking sequence variations. Genetic differentiation among populations classified into eight well-defined genetic clusters do not allow to unambiguously delineate two species. Instead, continuous level of genetic differentiation indicates interspecific gene flow or drift in isolation. Whereas the analysis of microsatellite variation allowed inferring recent interspecific gene flow, additional nucleotide variation in flanking sequences, by reducing homoplasy, pointed towards ancient interspecific gene flow followed by drift in isolation. The assessment of the weight of each polymorphism in the inference demonstrates the value of linked variation with contrasted mutational mechanisms and rates to refine historical demographic inference. Past population size decline inferred in some of these oak populations as well as low contemporary effective population size for most populations is a concern for the persistence of these populations of high evolutionary significance and conservation value.

Keywords: *Quercus faginea*, *Quercus canariensis*, relict populations, hybridization, drift, effective population size, taxonomy, conservation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Repeated climate changes during the Holocene drove continental-scale species migration with alternative periods of vicariance (drift in isolation) within glacial refugees during glacial periods and secondary contact and gene flow at mid-latitude following post-glacial recolonizations (Hewitt, 1999; Schmitt, 2007). This historical legacy shapes contemporary within-species genetic diversity distribution. As a consequence, current patterns of population genetic structure can shed light on evolutionary history of species (Hewitt, 2001; Petit et al., 2003). While mid-latitudinal landscapes have seen profound climate changes following glacial period leading to major shifts in ecosystems, the situation in low-latitudinal landscapes is less clear (Feliner, 2011). In these regions, including the Mediterranean basin, the effect of climate change is buffered by heterogeneous topoclimates in mountain ranges. As a result, many low-latitude mountain range that are now located at the rear-edge of species distribution acted a glacial refugee sheltering populations from climate variability during glacial cycles (Petit, Hampe & Cheddadi, 2005; Hampe & Jump, 2011). Some of these population effectively acted as a source for postglacial recolonization, while more isolated relict population have evolved locally without taking part to postglacial recolonization (Hampe & Petit, 2005). Unfortunately, continental-scale phylogeographic studies have often failed to sample these relict populations and thus cannot capture their unique evolutionary trajectory. Yet these relict populations represent reservoirs of genetic diversity (Hampe & Petit, 2005), including of adaptive genetic variation (Temunović et al., 2020). Populations located at the rear-edge of species distribution range are also particularly vulnerable to extinction risks posed by ongoing climate changes and increased anthropogenic pressures (Aguinagalde et al., 2005). These threats are especially critical for rear-edge population that persist as small and fragmented populations that face extinction risks due to high demographic and environmental stochasticity (Hampe & Jump, 2011; Alberto et al., 2013). Thus local scale study of rear-edge population should help better understand both past evolutionary history and its consequence for the persistence and conservation of these potentially unique populations.

Trees have been model organisms to understand broad-scale postglacial colonisation pattern following the end of the last glacial maximum (Hewitt, 1999). Their unique life history traits, such as long generation time, and the use of uniparentally inherited chloroplast DNA markers allowed to trace back glacial refugia and postglacial recolonization routes (Petit et al., 2002) or even Tertiary geological events (Magri et al., 2007). Nevertheless, studies at a more local scale in region known to harbour glacial refugia have identified fine-scale population genetic structure within refugia (Gómez & Lunt, 2007), highlighting the value of more spatially focused studies to understand complex

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

evolutionary processes such as vicariance and secondary contact and gene flow at play within these relict populations throughout the Pleistocene and the Holocene (Rodríguez-Sánchez et al., 2010; Feliner, 2014). While the Mediterranean basin is a major centre of biodiversity for the European biota, studies of the North African side of the Mediterranean basin that harbour biodiversity hotspot and potential glacial refugee (Médail & Diadema, 2009) remain scarce (Feliner, 2014). Understanding evolutionary processes in the region is particularly pressing as anthropogenic pressure is rising and likely increases threats posed by ongoing climate change (Hampe & Petit, 2005).

Quercus canariensis Willd. and *Quercus faginea* Lam. are two oak species that are good illustration of the complexity of evolutionary processes that shape present-day biodiversity within species and the threats faced by rear-edge populations. While the taxonomy of *Q. canariensis* is well defined, the taxonomic statute of *Q. faginea* has been the object of debates among taxonomists with two sub-species (*broteroi* and *faginea*) differing by morphological and environmental characteristics being recognized in Algeria among the four subspecies described within *Q. faginea* (Aissi et al., 2021). The origin of such high level of intraspecific variability is unclear and might involve a combination of genetic drift in isolated populations (Moracho et al., 2016) and interspecific gene flow. Whereas interspecific gene flow is frequent among related European oak species (Leroy et al., 2017), it is scarcely documented for *Q. faginea* or *Q. canariensis* (Mir et al., 2006). A better understanding of the origin of this diversity and of the recent demographic trajectory is paramount for evaluating the evolutionary value and conservation status of these populations. The species are located in the western part of the Mediterranean basin within the Mediterranean conifer and mixed forests ecoregion, classified as critically endangered due to deforestation and vulnerability to climate changes (PA1214 ecoregion (Olson et al., 2001)). *Q. canariensis* is listed as Near Threatened at European level in the IUCN's Red List (García Murillo & Harvey-Brown, 2017) due to population decline and increasing drought within the Iberian Peninsula. At the global level, the species is considered as Data Deficient because the species lacks information on population size and status across its range especially in North Africa (Gorener, Harvey-Brown & Barstow, 2017). The species is present in forest stands across Morocco, Algeria and Tunisia; in view of ongoing climate change, these stands deserve specific attention. In Algeria, *Q. canariensis* is located in either vast strands near the coastal area and in few isolated populations in more remote southern mountain locations.

Q. faginea is listed as Least Concern both at the European scale (Harvey-Brown, García Murillo & Buira, 2017) and at the global level (Jerome & Vasquez, 2018) due to widespread and stable abundance within its main range in Spain, Portugal and to a lesser extent to Morocco. In Algeria however, *Q. faginea* is only found in four locations in the form of small and isolated populations of a few tens of hectares including scattered individuals in mountainous landscape (Aissi, Beghami & Heuertz, 2019). Being located at the rear-edge of the distribution range of the species in area of

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

putative refugia and regional hotspot of biodiversity (Médail & Diadema, 2009) and under the direct threat of environmental changes, these populations deserve specific evaluation of their conservation status and value.

Recent developments of sequence-based microsatellite genotyping (Darby et al., 2016; De Barba et al., 2016; Vartia et al., 2016) have improved our understanding of molecular variation at microsatellite repeats and in surrounding flanking sequences (Barthe et al., 2012; Viruel et al., 2018; Lepais et al., 2020). A few previous studies demonstrated that jointly accounting for linked microsatellite and flanking sequence substitutions allows exploiting complementary information provided by polymorphism evolving at different rates and in different ways (Hey et al., 2004; Ramakrishnan & Mountain, 2004; Payseur & Cutter, 2006). Assuming complete linkage due to short physical distance, linked polymorphisms must have experienced the exact same genealogical history. As a result, microhaplotypes combining fast evolving microsatellite repeat variations with the much slowly evolving flanking DNA sequence should provide refined insights about past demographic history with improved temporal range and resolution (Ramakrishnan & Mountain, 2004).

While early works demonstrated that these compound loci can improve biological inference (Hey et al., 2004; Ramakrishnan & Mountain, 2004), they were rarely studied given the difficulty to genotype such compound loci (Mountain et al., 2002; Šarhanová et al., 2018; but see Barthe et al., 2012). These difficulties do not hold true anymore given the development of sequence-based microsatellite genotyping that greatly improves characterisation at microsatellites and flanking sequences (Bradbury et al., 2018; Viruel et al., 2018; Curto et al., 2019; Layton et al., 2020). However, to investigate such data, new statistical approaches that harness information from such microhaplotypes are needed (Payseur & Cutter, 2006).

We report a study of the molecular diversity at sequenced microsatellites to shed light on the taxonomic status of Algerian *Q. faginea* and *Q. canariensis* populations, on their past evolutionary history and on their recent demographic trajectory. Given high pollen dispersal distance, high effective population size and generation time typical of wind-pollinated tree species such as oaks, we expect high local genetic diversity and low regional genetic structure at neutral nuclear loci. This would imply significant gene flow among stands that function as a single population, or recent demographic changes that did not have time to imprint genetic structure. Alternatively, forest fragmentation, reduced effective population size and limited pollen dispersal between remnant forest stands might have led to population isolation and genetic drift resulting in low local genetic diversity and high regional genetic differentiation typical of relict populations (Hampe & Petit, 2005). These two hypothetical demographic histories can be further confounded by historical or contemporary interspecific gene flow that will additionally blurs species limits and contribute to reshuffle genetic diversity across populations. To identify the most likely processes that shape

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

current population genetic diversity, we first compared pattern of molecular variation at 36 sequenced microsatellites across the 11 Algerian *Q. faginea* and *Q. canariensis* populations using standard descriptive approaches. We then used simulation-based inference, flexible enough to model specific linked-polymorphisms observed at each loci and alternative past demographic scenarios. We derived specific summary statistics that account for linked microsatellite and flanking sequence variations. We then applied a Random Forest implementation of approximate Bayesian computation (ABC) that allows assessing the information content of each summary statistics and type of polymorphism to differentiate between concurrent demographic models and to estimate mutational and demographic parameters for the most likely demographic scenario. Implication of the results for the taxonomic status of the species as well as the conservation status and value of the studied populations are discussed in the light of their rear-edge context that makes them especially vulnerable to environmental changes.

Materials and methods

Sampling

Fresh leaves from thirty individuals were sampled in seven *Q. canariensis* and four *Q. faginea* population across Algeria (Table 1; Figure 3d). Geographical coordinates of each sampled individual were recorded and leaves were stored dried in silica gel until DNA extraction.

Table 1: Description of the sampled *Q. faginea* and *Q. canariensis* Algerian population.

Site	Taxonomic identity	Location	Latitude	Longitude	Number of sampled individuals	Proportion of standing individual sampled
S01	*Q. canariensis*	Hafir Tlemcen	34.8326	-1.3748	30	5%
S02	*Q. faginea*	Terni Tlemcen	34.7772	-1.3440	30	40%
S03	*Q. faginea*	Baloul Saïda	34.9844	0.4041	30	60%
S08	*Q. faginea*	Safalou Tiaret	35.4053	1.3486	30	10%
S07	*Q. canariensis*	Thniet El Hed Tessemssilt	35.8655	1.9776	30	5%
S06	*Q. canariensis*	Errich Bouira	36.4057	3.8664	30	5%
S10	*Q. canariensis*	Akfadou Tizi-Ouzou	36.7218	4.4546	29	<1%
S09	*Q. canariensis*	Hamza Jijel	36.7483	5.6065	28	<1%
S11	*Q. faginea*	Chélia Aurés	35.3679	6.6320	28	5%
S04	*Q. canariensis*	Mechrouha Souk Ahras	36.3856	7.8606	30	<1%
S05	*Q. canariensis*	Ghora Bougous El-Taref	36.6739	8.4702	29	<1%

Microsatellite genotyping by sequencing

DNA from each individual was extracted using Invisorb DNA Plant HTS 96 Kit (Invitek) and sequence-based microsatellite genotyping (SSRseq) was used to analyse polymorphism at 36 loci as described

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

elsewhere (Lepais et al., 2020). In short, 60 genomic and EST-derived microsatellites (Kampfer et al., 1998; Durand et al., 2010) were co-amplified in a single multiplexed PCR and sequenced on an Illumina MiSeq sequencer (Supporting Material File 1). After sequence demultiplexing, a bioinformatics pipeline (Lepais et al., 2020) integrating the FDSTools analysis toolkit (Hoogenboom et al., 2016) was used to convert raw sequence into genotypic data (microhaplotypes) integrating all polymorphisms identified within the microsatellite itself (number of repeats, SNP and Indel) and in the flanking sequences (SNP and Indel). The bioinformatics pipeline also compares genotypes from repeated individuals to estimate allelic error rate and compute the overall missing data rate for each locus. This information about data quality allows optimising the bioinformatics analysis strategy. In particular, it can help identify loci that cannot be reliably genotyped or for which the genotyping accuracy is only acceptable when analysing the polymorphism within the repeat motif itself, i.e. not accounting for the polymorphism in the flanking sequence (Lepais et al., 2020). Finally, for each locus, every allele differing from the other by any polymorphism was coded under an arbitrary three-digit scheme with a unique number assigned to each microhaplotype within each locus.

Population genetic diversity and structure

Assuming Hardy-Weinberg equilibrium, we tested for heterozygote deficit or excess for each locus in each population using Genetix (Belkhir et al., 2004) to identify outlier loci presenting extreme inbreeding coefficient that could indicate presence of null allele (positive Fis) or paralogs (negative Fis).

As clonal reproduction has been reported for oaks in harsh environmental conditions (Alberto et al., 2010), we identified clones using Colony (Wang, 2016) and kept only one genotype from each detected genet.

The studied loci were partly derived from expressed transcripted sequences (Durand et al., 2010) or selected to differentiate *Q. robur* and *Q. petraea* (Scotti-Saintagne et al., 2004; Lepais, Leger & Gerber, 2006), therefore all may not behave neutrally. Bayescan v2.1 (Foll & Gaggiotti, 2008) was used to identify and remove loci departing from neutrality as most of the following analyses assume neutral genetic evolution.

Gene diversity (Nei's unbiased heterozygosity), allelic richness after rarefaction to 16 diploid individuals, Fis and Fst were computed for each sampled location with FSTAT (Goudet, 1995), genetic differentiation accounting for allele distance in number of repeat at microsatellite alleles (Rst) was computed using SPAGeDi (Hardy & Vekemans, 2002) and genetic differentiation between species tested using 10000 permutations of individual between species. Different genotypic datasets accounting either for the whole haplotype, the flanking sequence haplotype, the number of microsatellite repeat or the whole allele size were used to highlight potential difference in genetic

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

structure between polymorphisms. Statistical difference of genetic diversity and differentiation within species were tested by comparing locus-level statistics using a Mann-Whitney test. The most informative whole haplotype genotypic dataset was used for subsequent analysis.

The contemporary effective population size was estimated for each sampled location using the linkage disequilibrium method implemented in NeEstimator v2.1 (Do et al., 2013) assuming random matting, removing alleles with frequency below 0.02 (Waples & Do, 2010) and using the parametric chi-squared method to estimate 95% confidence interval.

Structure v 2.3.4 (Pritchard, Stephens & Donnelly, 2000; Falush, Stephens & Pritchard, 2003) was used to explore the genetic structure between and within species. We applied the Admixture model and the Correlated Allele Frequency model with 20 replicated runs for an assumed number of genetic clusters (k) ranging from 1 to 12. To remove runs that did not converged and produced outlier results, the 15 most likely runs for each k value were kept for further inspection. Structure Harverster (Earl & vonHoldt, 2012) was used to inspect likelihood trends of the runs and Clumpak (Kopelman et al., 2015) to identify and filter-out sub-optimal runs.

The geographical context of the population genetic structure was illustrated using a network graph built with EDENetworks (Kivelä, Arnaud-Haond & Saramäki, 2015) with node location representing the geographical coordinates of samples locations and edges linking sites showing pairwise genetic differentiation value (Fst) lower than 0.04 (determined by automatic thresholding), with edge width inversely proportional to Fst.

Population demographic history

While the Structure results revealed that each of the 11 studied populations can be unambiguously assigned to one of the eight delineated genetic clusters (see Results), each population cannot be unambiguously assigned to one of the two expected species due to genetic continuity of the studied populations. We thus investigated the most likely demographic history that led to the observed continuous pattern of genetic differentiation. One hypothesis is that the populations diverged separately from the main core populations and that different levels of genetic drift across populations led to a continuous gradient of genetic differentiation that blurs species delineation. A second hypothesis is that gene flow taking place after a divergence period (secondary contact) produced admixed genotypes that, following recombination and drift, have mingled to form a variety of intermediate populations. While interspecific gene flow following secondary contact has been found to be the rule for sympatric European white oak species (Leroy et al., 2017), the studied species are not sympatric due to different ecological requirements. As a result, interspecific gene flow could only happen during secondary contact or through continuous levels of low interspecific gene flow over long distances. We simulated concurrent demographic models consisting of

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

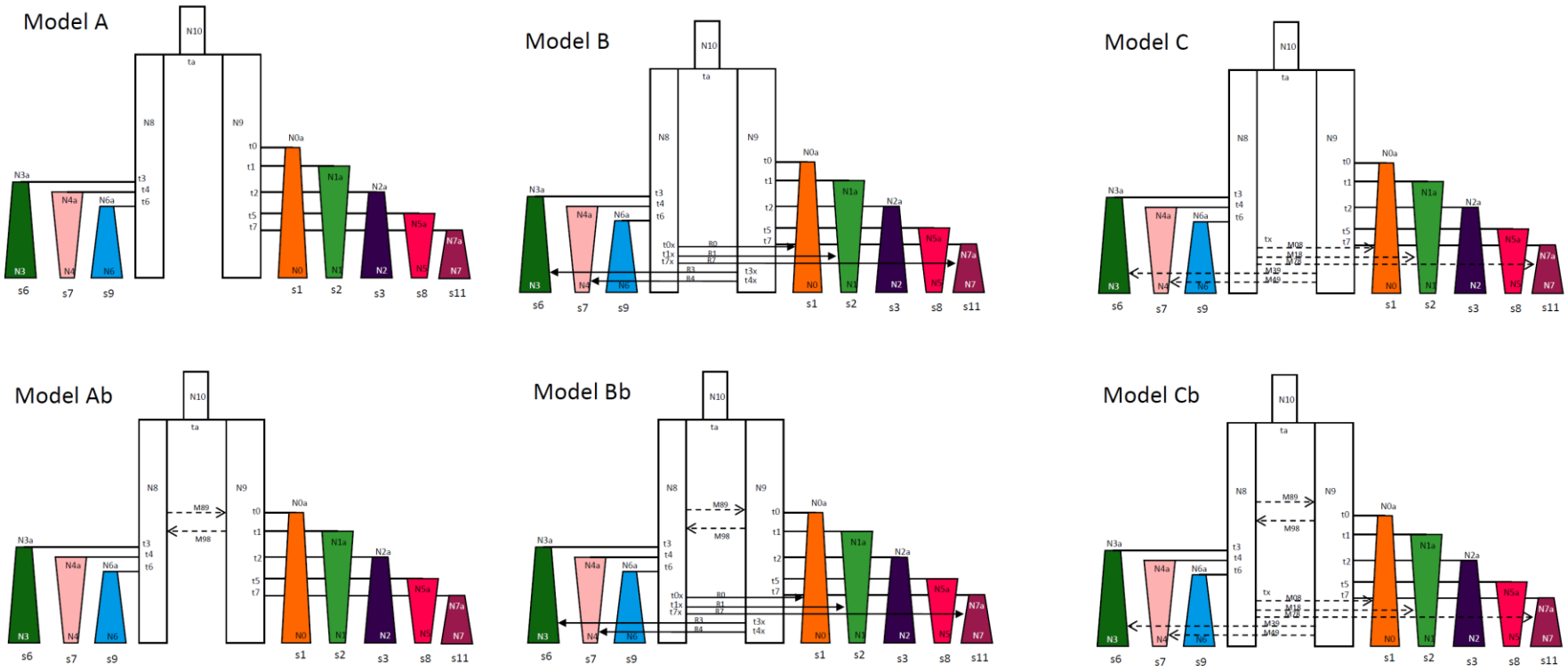
population divergence without gene flow (model A), population divergence with a single interspecific gene flow event (model B) or with continuous interspecific gene flow (model C) after a period of divergence without gene flow (Figure 1). Three additional models include long-term continuous interspecific gene flow between core populations of the two species (model Ab, Bb and Cb, respectively, Figure 1). Gene flow between current populations and unsampled core range of the other species was modelled as directional from the core to the isolated heterospecific population because we were interested in the effect of interspecific gene flow on the studied population. In addition, we allowed interspecific gene flow only for populations showing hypothetical admixture from Structure results (Figure 3). Prior distribution of demographic parameters has been setup wide to reflect the lack of knowledge about the history of these populations (Table 4, Supporting File 1) and checked for their capacity to produce realistic genotypic data by comparing observed and simulated genotyping datasets (Cornuet, Ravigné & Estoup, 2010; Leroy et al., 2017, Supporting Material File 3).

Molecular variability was simulated following each of the demographic model using the coalescent implemented in fastsimcoal2 v2.6.0.3 (Excoffier et al., 2013). We simulated realistic molecular diversity that closely matches the characteristics of each of the studied loci. We focused on two kind of variability that can be simulated: variation in repeat number at the 36 microsatellites and substitutions in the flanking sequence for the 20 loci for which flanking sequence could be reliably analysed (called HapSTR loci thereafter) and assuming complete linkage between the microsatellites and their flanking sequences. This procedure accounts for 90% of the source of variations in the dataset (Lepais et al., 2020), leaving additional polymorphisms (SNP in the repeated motif or indel in the flanking sequence) that is far more complex to model due to uncertainty of the mutational mechanisms and rates.

Each microsatellite could mutate at rates ranging from 10^{-5} to 10^{-2} mutation/generation/haploid genome following Generalized Mutation Model with parameter p between 0.1 and 0.5. Similarly, substitution in the flanking sequence could mutate at rates ranging between 10^{-9} and 10^{-6} following neutral model of evolution. Each locus system had its own prior distributions for these parameters allowing them to have a wide range of mutational dynamics.

To compare the six alternative demographic models, 12,000 genotypic datasets were simulated for each model with demographic and mutational parameter values sampled from prior distributions. To summarize the distribution of genetic diversity among populations and locus, a total of 530 summary statistics were derived (Supporting Material File 2). We used arlsumstat (Excoffier & Lischer, 2010) to compute 294 summary statistics that take into account the different sources of molecular variation.

Figure 1: Alternative *Q. faginea* and *Q. canariensis* Algerian populations demographic history models simulated. Present-day populations diverged from their respective species main distribution area without any additional interspecific gene flow (models A and Ab), or with interspecific gene flow represented either by an instantaneous single event of interspecific gene flow (models B and Bb) or a continuous interspecific gene flow (models C and Cb). Models Ab, Bb and Cb account for long-term interspecific gene flow between species whereas models A, B and C assume that species diverged without interspecific gene flow. Unsampled (ghost and ancestral) populations are indicated in white, samples population are coloured according to their main Structure's genetic cluster. Arrow with continuous line indicates single instantaneous interspecific gene flow, arrow with dashed line indicates continuous interspecific gene flow. N: effective population size, t: time of event, tx: time of secondary contact; R and M: instantaneous and continuous interspecific migration rate, s: sampled population code. Rectangle indicates fixed effective population size, isosceles trapezoid indicates variable effective population size since time of divergence; shapes and sizes are not to scale in term of time of event or effective population size.



bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A first set of 98 statistics were computed with microsatellites and, when available, flanking sequences combined as microhaplotypes "considered as mutationally equidistant from each other" (standard data in Arlequin nomenclature) using the 36 loci. A second set of 135 statistics were computed based on the number of repeats from all 36 loci (microsatellite data type in Arlequin). Finally, for the 20 HapSTR loci, an additional set of 61 statistics were computed based on substitution in the flanking sequences. The covariation in diversity at linked polymorphisms might contain useful information about mutational and demographic processes (Payseur & Cutter, 2006). We thus computed 224 additional summary statistics taking advantage of linked polymorphism at HapSTR loci level using a custom R script (Supporting Material File 4). First, 136 of these statistics were locus-level variability such as variance in allele size for the microsatellite (V), number of microsatellite alleles, number of flanking haplotypes, nucleotide diversity (theta). Second, 88 statistics interrogate the association between the repeated motif number and the flanking haplotype, such as the correlation between theta and V (Payseur & Cutter, 2006) or the mean number of microsatellite alleles per flanking sequence haplotype across microhaplotypes.

We performed model choice by comparing simulation-based summary statistics from the six models to the observed summary statistics from the real dataset within the Approximate Bayesian Computation framework (Beaumont, Zhang & Balding, 2002) and using the Random Forest approach implemented in the R package abcrf (Marin et al., 2015; Raynal et al., 2016).

We first built a classification random forest model using 1000 trees and a training dataset consisting of simulation-based summary statistics obtained from all models. We estimated the classification prior error rate for each model using an "out-of-bag" procedure to estimate the power of the genetic data to differentiate between the demographic models given the model and prior specifications (Marin et al., 2015). Then, we used the summary statistics computed based on the observed genotypic data to predict the demographic model that best fit the data using a regression forest with 1000 trees. In addition, we also repeated the model choice analysis using only the 231 summary statistics derived from the microsatellite motifs from the 36 loci.

We then used the overall most likely scenario to simulate 100,000 additional genetic datasets using parameters and prior distributions described above to estimate demographic model parameters. We built a regression random forest model implemented in the R package abcrf based on the summary statistics using 1000 trees. The out-of-bag procedure (Raynal et al., 2016) was used to assess the prediction error and inference power by comparing true (simulated) and estimated parameter values. We then estimated the posterior median, 0.05 and 0.95 quantiles of demographic and mutational parameters using random forest regression model based on all summary statistics computed from of the observed haplotypic dataset.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Finally, we assessed the value of different polymorphisms for demographic inference by quantifying summary statistics importance for both model choice and parameter estimation (Estoup et al., 2018). To better interpret the significance of variable importance, we introduced 12 additional random summary statistics ("noise variable") in the datasets (Chapuis et al., 2020). These consisted in four random floating point variables sampled between 0 and 1, four random floating variables sampled between 0 and 10 and four random natural number variable sampled between 1 and 30. Being random, these noise variables should not contribute to the inference, and thus will provide a benchmark to assess the significance of the variable importance metrics reported for each summary statistics (Chapuis et al., 2020).

Results

Genotyping

After filtering for missing data, genotyping error, Hardy-Weinberg equilibrium, neutrality and clonality (see Supporting Material File 2), the final genotypic dataset consisted in 36 high quality loci (20 linked flanking sequence and microsatellite loci, called HapSTR thereafter, and 16 loci that failed been reliably analysed across the whole haplotype and for which only the microsatellite motif was analysed, called STR thereafter, Supporting Material File 2) genotyped in 318 individuals.

Molecular variation over loci

In addition to variation in the number of repeats, polymorphisms within the repeated microsatellite motif and the flanking sequences were the rule rather than the exception (Supporting Material File 2). The number of alleles when considering allele size only (438 alleles in total) is half the number of alleles when considering all source of molecular variations (899 haplotypes in total, Supporting Material File 2). As a result, size homoplasy amounts to 51% overall (56% for HapSTR and 31% for STR). Across the 36 loci, 32 showed SNP within the repeat motif (for a total of 115 SNP), nine show indel within the repeat motif (for a total of 10 indel) and nine consisted of more than one repeat motif (compound microsatellites). Nucleotide substitutions among the flanking sequence for the 20 HapSTR were widespread with an average of 7.5 SNP per loci (min: 3, max: 14) and an average of 1 substitution every 6.7 nucleotides (15% of flanking sequence position showed a substitution). Half of the HapSTR showed an indel with an average of 0.9 indel per loci (min: 0, max: 4).

Molecular variation across populations

Interspecific genetic differentiation (Fst) varies from 0.030 when computed based on flanking sequence haplotype to 0.045 when computed from the whole haplotype (Table 2). Differentiation among *Q. canariensis* populations (average: 0.026) was lower than among *Q. faginea* populations

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(average: 0.043) irrespectively of the polymorphism considered. The comparison of Fst computed from the whole haplotypes among *Q. canariensis* populations (Fst=0.026) and among *Q. faginea* populations (Fst=0.049) was statistically significant (Mann-Whitney z-score = -3.73, p-value = 0.0002). The inflation of differentiation among *Q. faginea* populations is mostly driven by differentiation at microsatellite variation which translated into higher Fst estimates based on allele size, repeat number or whole haplotype (Table 2). By contrast, differentiation at flanking sequence is not statistically different among *Q. canariensis* and *Q. faginea* populations (z-score = -0.35, p.value = 0.7363) and *Q. canariensis* differentiation is similar at the microsatellite and the flanking sequence (0.027 and 0.024 respectively, Table 2).

Table 2: Between and within species genetic diversity and differentiation computed using different polymorphisms.

	Parameter \ Dataset		Allele size	Repeat number	Flanking sequence SNP haplotype	Whole haplotype
Genetic differentiation	Fst	Interpsecific	0.044	0.043	0.030	0.045
		Q. canariensis	**0.028**	**0.027**	0.024	**0.026**
		Q. faginea	**0.047**	**0.046**	0.030	**0.049**
		Mann-Whitney z-score (p-value)	*-2.78 (0.0054)*	*-2.80 (0.0051)*	*-0.35 (0.7263)*	*-3.73 (0.0002)*
	Rst	Interpsecific		0.035		
		Q. canariensis		0.028		
		Q. faginea		0.048		
		Mann-Whitney z-score (p-value)		-1.52 (0.1285)		
Genetic diversity	Variance in allele size	*Q. canariensis*		4.31		
		Q. faginea		4.65		
		Mann-Whitney z-score (p-value)		-0.34 (0.7279)		
	Nucleotidic diversity	*Q. canariensis*			0.0073	
		Q. faginea			0.0065	
		Mann-Whitney z-score (p-value)			-0.20 (0.8415)	

Genetic differentiation based on microsatellite allele distances (Rst) was not statistically significant from genetic differentiation considering only allele identity (Fst) at the intraspecific and interspecific level (all permutation tests not significant) indicating that mutation is not the driving factor of microsatellite variation distribution across population and species. This suggests that other factors, such as drift or interspecific gene flow, are involved in the increased differentiation at microsatellite for *Q. faginea*.

Genetic diversity did not differ between species (Table 2): variance of allele size at microsatellites amount to 4.31 and 4.65 and nucleotide diversity at flanking sequence to $7.3.10^{-3}$ and $6.5.10^{-3}$ for *Q. canariensis* and *Q. faginea*, respectively. Locus-level comparison showed low and not significant correlation between variance of allele size at microsatellite and nucleotide diversity at linked flanking sequence (Pearson product moment correlation: 0.15 (t = 0.62, df = 16, p-value = 0.54) and 0.20 (t = 0.84, df = 17, p-value = 0.41) for *Q. canariensis* and *Q. faginea* respectively).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

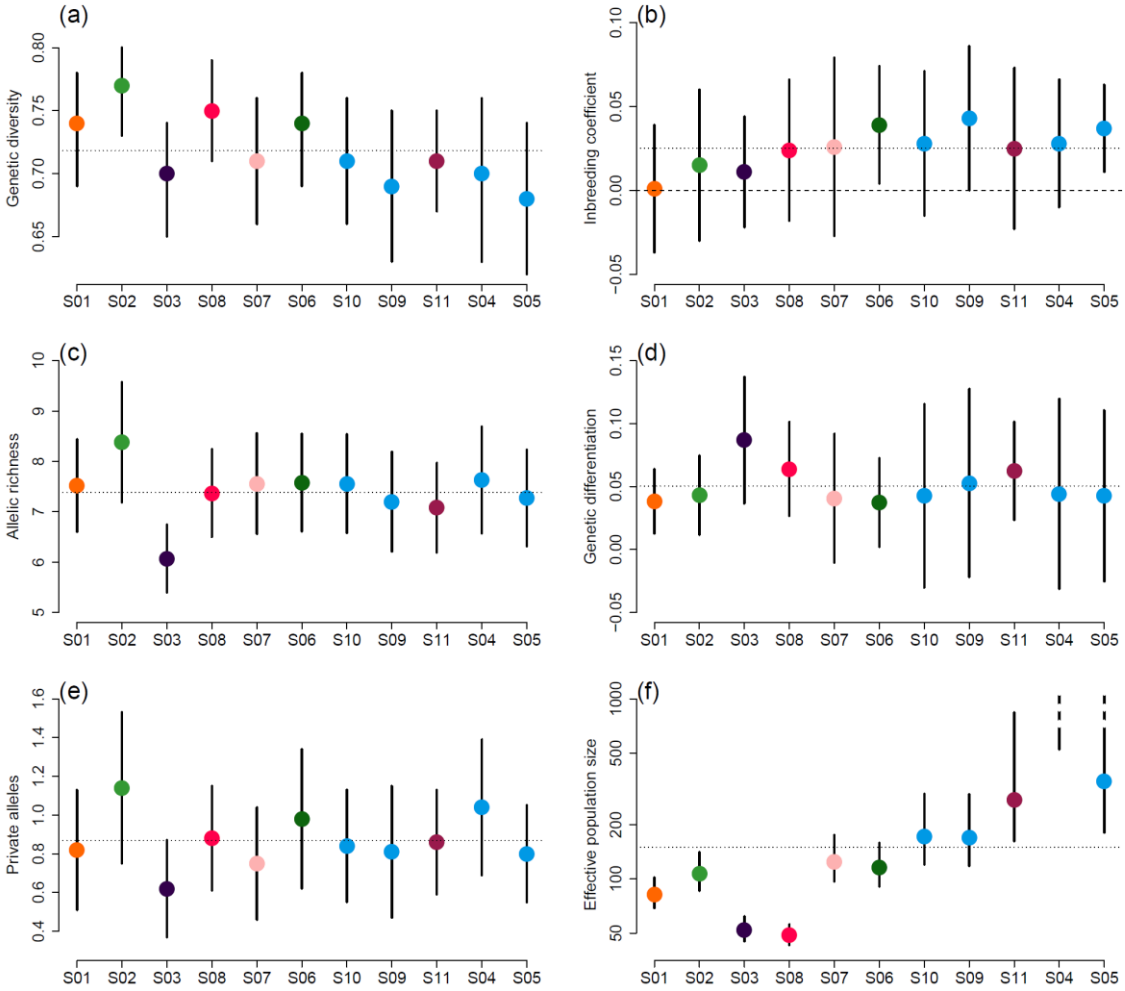
Microhaplotype genetic diversity distribution across species and populations

Genetic diversity expressed as heterozygosity, allelic richness and private alleles showed similar values across populations (Figure 2), except for populations S02 and S03 that harbour higher and lower diversity, respectively (Figure 2a, 2c and 2e). Inbreeding coefficients present a geographical pattern of variation from lower values to the west to higher values to the east of the studied area but were all not significant from 0 after Bonferroni correction for multiple testing (Figure 2b). Population-specific genetic differentiation is the highest for S03, S08 and S11 populations and shows the highest variability for *Q. canariensis* S04, S05, S09 and S10 populations (Figure 2d) which is consistent with *Q. canariensis* populations showing low divergence from each other and *Q. faginea* populations showing higher distinctiveness in the genetic differentiation network (Figure 3d). Contemporary effective population size shows high variability with estimates ranging as low as 50 individuals for S03 and S08 population to 350 individuals for S05 (Figure 2f). Excluding the S04 population for which sampled size is too low to get a point estimate indicating a higher effective population size (95% lower bound confidence interval estimated at 524 individuals), the averaged contemporary effective population size amount to 150 individuals.

Structure analysis assuming two genetic clusters broadly delineate the two species (Figure 3a): *Q. canariensis* S04, S05, S06, S09, S10 populations in the one hand and *Q. faginea* S03 and S08 populations in the other hand clearly fall apart in their own specific cluster. However, *Q. canariensis* S01, S06 and S07 populations and *Q. faginea* S02 and S11 populations showed admixture or unsorted ancient ancestry. While the population S01 had been morphologically affected as *Q. canariensis*, it was found to be more genetically similar to *Q. faginea.* Assuming a higher number of genetic clusters resulted in a more refined view of the intraspecific genetic structure (Figure 3b, Supporting Material File 1). Assuming eight genetic clusters (Supporting Materiel File2) results in a clear delineation of 7 populations into specific clusters in addition to a genetic cluster regrouping 4 *Q. canariensis* populations (Figure 3b). Differentiation between genetic clusters do not showed a bipartite hierarchical structure expected when analysing two species (Figure 3c). Instead, five of the cluster showed gradual level of differentiation starting from the *Q. canariensis* side (right part of the tree in Figure 3c), while three genetic cluster on the *Q. faginea* side showed high genetic divergence (right part of the tree in Figure 3c). Populations with intermediate ancestry coefficient assuming two genetic clusters (S02, S01, S06, S07 and S11) now constitute independent clusters that located at intermediate positions between the two species extremities (Figure 3c).

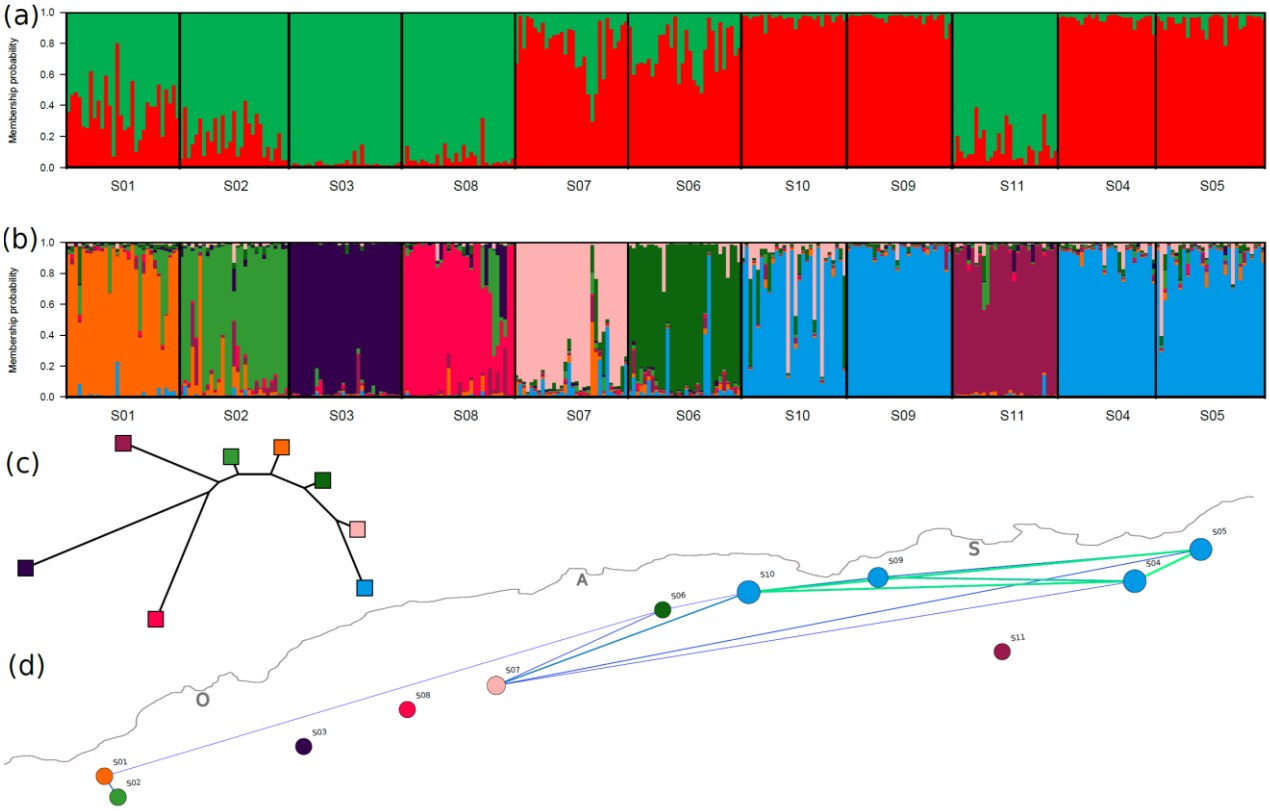
bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2: Genetic diversity of *Q. faginea* (S02, S03, S08, S11) and *Q. canariensis* (S01, S04, S05, S06, S07, S09, S10) Algerian populations: (a) genetic diversity, (b) inbreeding coefficient, (c) allelic richness, (d) genetic differentiation (summarized by averaging all pairwise genetic differentiation estimates that involved the focal population), (e) private alleles and (f) contemporary effective population size. Point represents the mean over loci and the bar shows estimate variability (plus and minus two standard deviations over loci). Population are ordered from west to east along the x-axis, colours refer to Structure results, horizontal dashed lines represent the average value across populations and (f) vertical dashed lines uncertain effective population size estimates.



bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3: Population genetic structure of *Q. faginea* and *Q. canariensis* Algerian populations: Bayesian genetic clustering Structure results assuming (a) two or (b) eight genetic clusters, (c) phylogenetic tree illustrating the allele frequency divergence (Structure's Net nucleotide distance) among the eight genetic clusters; (d) EDENetwork network graph projected on a geographic map with nodes indicating geographical location of the sampled sites, node size is proportional to the number of connection with other nodes and colours reflect the dominant genetic cluster, edges link population with a pairwise genetic differentiation (Fst) below 0.04 with edge width and greenness inversely proportional to genetic differentiation. O: Oran, A: Alger and S: Skikda.



Demographic inferences

Model choice

When using the information from the microsatellite and the linked flanking sequence (HapSTR), the most likely model to explain the observed data was model Ab (posterior probability: 0.47, Table 3a) which consisted in species divergence with gene flow followed by independent divergence of populations without additional more recent interspecific gene flow (Figure 1). The out-of-bag procedure estimated that given the choice of the model Ab and the uncertainty in differentiating the models given the data and the prior distribution settings, there is 49% probability that the true model is indeed the model Ab, and a 18% or 17% probability that the true model is model Cb and Bb respectively (Table 3b).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 3: Random forest ABC Model choice (a) by using jointed microsatellite and flanking sequence information (HapSTR) or information derived only from the microsatellite variations (SSR only). Power to differentiate alternative demographic models for HapSTR dataset (b) given the model and prior range specifications. The table reports how many of the 12,000 simulated datasets generated under a specific scenario (in rows) are classified into each demographic scenario (in column). The overall scenario classification error is computed based on the number of incorrectly classified dataset; the last column shows the percentage of simulated datasets classified as Ab, the most likely scenario for the observed genetic dataset. Bold numbers indicate >10% incorrect classification, underlined number indicates correct classifications.

(a)

	Votes Model A	Votes Model B	Votes Model C	Votes Model Ab	Votes Model Bb	Votes Model Cb	Most likely model	Posterior probability
HapSTR	163	103	114	**271**	191	158	Ab	0.47
SSR only	31	113	82	169	286	**319**	Cb	0.40

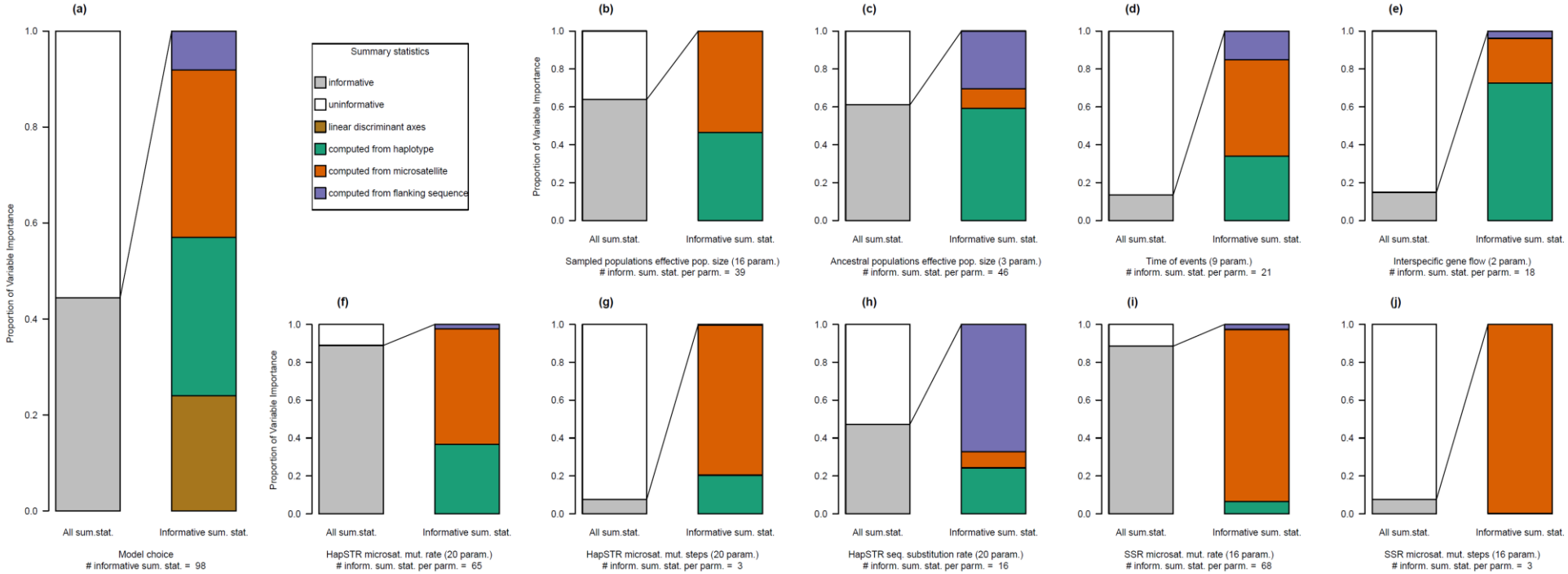
(b)

HapSTR-Classified models HapSTR-Simulated models	Model A	Model B	Model C	Model Ab	Model Bb	Model Cb	total	Scenario classification error	Probability given the selected best model is Ab
Model A	**<u>10142</u>**	183	290	1042	183	160	12000	15.5%	8%
Model B	54	**<u>8017</u>**	**2087**	623	765	454	12000	33.2%	5%
Model C	479	**3905**	**<u>5905</u>**	562	520	629	12000	50.8%	4%
Model Ab	450	228	227	**<u>6685</u>**	**2419**	**1991**	12000	44.3%	**<u>49%</u>**
Model Bb	35	559	305	**2248**	**<u>5488</u>**	**3365**	12000	54.3%	17%
Model Cb	37	366	522	**2380**	**3769**	**<u>4926</u>**	12000	59.0%	18%
total	11197	13258	9336	13540	13144	11525		42.8%	

A total of 98 summary statistics had variable importance higher than randomly generated statistics, and thus were considered informative to differentiate alternative scenario (Figure 4a). These informative summary statistics represented 44% of the total variable importance. Among them, 24% derived from linear discriminant axes, 35% from summary statistics computed from microsatellite variation, 33% from haplotype and only 8% from the substitution within the flanking sequences (Figure 4a).

It should be noted that when considering the microsatellite variability only, the most likely scenario was model Cb (posterior probability: 0.40, Table 3b) which consisted in species divergence with gene flow followed by independent divergence of populations followed by interspecific gene flow after a period of strict isolation (Figure 1). However, because HapSTR integrates more molecular information, this strategy is more decisive (posterior probability of 0.48 and 0.40 for HapSTR and STR respectively) and powerful (overall classification error of 42.7% and 44.4% for HapSTR and STR respectively), we further investigate the scenario Ab to estimate its demographic parameters using the HapSTR dataset.

Figure 4: Contribution of different types of polymorphism measured as Variable Importance (y-axis) for (a) model choice and parameter estimation: (b) samples and (c) unsampled population effective population size, (d) time of event, (e) interspecific gene flow, HapSTR microsatellite (f) mutation rate, (g) mutation steps and substitution rate (h), SSR microsatellite mutation rate (i) and mutation steps (j). In each subfigure, the leftmost bar shows the proportion of total Variable Importance resulting from informative summary statistics, while the rightmost bar shows the contribution of the different groups of summary statistics derived from different type of polymorphisms. The abscise legend reports the number of estimated demographic or mutational parameter in each group and the mean number of informative summary statistics per parameter.



bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Demographic and mutational parameter estimations

Microsatellite mutation and substitution rates were accurately estimated (all posterior MNSE lower than 0.11; Supporting Material File 2) with an average of 65 and 68 informative summary statistics representing 89% and 88% of the total variable importance for microsatellite mutation rate at HapSTR loci and SSR loci respectively (Figure 4f and 4i) and an average of 16 informative parameters for substitution rate that represents 47% of the total variable importance (Figure 4h). On the contrary, the parameter p controlling the number of mutational steps at microsatellite was difficult to estimate (posterior MNSE between 2.6 and 7.6; Supporting Material File 2) and had only an average of 3 informative summary statistics per parameters that represented 7.6% of the total variable importance for all loci (Figure 4g and 4j).

Estimated microsatellite mutation rates varied from 2.2×10^{-5} to 2.1×10^{-3} mutation per generation per haploid genome (mean: 3.1×10^{-4}, 95% CI: $4.5 \times 10^{-5} – 2.4 \times 10^{-3}$) and flanking sequence substitution rate varied from 4.5×10^{-8} to $4.8\ 10^{-7}$ (mean: 2.5×10^{-7}, 95% CI: $2.8 \times 10^{-8} – 9.0 \times 10^{-7}$, Supporting Material File 2).

Posterior effective population size parameters showed restricted ranges compared to prior and reduced estimation error (posterior NMAE below 0.32, Table 4). Contemporary and ancestral effective population sizes of sampled populations were estimable with an average of 39 informative summary statistics (Figure 4b) that represented 64% of the total variable importance. Effective population size of unsampled populations (i.e. *Q. canariensis* and *Q. faginea* core range, and their ancestor) were estimable with an average of 46 informative summary statistics (Figure 4c) that represented 61% of the total variable importance.

Effective population size estimates showed some variation between populations (Table 4). Effective population size at divergence time were generally higher than recent effective population size, with clear sign of effective population size decreases for S08 (ratio N5a/N5 = 9.3, N5 = 3908 haploid genomes), S11 (ratio of N7a/N7 = 10.3, N7 = 1898 haploid genomes) and S3 (ratio of N2a/N2 = 24.3, N2 = 1667 haploid genomes). Populations that did not experienced significant effective population size change showed relatively moderate (~10000 for S09, ~20000 for S01, S06 and S07) or high (~40000 for S02) effective population size.

Overall, only 36 out of the 518 summary statistics were not informative for demographic or mutational parameter estimations, most of them were associated with the flanking sequence diversity or divergence and the correlation between diversity at microsatellite and flanking sequence (Supporting Material File 2).

Table 4: Demographic parameter prior specification and posterior estimations for the most likely model Ab.

		Prior distribution			Posterior distribution							
Parameter	Sampled site	distribution	min	max	median	lower 95% confidence interval	upper 95% confidence interval	posterior NMAE	median effective population size natural values	lower 95% confidence interval	upper 95% confidence interval	Ancient / contemporary effective population size
N0a	S01	log uniform	1	6	4.407	1.820	5.963	0.28	25546	66	918883	
N0	S01	log uniform	1	6	4.224	2.744	5.890	0.14	16753	555	777048	1.5
N1a	S02	log uniform	1	6	4.587	2.064	5.936	0.27	38661	116	863569	
N1	S02	log uniform	1	6	4.600	2.988	5.911	0.13	39792	972	814631	1.0
N2a	S03	log uniform	1	6	4.607	1.663	5.920	0.28	40447	46	831414	
N2	S03	log uniform	1	6	3.222	1.863	5.487	0.18	1667	73	307105	24.3
N3a	S06	log uniform	1	6	4.432	2.072	5.908	0.26	27060	118	808269	
N3	S06	log uniform	1	6	4.334	2.713	5.935	0.14	21576	516	860303	1.3
N4a	S07	log uniform	1	6	4.400	2.176	5.926	0.26	25102	150	842604	
N4	S07	log uniform	1	6	4.226	2.582	5.921	0.16	16814	382	834116	1.5
N5a	S08	log uniform	1	6	4.559	1.531	5.909	0.32	36232	34	811686	
N5	S08	log uniform	1	6	3.592	2.004	5.818	0.19	3908	101	658238	9.3
N6a	S09	log uniform	1	6	4.128	1.708	5.899	0.30	13424	51	792957	
N6	S09	log uniform	1	6	4.077	2.412	5.915	0.17	11948	258	821561	1.1
N7a	S11	log uniform	1	6	4.292	1.398	5.935	0.31	19576	25	861619	
N7	S11	log uniform	1	6	3.278	1.785	5.936	0.19	1898	61	863933	10.3
N8		log uniform	1	6	4.517	2.332	5.919	0.18	32859	215	829199	
N9		log uniform	1	6	4.577	2.666	5.914	0.16	37758	463	820374	
N10		log uniform	1	6	3.751	1.079	5.538	0.30	5631	12	345095	
t0	S01	uniform	10	400	152	32	320	0.76				
t1	S02	uniform	10	400	152	32	346	0.77				
t2	S03	uniform	10	400	161	32	342	0.76				
t3	S06	uniform	10	400	155	30	337	0.69				
t4	S07	uniform	10	400	161	32	344	0.75				
t5	S08	uniform	10	400	167	29	330	0.74				
t6	S09	uniform	10	400	153	29	335	0.77				
t7	S11	uniform	10	400	167	31	346	0.72				
ta		uniform	401	10000	5462	723	9772	0.81				
M89		uniform	0.0001	0.001	0.00490	0.00033	0.00978	1.62				
M98		uniform	0.0001	0.001	0.00449	0.00025	0.00983	2.10				

NMAE (normalized mean absolute error): average difference between estimated and true value normalized over the average of all true values.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

Sequence information around microsatellite loci can uncover molecular variations caused by different mutation dynamics and mechanisms that generate complementary patterns across populations and species (Hey et al., 2004; Barthe et al., 2012). If the dataset combining all polymorphisms under microhaplotypes showed the most clear-cut results to depict the geographical distribution of the genetic clusters, we also observed different patterns of diversity distribution across populations according to the polymorphisms considered. For instance, microsatellite variation drives the build-up of genetic differentiation among *Q. faginea* populations, but not among *Q. canariensis* populations. Given that microsatellite loci evolve faster than DNA sequence, increased differentiation at fast-evolving loci might suggest recent demographic events affecting *Q. faginea* populations. Pattern of microsatellite differentiation is not linked to newly acquired mutations because accounting for phylogenetic distance between microsatellite alleles does not yield a higher population differentiation. However, highly polymorphic multi-allelic microsatellite should be more sensitive to the recent effect of genetic drift. Remarkably, estimates of genetic differentiation among *Q. faginea* populations were similar to interspecific genetic differentiation, suggesting that interspecific gene flow might blur species boundaries with morphologically undetected hybridization and directional introgression introducing heterospecific alleles within *Q. faginea* populations. A combination of interspecific gene flow and recent demographic events might account for this pattern of molecular variation, highlighting the limits of descriptive analysis to resolve complex demographic history. Nevertheless, contrasted variability between differently evolving polymorphisms suggests complementary information about past population history. Simulation-based inferences that allow for realistic modelling of molecular evolution at linked loci in populations should leverage such complementary information for improved historical inferences.

We found alternative most likely demographic scenarios when considering microsatellites alone or in conjunction with variation in their flanking sequences. Recent interspecific gene flow was inferred when using microsatellite variation while divergence without interspecific gene flow was the most probable scenario when accounting for flanking sequence variation. Homoplasy (identical allelic state derived from different ancestor alleles) at microsatellite loci explains this discrepancy: with the underestimation of interspecific divergence due to homoplasy (Estoup, Jarne & Cornuet, 2002) interspecific gene flow needs to be invoked to explain the observed pattern of microsatellite variation between species. Indeed, homoplasy at STR loci is expected to overestimate hybridization rate and underestimates the timing of interspecific gene flow events (Dickey et al., 2013; Henriques, von der Heyden & Matthee, 2016).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Contrary to previous implementation of ABC, Random Forest ABC is not sensitive to the number and relevance of the summary statistics (Marin et al., 2018) and can assess the weight of each summary statistics in the inference (Raynal et al., 2016; Estoup et al., 2018). We leveraged these new opportunities to compare the value of microsatellite variations and flanking sequence substitutions or the combination of both to inform biological inferences. For model choice, all types of polymorphisms provide some information to differentiate the concurrent demographic models. For population demographic parameters, population-level diversity and pairwise difference between populations computed from haplotype and microsatellite inform about effective population size of sampled populations, while for ancient events such as effective population size of ancestral populations, the level of genetic diversity within flanking sequence brings some additional information. While population-level correlation between genetic diversity at flanking sequence and microsatellite was suggested to be sensitive about past demographic events (Payseur & Cutter, 2006), nine of those summary statistics were among the 33 uninformative summary statistics out of the total of 518 summary statistics derived in this study. Microsatellite mutation rate at HapSTR and SSR loci and substitution rate at flanking sequence are overall informed by population level summary statistics, while locus-specific mutation parameters are specifically informed by the corresponding locus-level summary statistics. These results confirm the relevance of inference at the locus level in order to capture realistic picture of mutational dynamic variability across loci (Chapuis et al., 2020). Beside assessing the informative content of molecular variation to infer past demographic history, the aims of the study was to evaluate the role of (1) interspecific gene flow and (2) recent demographic dynamics in structuring population genetic diversity.

The genetic clustering analysis assuming two genetic clusters showed highly admixed populations, which might be interpreted as evidence for recent interspecific gene flow. However, these admixed populations differentiated into specific genetic cluster when assuming eight genetic clusters. Contrary to expectations, the genetic distance between the eight genetic clusters does not clearly delineate species boundary, but instead shows a continuous gradient in genetic distance that might be due to ancient hybridization or retention of ancestral polymorphism. Notably, within-species genetic differentiation of 0.028 and 0.047 for *Q. canariensis* and *Q. faginea* microsatellite allele sizes are intermediate between typical estimates of microsatellite genetic differentiation among European white oaks core range populations (Muir et al., 2004; Alberto et al., 2010; Neophytou et al., 2015) and isolated marginal populations (Buschbom, Yanbaev & Degen, 2011; Moracho et al., 2016). Our results indeed suggested that interspecific gene flow predated population divergence in isolation. Divergence by drift was found to be the driving factor shaping genetic structure of small isolated *Q. faginea* and *Q. canariensis* populations. Recent drift in isolation was exacerbated in declining *Q. faginea* populations and affected to a lesser extent the remaining *Q. faginea* and some

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Q. canariensis populations. This has led to the formation of independent genetic clusters in a context of constant effective population size, allowing them to retain some admixed characteristics from ancient hybridization events as shown by intermediate genetic characteristics of their respective genetic clusters. At the opposite, costal *Q. canariensis* populations form a single cohesive genetic cluster despite the significant geographic distance separating them. In addition, these populations show no sign of ancient admixture, contrary to smaller *Q. canariensis* populations. For sympatric species, species imbalance can increase hybridization rate and directional introgression (Lepais et al., 2009). In a parapatric context, pollen limitation in small isolated population could increase opportunity for hybridization by long distance heterospecific pollen flow. This would explain the intermediate genetic characteristics of the smallest *Q. canariensis* populations. However, the most likely demographic scenario rather suggested that interspecific gene flow events predated population divergence. Thus other factors than purely demographic ones might explain the retention of admixed ancestry in smaller populations such as geographical proximity with heterospecific populations, or increased selection against hybrids in large *Q. canariensis* or declining *Q. faginea* populations. In addition, populations that experienced stable environment for extended periods of time are expected to retain ancestral characters (Petit, Hampe & Cheddadi, 2005), and thus in our context, the footprint of ancient interspecific gene flow.

Irrespectively of the precise complex demographic history, it is clear that small *Q. faginea* and *Q. canariensis* populations have recently evolved in isolation with random genetic drift. As a result, each population constitutes an original and significant contribution to the total genetic diversity of the species and provides a good illustration that rear-edge populations effectively act as reservoirs of evolutionary history (Petit, Hampe & Cheddadi, 2005; Médail & Diadema, 2009), as suggested by the description of subspecies in the region (Aissi et al., 2021). Given their small population size, these populations are vulnerable to rapid decline, as illustrated by our simulations showing demographic decline that did not translate yet into a decrease of genetic diversity. While quantitative estimates of historical effective population sizes (coalescent-based estimates) depend on model specification including mutation rates (Fady & Bozzano, 2021), ratios of ancestral over current effective population size do not depend on mutation rates and provide valuable information about past demographic trends. In contrast, contemporary effective population sizes estimated using linkage disequilibrium is independent on mutation rate (Hoban et al., 2021) and provide a more compelling evaluation of the conservation status of populations (Allendorf & Luikart, 2007; Hoban et al., 2020). Historical declines in some *Q. faginea* populations was not always congruent with small contemporary effective indicating that these populations might not follow the same recent demographic trends. Nevertheless, most of the studied populations have effective population size confidence interval below 500 individuals, a critical level for the long term maintenance of evolutionary potential

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(Jamieson & Allendorf, 2012; Hoban et al., 2020), and estimates of effective population sizes as low as 50 individuals in isolated *Q. faginea* populations suggest short-term risk of detrimental effect of inbreeding depression on population persistence (Jamieson & Allendorf, 2012). If these populations are probably more likely to be in immediate threat posed by demographic stochasticity rather than generic risks, long term retention of evolutionary potential is however paramount for adaptation to environmental changes (Allendorf & Luikart, 2007) and is clearly challenged by low effective population size of the studied rear-edge *Q. faginea* and *Q. canariensis* populations.

Being at the forefront of harmful effect caused by environmental changes, these populations deserve special conservation action because any loss of one of such small population will immediately translate into a net loss of genetic diversity at the species level. This adverse conservation status is recognised for *Q. canariensis* classified as Near Threatened at European level in the IUCN Red List (García Murillo & Harvey-Brown, 2017; Rivers, 2019); even though the species is classified as Data Deficient at global level (Gorener, Harvey-Brown & Barstow, 2017). However, *Q. faginea* is only classified as Least Concern at European and global scales (Harvey-Brown, García Murillo & Buira, 2017; Jerome & Vasquez, 2018; Rivers, 2019). While more abundant at the distribution range scale, *Q. faginea* is located in isolated patches that evolve in isolation, resulting in a mosaic distribution of the genetic diversity with evolutionary potential that should deserve better scrutiny. In the light of these results and because of historical taxonomic confusions with *Q. canariensis* leading to overestimation of *Q. faginea* distribution in Morocco and Algeria, a regional assessment at the African level as well as a global reassessment of conservation status of the species would be of great interest. Appropriate conservation strategy at the infra-specific level should be considered to maintaining the highest number of such local populations (Hampe & Petit, 2005). The complex demographic history uncovered in this local study might explain range-wide taxonomic difficulties with multiple subspecies proposed within *Q. faginea.* However, we cannot provide a clearer *Q. faginea* taxonomic resolution (for instance, the relationship between identified genetics clusters and subspecies) due to lack of reference samples from *Q. faginea* subspecies and availability of molecular and morphological data at individual level. Additional studies at the species distribution range level, including core and marginal populations should provide a better understanding of the history and evolutionary significance of intraspecific genetic diversity within these Mediterranean oak species.

Acknowledgments

We thank Rémy Petit for advices and comments on a previous version of the manuscript. Sequence-based microsatellite genotyping was performed at the Genome Transcriptome Facility of Bordeaux

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(Grants from Investissements d'Avenir, Convention attributive d'aide EquipEx Xyloforest ANR-10-EQPX-16-01). Computer time for the demographic simulations and ABC inferences was provided by the computing facilities MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour.

References

Aguinagalde I, Hampe A, Mohanty A, Martín JP, Duminil J, Petit RJ. 2005. Effects of life-history traits and species distribution on genetic structure at maternally inherited markers in European trees and shrubs. *Journal of Biogeography* 32:329–339. DOI: 10.1111/j.1365-2699.2004.01178.x.

Aissi A, Beghami Y, Heuertz M. 2019. Le chêne faginé (*Quercus faginea*, Fagaceae) en Algérie : potentiel germinatif et variabilité morphologique des glands et des semis. *Plant Ecology and Evolution* 152:437–449. DOI: 10.5091/plecevo.2019.1553.

Aissi A, Beghami Y, Lepais O, Véla E. 2021. Morphological and taxonomic analysis of *Quercus faginea* (Fagaceae) complex in Algeria. *Botany* 99:99–113.

Alberto FJ, Aitken SN, Alía R, González-Martínez SC, Hänninen H, Kremer A, Lefèvre F, Lenormand T, Yeaman S, Whetten R, Savolainen O. 2013. Potential for evolutionary responses to climate change - evidence from tree populations. *Global change biology* 19:1645–61. DOI: 10.1111/gcb.12181.

Alberto F, Niort J, Derory J, Lepais O, Vitalis R, Galop D, Kremer A. 2010. Population differentiation of sessile oak at the altitudinal front of migration in the French Pyrenees. *Molecular Ecology* 19:2626–2639. DOI: 10.1111/j.1365-294X.2010.04631.x.

Allendorf FW, Luikart G. 2007. *Conservation and the genetics of populations*. Blackwell Pub.

De Barba M, Miquel C, Lobréaux S, Quenette PY, Swenson JE, Taberlet P. 2016. High-throughput microsatellite genotyping in ecology: Improved accuracy, efficiency, standardization and success with low-quantity and degraded DNA. *Molecular Ecology Resources* 17:492–507. DOI: 10.1111/1755-0998.12594.

Barthe S, Gugerli F, Barkley NA, Maggia L, Cardi C, Scotti I. 2012. Always look on both sides: phylogenetic information conveyed by simple sequence repeat allele sequences. *PLoS ONE* 7:e40699. DOI: 10.1371/journal.pone.0040699.

Beaumont M a, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–35.

Belkhir K, Borsa P, Raufaste N, Bonhomme F. 2004. GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations.

Bradbury IR, Wringe BF, Watson B, Paterson I, Horne J, Beiko R, Lehnert SJ, Clément M, Anderson EC,

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Jeffery NW, Duffy S, Sylvester E, Robertson M, Bentzen P. 2018. Genotyping-by-sequencing of genome-wide microsatellite loci reveals fine-scale harvest composition in a coastal Atlantic salmon fishery. *Evolutionary Applications* 11:918–930. DOI: 10.1111/eva.12606.

Buschbom J, Yanbaev Y, Degen B. 2011. Efficient long-distance gene flow into an isolated relict oak stand. *Journal of Heredity* 102:464–472. DOI: 10.1093/jhered/esr023.

Chapuis M, Raynal L, Plantamp C, Meynard CN, Blondin L, Marin J, Estoup A. 2020. A young age of subspecific divergence in the desert locust inferred by ABC Random Forest. *Molecular Ecology*. DOI: 10.1111/mec.15663.

Cornuet J-M, Ravigné V, Estoup A. 2010. Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC bioinformatics* 11:401. DOI: 10.1186/1471-2105-11-401.

Curto M, Winter S, Seiter A, Schmid L, Scheicher K, Barthel LMF, Plass J, Meimberg H. 2019. Application of a SSR-GBS marker system on investigation of European Hedgehog species and their hybrid zone dynamics. *Ecology and Evolution* 9:2814–2832. DOI: 10.1002/ece3.4960.

Darby BJ, Erickson SF, Hervey SD, Ellis-Felege SN. 2016. Digital fragment analysis of short tandem repeats by high-throughput amplicon sequencing. *Ecology and Evolution* 6:4502–4512. DOI: 10.1002/ece3.2221.

Dickey AM, Hall PM, Shatters RG, Mckenzie CL. 2013. Evolution and homoplasy at the Bem6 microsatellite locus in three sweetpotato whitefly (*Bemisia tabaci*) cryptic species. *BMC Research Notes* 6:249. DOI: 10.1186/1756-0500-6-249.

Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR. 2013. NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Molecular Ecology Resources* 14:209–214. DOI: 10.1111/1755-0998.12157.

Durand J, Bodenes C, Chancerel E, Frigerio JM, Vendramin G, Sebastiani F, Buonamici A, Gailing O, Koelewijn HP, Villani F, Mattioni C, Cherubini M, Goicoechea P, Herran A, Ikaran Z, Cabane C, Ueno S, Alberto F, Dumoulin PY, Guichoux E, Daruvar A de, Kremer A, Plomion C. 2010. A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. *BMC Genomics* 11. DOI: 10.1186/1471-2164-11-570.

Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4:359–361. DOI: 10.1007/s12686-011-9548-7.

Estoup A, Jarne P, Cornuet J-M. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology* 11:1591–1604. DOI: 10.1046/j.1365-294X.2002.01576.x.

Estoup A, Raynal L, Verdu P, Marin J-M. 2018. Model choice using Approximate Bayesian

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Computation and Random Forests: analyses based on model grouping to make inferences about the genetic history of Pygmy human populations. *Journal de la Société Française de Statistique* 159:167–190.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genetics* 9:e1003905. DOI: 10.1371/journal.pgen.1003905.

Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* 10:564–567. DOI: 10.1111/j.1755-0998.2010.02847.x.

Fady B, Bozzano M. 2021. Effective population size does not make a practical indicator of genetic diversity in forest trees. *Biological Conservation* 253:108904. DOI: 10.1016/J.BIOCON.2020.108904.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–87.

Feliner GN. 2011. Southern European glacial refugia : A tale of tales. *Taxon* 60:365–372.

Feliner GN. 2014. Patterns and processes in plant phylogeography in the Mediterranean Basin . A review. *Perspectives in Plant Ecology , Evolution and Systematics* 16:265–278. DOI: 10.1016/j.ppees.2014.07.002.

Foll M, Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* 180:977–993. DOI: 10.1534/genetics.108.092221.

García Murillo P., Harvey-Brown Y. 2017. *Quercus canariensis*. In: *The IUCN Red List of Threatened Species*. e.T78809256A80570536.

Gómez A, Lunt DH. 2007. Refugia within refugia: Patterns of phylogeographic concordance in the Iberian peninsula. In: Weiss S, Ferrand N eds. *Phylogeography of Southern European Refugia*. Dordrecht: Springer Netherlands, 155–188. DOI: 10.1007/1-4020-4904-8_5.

Gorener V, Harvey-Brown Y, Barstow M. 2017. *Quercus canariensis*. *The IUCN red list of threatened species* e.T7880925. DOI: https://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T78809256A78809271.en.

Goudet J. 1995. FSTAT (Version 1.2): A computer program to calculate F-statistics. *Journal of Heredity* 86:485–486. DOI: 10.1093/oxfordjournals.jhered.a111627.

Hampe A, Jump AS. 2011. Climate relicts: Past, present, future. *Annual Review of Ecology, Evolution, and Systematics* 42:313–333. DOI: 10.1146/annurev-ecolsys-102710-145015.

Hampe A, Petit RJ. 2005. Conserving biodiversity under climate change: the rear edge matters. *Ecology letters* 8:461–7. DOI: 10.1111/j.1461-0248.2005.00739.x.

Hardy OJ, Vekemans X. 2002. spagedi: a versatile computer program to analyse spatial genetic

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

structure at the individual or population levels. *Molecular Ecology Notes* 2:618–620. DOI: 10.1046/j.1471-8286.2002.00305.x.

Harvey-Brown Y, García Murillo PG, Buira A. 2017. *Quercus faginea*. In: *The IUCN Red List of Threatened Species*. e.T78916251A80570540.

Henriques R, von der Heyden S, Matthee CA. 2016. When homoplasy mimics hybridization: a case study of Cape hakes (*Merluccius capensis* and *M. paradoxus*). *PeerJ* 4:e1827. DOI: 10.7717/peerj.1827.

Hewitt GM. 1999. Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society* 68:87–112. DOI: 10.1111/j.1095-8312.1999.tb01160.x.

Hewitt GM. 2001. Speciation, hybrid zones and phylogeography - or seeing genes in space and time. *Molecular Ecology* 10:537–49.

Hey J, Won YJ, Sivasundar A, Nielsen R, Markert JA. 2004. Using nuclear haplotypes with microsatellites to study gene flow between recently separated Cichlid species. *Molecular Ecology* 13:909–919. DOI: 10.1046/j.1365-294X.2003.02031.x.

Hoban S, Bruford M, D'Urban Jackson J, Lopes-Fernandes M, Heuertz M, Hohenlohe PA, Paz-Vinas I, Sjögren-Gulve P, Segelbacher G, Vernesi C, Aitken S, Bertola LD, Bloomer P, Breed M, Rodríguez-Correa H, Funk WC, Grueber CE, Hunter ME, Jaffe R, Liggins L, Mergeay J, Moharrek F, O'Brien D, Ogden R, Palma-Silva C, Pierson J, Ramakrishnan U, Simo-Droissart M, Tani N, Waits L, Laikre L. 2020. Genetic diversity targets and indicators in the CBD post-2020 Global Biodiversity Framework must be improved. *Biological Conservation* 248:108654. DOI: 10.1016/J.BIOCON.2020.108654.

Hoban S, Paz-Vinas I, Aitken S, Bertola LD, Breed MF, Bruford MW, Funk WC, Grueber CE, Heuertz M, Hohenlohe P, Hunter ME, Jaffé R, Fernandes ML, Mergeay J, Moharrek F, O'Brien D, Segelbacher G, Vernesi C, Waits L, Laikre L. 2021. Effective population size remains a suitable, pragmatic indicator of genetic diversity for all species, including forest trees. *Biological Conservation* 253:108906. DOI: 10.1016/J.BIOCON.2020.108906.

Hoogenboom J, de Knijff P, Laros JFJ, de Leeuw RH, van der Gaag KJ, Sijen T. 2016. FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise. *Forensic Science International: Genetics* 27:27–40. DOI: 10.1016/j.fsigen.2016.11.007.

Jamieson IG, Allendorf FW. 2012. How does the 50/500 rule apply to MVPs? *Trends in Ecology & Evolution* 27:578–584. DOI: 10.1016/J.TREE.2012.07.001.

Jerome D, Vasquez F. 2018. *Quercus faginea*. *The IUCN red list of threatened species threatened species* e.T7891625. DOI: https://dx.doi.org/10.2305/IUCN.UK.2018-2.RLTS.T78916251A78916554.en.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Kampfer S, Lexer C, Glössl J, Steinkellner H. 1998. Characterization of (GA)n microsatellite loci from *Quercus robur*. *Hereditas* 129:183–186. DOI: 10.1111/j.1601-5223.1998.00183.x.

Kivelä M, Arnaud-Haond S, Saramäki J. 2015. EDENetworks: A user-friendly software to build and analyse networks in biogeography, ecology and population genetics. *Molecular Ecology Resources* 15:117–122. DOI: 10.1111/1755-0998.12290.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. 2015. Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources* 15:1179–1191. DOI: 10.1111/1755-0998.12387.

Layton KKS, Dempson B, Snelgrove PVR, Duffy SJ, Messmer AM, Paterson IG, Jeffery NW, Kess T, Horne JB, Salisbury SJ, Ruzzante DE, Bentzen P, Côté D, Nugent CM, Ferguson MM, Leong JS, Koop BF, Bradbury IR. 2020. Resolving fine-scale population structure and fishery exploitation using sequenced microsatellites in a northern fish. *Evolutionary Applications*:eva.12922. DOI: 10.1111/eva.12922.

Lepais O, Chancerel E, Boury C, Salin F, Manicki A, Taillebois L, Dutech C, Aissi A, Bacles CFE, Daverat F, Launey S, Guichoux E. 2020. Fast sequence-based microsatellite genotyping development workflow. *PeerJ* 8:e9085. DOI: 10.7717/peerj.9085.

Lepais O, Leger V, Gerber S. 2006. Short note: high throughput microsatellite genotyping in oak species. *Silvae Genetica* 55:238.

Lepais O, Petit RJ, Guichoux E, Lavabre JE, Alberto F, Kremer A, Gerber S. 2009. Species relative abundance and direction of introgression in oaks. *Molecular Ecology* 18:2228–2242.

Leroy T, Roux C, Villate L, Bodénès C, Romiguier J, Paiva JAP, Dossat C, Aury JM, Plomion C, Kremer A. 2017. Extensive recent secondary contacts between four European white oak species. *New Phytologist* 214:865–878. DOI: 10.1111/nph.14413.

Magri D, Fineschi S, Bellarosa R, Buonamici A, Sebastiani F, Schirone B, Simeone MC, Vendramin GG. 2007. The distribution of *Quercus suber* chloroplast haplotypes matches the palaeogeographical history of the western Mediterranean. *Molecular Ecology* 16:5259–66. DOI: 10.1111/j.1365-294X.2007.03587.x.

Marin J-M, Estoup A, Cornuet J-M, Robert CP, Pudlo P, Gautier M. 2015. Reliable ABC model choice via random forests. *Bioinformatics* 32:859–866. DOI: 10.1093/bioinformatics/btv684.

Marin J, Pudlo P, Estoup A, Robert C. 2018. Likelihood-free model choice. In: Sisson S a, Fan Y, Beaumont M eds. *Handbook of Approximate Bayesian Computation*. 153.

Médail F, Diadema K. 2009. Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *Journal of Biogeography* 36:1333–1345. DOI: 10.1111/j.1365-2699.2008.02051.x.

Mir C, Toumi L, Jarne P, Sarda V, Di Giusto F, Lumaret R. 2006. Endemic North African *Quercus afares* Pomel originates from hybridisation between two genetically very distant oak species (*Q. suber*

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

L. and *Q. canariensis* Willd.): evidence from nuclear and cytoplasmic markers. *Heredity* 96:175–84. DOI: 10.1038/sj.hdy.6800782.

Moracho E, Moreno G, Jordano P, Hampe A. 2016. Unusually limited pollen dispersal and connectivity of Pedunculate oak (*Quercus robur*) refugial populations at the species' southern range margin. *Molecular Ecology* 25:3319–3331. DOI: 10.1111/mec.13692.

Mountain JL, Knight A, Jobin M, Gignoux C, Miller A, Lin AA, Underhill PA. 2002. SNPSTRs: Empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome Research* 12:1766–1772. DOI: 10.1101/gr.238602.

Muir G, Lowe AJ, Fleming CC, Vogl C. 2004. High nuclear genetic diversity, high levels of outcrossing and low differentiation among remnant populations of *Quercus petraea* at the margin of its range in Ireland. *Annals of Botany* 93:691–697. DOI: 10.1093/aob/mch096.

Neophytou C, Gärtner SM, Vargas-Gaete R, Michiels H-G. 2015. Genetic variation of Central European oaks: shaped by evolutionary factors and human intervention? *Tree Genetics & Genomes* 11:79. DOI: 10.1007/s11295-015-0905-7.

Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GVN, Underwood EC, D'amico JA, Itoua I, Strand HE, Morrison JC, Loucks CJ, Allnutt TF, Ricketts TH, Kura Y, Lamoreux JF, Wettengel WW, Hedao P, Kassem KR. 2001. Terrestrial Ecoregions of the World: A New Map of Life on EarthA new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience* 51:933–938. DOI: 10.1641/0006-3568(2001)051[0933:teotwa]2.0.co;2.

Payseur BA, Cutter AD. 2006. Integrating patterns of polymorphism at SNPs and STRs. *Trends in Genetics* 22:424–429. DOI: 10.1016/j.tig.2006.06.009.

Petit RJ, Aguinagalde I, de Beaulieu J-L, Bittkau C, Brewer S, Cheddadi R, Ennos R, Fineschi S, Grivet D, Lascoux M, Mohanty A, Müller-Starck G, Demesure-Musch B, Palmé A, Martín JP, Rendell S, Vendramin GG. 2003. Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* 300:1563–5. DOI: 10.1126/science.1083264.

Petit RJ, Brewer S, Bordács S, Burg K, Cheddadi R, Coart E, Cottrell J, Csaikl UM, van Dam B, Deans JD, others. 2002. Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. *Forest Ecology and Management* 156:49–74.

Petit RJ, Hampe A, Cheddadi R. 2005. Climate changes and tree phylogeography in the Mediterranean. *Taxon* 54:877–885.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–59.

Ramakrishnan U, Mountain JL. 2004. Precision and accuracy of divergence time estimates from STR

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and SNPSTR variation. *Molecular Biology and Evolution* 21:1960–1971. DOI: 10.1093/molbev/msh212.

Raynal L, Marin J-M, Pudlo P, Ribatet M, Robert CP, Estoup A. 2016. ABC random forests for Bayesian parameter inference. DOI: 10.24072/pci.evolbiol.100036.

Rivers M. 2019. *European Red List of Trees*. IUCN, International Union for Conservation of Nature. DOI: 10.2305/iucn.ch.2019.erl.1.en.

Rodríguez-Sánchez F, Hampe A, Jordano P, Arroyo J. 2010. Past tree range dynamics in the Iberian Peninsula inferred through phylogeography and palaeodistribution modelling: A review. *Review of Palaeobotany and Palynology* 162:507–521. DOI: 10.1016/J.REVPALBO.2010.03.008.

Šarhanová P, Pfanzelt S, Brandt R, Himmelbach A, Blattner FR. 2018. SSR-seq: Genotyping of microsatellites using next-generation sequencing reveals higher level of polymorphism as compared to traditional fragment size scoring. *Ecology and Evolution* 8:10817–10833. DOI: 10.1002/ece3.4533.

Schmitt T. 2007. Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology* 4:11. DOI: 10.1186/1742-9994-4-11.

Scotti-Saintagne C, Mariette S, Porth I, Goicoechea PG, Barreneche T, Bodénès C, Burg K, Kremer A. 2004. Genome scanning for interspecific differentiation between two closely related oak species [*Quercus robur* L. and *Q. petraea* (Matt.) Liebl.]. *Genetics* 168:1615–26. DOI: 10.1534/genetics.104.026849.

Temunović M, Garnier-Géré P, Morić M, Franjić J, Ivanković M, Bogdan S, Hampe A. 2020. Candidate gene SNP variation in floodplain populations of pedunculate oak (*Quercus robur* L.) near the species' southern range margin: Weak differentiation yet distinct associations with water availability. *Molecular Ecology* 29:2359–2378. DOI: 10.1111/mec.15492.

Vartia S, Villanueva-Cañas JL, Finarelli J, Farrell ED, Collins PC, Hughes GM, Carlsson JEL, Gauthier DT, McGinnity P, Cross TF, FitzGerald RD, Mirimin L, Crispie F, Cotter PD, Carlsson J. 2016. A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. *Royal Society Open Science* 3:150565. DOI: 10.1098/rsos.150565.

Viruel J, Haguenauer A, Juin M, Mirleau F, Bouteiller D, Boudagher-Kharrat M, Ouahmane L, La Malfa S, Médail 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. DOI: 10.1002/aps3.1201.

Wang J. 2016. Individual identification from genetic marker data: developments and accuracy comparisons of methods. *Molecular Ecology Resources* 16:163–175. DOI: 10.1111/1755-0998.12452.

Waples RS, Do C. 2010. Linkage disequilibrium estimates of contemporary Ne using highly variable

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.12.452011; this version posted July 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

genetic markers: A largely untapped resource for applied conservation and evolution. *Evolutionary Applications* 3:244–262. DOI: 10.1111/j.1752-4571.2009.00104.x.