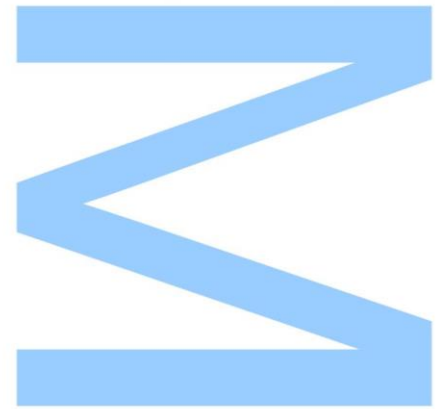




***Quercus canariensis***  
**facing climate change:**  
**Genomics and**  
**niche modelling**  
**as tool for**  
**conservation**



**Tiago Miguel Oliveira Costa**

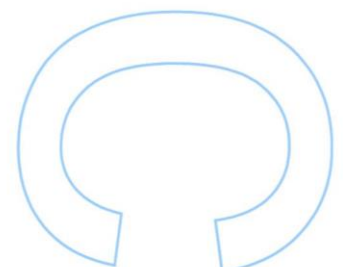
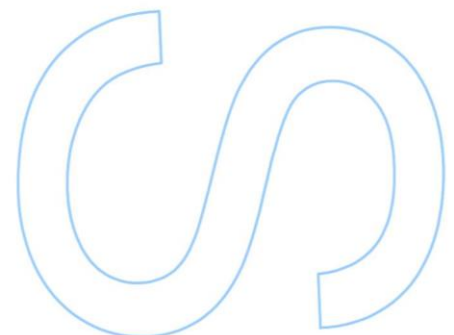
Mestrado em Biodiversidade, Genética e Evolução

Departamento de Biologia

2020

**Orientador**

Antonio Jesús Muñoz Pajares, Investigador, Facultad de Ciencias de la UGR

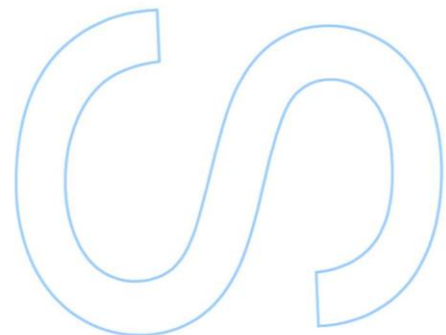
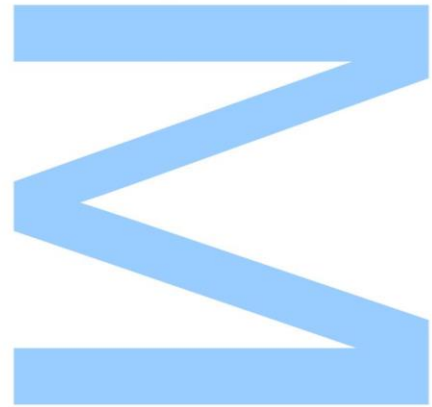




Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, \_\_\_\_/\_\_\_\_/\_\_\_\_



## Resumo

As alterações climáticas influenciadas pelas atividades humanas estão a impactar negativamente as populações de *Quercus canariensis* uma espécie de carvalho bastante esquecida, endémica da Península Ibérica e do Norte de África, causando o declínio de indivíduos ou afetando a distribuição das espécies. Além disso, as mudanças climáticas também impactam a diversidade genética intraespecífica e a estrutura genética das populações. Essas respostas dependem da existência de variação genética para prevenir a extinção da espécie.

Este trabalho tem como objetivo, primeiramente, estudar sobre o status taxonómico de *Q. canariensis*, incluindo relações evolutivas e quantificar a quantidade atual de diversidade genética. Utilizando todo o cpDNA e o DNA nuclear dos ITS, sequenciado por meio da tecnologia NGS (Illumina). Com foco em três dos maiores hotspots de distribuição de *Q. canariensis* na Península Ibérica.

Em segundo lugar, prever como as mudanças climáticas influenciaram a distribuição de *Q. canariensis* no passado e como irá influenciar no futuro. Ao realizar modelagem de nicho ecológico no software MaxEnt, para os cinco períodos: Último Máximo Glacial, Holoceno Médio, Presente, 2050 e 2070. Ao fazer isso, pode ajudar os esforços de conservação priorizando áreas que se preveem se tornarão parte da distribuição das espécies.

A estrutura filogenética foi inferida usando o método de Máxima Verossimilhança para dois conjuntos de dados compostos por ou amostras de *Q. canariensis* e outras espécies de *Quercus* (conjunto de dados cpDNA composto por 54 sequências e o conjunto de dados dos ITS composto por 160 sequências). Os mesmos conjuntos de dados também foram usados para realizar uma Análise de Coordenadas Principais. A diversidade de nucleotídeos foi calculada usando três conjuntos de dados compostos por apenas indivíduos de *Q. canariensis* dos três hotspots da Península Ibérica.

Os modelos climáticos combinaram 169 ocorrências e cinco variáveis climáticas no MaxEnt para inferir a adequação climática para o passado, presente e futuro. Os modelos climáticos foram conduzidos na escala de 5x5 km<sup>2</sup>, cobrindo toda a extensão da espécie.

A análise da filogenia revelou para o conjunto de dados do cpDNA, uma subestrutura intraespecífica entre os três hotspots de *Q. canariensis* e outras espécies de *Quercus* da Península Ibérica. Os indivíduos da Catalunha estão claramente separados dos indivíduos da Andaluzia e de Portugal. Isso é corroborado na análise de diversidade de PCoA e de nucleotídeos de conjuntos de dados do cpDNA. Enquanto indivíduos da Andaluzia e de Portugal estão misturados nos resultados de filogenia, PCoA e diversidade de nucleotídeos.

Apesar do conjunto de dados ITS também identificar com sucesso indivíduos pertencentes a diferentes grupos taxonómicos (nomeadamente, subgéneros e seções), nem o PCoA nem a filogenia forneceram informações conclusivas sobre a estrutura genética dos hotspots de *Q. canariensis* estudados. Isso sugere que regiões adicionais do DNA nuclear devem ser analisadas juntamente com o ITS para obter resultados mais consistentes.

Os resultados do ENM indicam que Cádiz, Tânger e a costa da Argélia são bons candidatos para os esforços de conservação do futuro. Para evitar a extinção de *Q. canariensis* de Monchique, é necessário tomar medidas de conservação agora. No que diz respeito ao hotspot da Catalunha, as áreas de altitude elevada tornar-se-ão mais adequadas no futuro, tornando-se assim as áreas em torno das montanhas dos Pirenéus adequadas para fins de conservação.

Palavras-chave: Mudanças climáticas, modelagem de nicho ecológico, Sequenciamento de Nova Geração, diversidade de nucleotídica, filogenia, taxonomia

# Abstract

Climate changes influenced by human activities are negatively impacting populations of *Quercus canariensis* a largely overlooked species of oak, endemic to the Iberian Peninsula and North Africa, causing the decline of individuals or affecting the distribution of the species. Additionally, climate change also impacts intraspecific genetic diversity, and the genetic structure of natural and populations. These responses depend on the existence of genetic variation to prevent the extinction of the species.

This work aims to firstly study about the taxonomic status of *Q. canariensis*, including evolutionary relationships and quantify the current amount of genetic diversity. Using the whole cpDNA and the ITS nuclear DNA, sequenced by means of NGS technology (Illumina). With focus on the three main of *Q. canariensis* distribution hotspots in the Iberian Peninsula.

The second aim of this study is to predict how climate change have influenced the distribution of *Q. canariensis* in the past and how it will influenced it the future. By performing ecological niche modelling on MaxEnt software, for the five periods: Last Glacial Maximum, Mid Holocene, Present, 2050 and 2070. By doing so can help conservations efforts prioritising areas are predicted to became part of the species distribution.

Phylogenetic structure was inferred using Maximum Likelihood method for two datasets comprised of or samples of *Q. canariensis* and other species of *Quercus* (cpDNA dataset comprised 54 sequences and the ITS dataset comprised 160 sequences). The same datasets were also used to perform a Principal Coordinate Analysis. Nucleotide diversity was calculated using three datasets comprised of only *Q. canariensis* individuals from the three hotspots from the Iberian Peninsula.

Climatic models combined 169 occurrences and five climatic variables in MaxEnt to infer climatic suitability for past, present and future. Climatic models were conduct in a scale of 5x5 km<sup>2</sup>, covering the total range of the species.

Phylogenetic analysis revealed, for the cpDNA dataset, an intraspecific substructure between the three *Q. canariensis* hotspots and other species *Quercus* from the Iberian Peninsula. Catalonia individuals are clearly separated from individuals from Andalusia and Portugal. This is corroborated in the PCoA and nucleotide diversity analysis of cpDNA dataset. While individuals from Andalusia and Portugal are intermingled in both the phylogeny and the, PCoA and show similar values of nucleotide diversity.

Despite the ITS dataset also successfully identified individuals belonging to different taxonomic groups (namely, subgenera and sections), neither the PCoA nor the phylogeny provided conclusive information about the genetic structure of the studied *Q. canariensis* hotspots. This suggest that additional nuclear DNA regions must be analysed together with the ITS to obtain more consistent results.

The ENM results points to Cadiz, Tangier and Algerian coast are good candidates for the future conservations efforts. To prevent the extinction of *Q. canariensis* from Monchique conservation action need to be taken now. Regarding the Catalonia hotspot, higher altitude areas will become more suitable in the future, thus becoming areas around the Pyrenees mountains suitable for conservation purposes.

Keywords: Climate change, ecological niche modelling, next generation sequencing, nucleotide diversity, phylogeny, taxonomy

# Index

1. Introduction.....	1
1.1 - Algerian Oak ( <i>Quercus canariensis</i> Willd.) .....	1
1.2 - Conservation importance.....	2
1.3 - Climate Change.....	3
1.4 - Next Generation Sequencing (NGS).....	4
1.5 - Ecological Niche based models (ENMs).....	5
1.6 – Objectives.....	7
2. Materials and Methods.....	8
2.1 – Genetic Diversity analyses.....	8
2.1.1- Sampling.....	8
2.1.2 – DNA extraction.....	8
2.1.3 - Genomic library construction and Sequencing.....	9
2.1.4 - Whole genome sequencing data analysis.....	9
2.1.5 – Phylogenetic tree analysis.....	11
2.1.6 – Principal Coordinate Analysis (PCoA).....	12
2.1.7 – Nucleotide diversity.....	12
2.2 - Climatic modelling.....	13
3. Results and Discussion.....	15
3.1 – cpDNA assembly and annotation.....	15
3.2. – Genetic Diversity analysis.....	17
3.2.1 - Phylogenetic analysis.....	17
3.2.2– PCoA (cpDNA).....	21
3.2.3 - PCoA (ITS).....	23
3.5.4- Nucleotide Diversity ( $\pi$ ).....	24
3.3 – Inference of distribution patterns through ENM.....	26

4 – Conclusion.....	31
5 – Future work.....	31
References.....	34
Appendix.....	41



# Table Index

Table 1 - Bioclimatic variables used for building ENMs of *Q. canariensis*, with respective units.....14

Table 2 - Results of nucleotide diversity ( $\pi$ ) calculated with “pegas” package in R, for the cpDNA.....25

Table 3 – Results of nucleotide diversity ( $\pi$ ) calculated with PoPoolation, for the cpDNA.....25

Table 4 - Results of nucleotide diversity ( $\pi$ ) calculated with “pegas” package in R, for the ITS.....26

Table 5 - Results of nucleotide diversity ( $\pi$ ) calculated with PoPoolation, for the ITS.....26

Table 6 - Average (standard deviation) percent contribution of each variable to the model replicates.....26

# Figure Index

Fig. 1 - (A) - *Q. canariensis* species distribution (source: IUCN Red List of Endangered Species: <https://www.iucnredlist.org/fr/species/78809256/78809271>); (B) – Specimen of *Q. canariensis*; (C) - Achene (acorn) fruit of *Q. canariensis*; (D) – Leaf from *Q. canariensis*. Map Image from: <https://www.naturalearthdata.com/downloads/10m-raster-data/10m-cross-blend-hypso>.....2

Fig. 2 - Locations of the three hotspots areas in *Q. canariensis* (yellow circles). (Map Image source: <https://www.naturalearthdata.com/downloads/10m-raster-data/10m-cross-blend-hypso>) .....8

Fig. 3 – Schematic representation of the methods used to edit the sequence fragment with the inversion.....11

Fig. 4 – Occurrence records of *Q. canariensis* used to build the ENMs (pre-rarefication). Occurrences represented by yellow dots. (Map Image source: <https://www.naturalearthdata.com/downloads/10m-raster-data/10m-cross-blend-hypso>) .....13

Fig. 5 - - Schematic representation of the chloroplast features, produced in CPGAVAS2. Chloroplast of s102\_14 (161237bp), representative of individual with inversion (left) and *Q. dentata* (161250bp), representative of individuals without inversion. The inversion spans between *rps15* and *ndhF* genes (red arrows). Inverted repeats, long single copy section and short single copy section are marked with “IR”, “SSC” and “LSC” respectively.....16

Fig. 6 – Maximum Likelihood phylogenetic tree for individuals of the *Quercus* cpDNA. Nodes with bootstrap support values (BS). Tree generated from “align02” .....18

Fig. 7 - Maximum Likelihood phylogenetic tree for individuals of the *Quercus* cpDNA. Nodes with bootstrap support values (BS). Tree generated from “align03” .....19

Fig. 8 - Maximum Likelihood phylogenetic tree for individuals of the *Quercus* ITS sequences, BS values higher than 75 marked with “ \* “, (some of clades are collapsed to facilitate visualization).....20

Fig. 9 - Principal coordinates analysis plot. Results of the cpDNA dataset showing the first two principal coordinates that, combined, explain 9% of the observed genetic variation.....22

Fig. 10 - Principal coordinates analysis plot. Results of cpDNA sub-dataset showing the first two principal coordinates that, combined, explain 10% of the observed genetic variation. ....22

Fig. 11 - Principal coordinates analysis plot. Results of ITS dataset showing the first two principal coordinates that, combined, explain 6% of the observed genetic variation.....23

Fig. 12 - Principal coordinates analysis plot. Results of the ITS sub-dataset showing the first two principal coordinates that, combined, explain 15% of the observed genetic variation.....24

Fig. 13 - Predicted current (middle), past (left) and future (right) periods habitat for *Q. canariensis*, using CCSM4 model.....28

Fig. 14 - Predicted current (middle), past (left) and future (right) periods habitat for *Q. canariensis*, using MPI-ESM model.....29

Fig. 15 - Predicted current (middle), past (left) and future (right) periods habitat for *Q. canariensis*, using MIROC-ESM model.....30

## Abbreviation List

AND – Andalusia

BS – Bootstrap

CAT – Catalonia

ENM – Ecological Niche Modelling

ITS – Internal transcribed spacer

LGM – Last Glacial Maximum

MH – Mid Holocene

ML – Maximum Likelihood

NGS – Next Generation Sequencing

PCoA – Principal Coordinate Analysis

POR - Portugal

WGS – Whole genome Sequencing

# 1. Introduction

## 1.1 - Algerian Oak (*Quercus canariensis* Willd.)

*Quercus canariensis* Willd, the Algerian oak or Mirbeck's oak is a plant part of the *Quercus* genus and the *Fagaceae* family, native to southern Portugal, Spain, Tunisia, Algeria, and Morocco (Garcia-Lopez et al., 2005). It is present in Spain is limited to a few small areas in Andalusia and some parts of the Catalan coastal area (Fig. 1), where it is highly hybridized with the other oaks present in the region. In fact, extensive hybridization and introgression is a defining characteristic of the *Quercus* genus (Rushton et al., 1993). Contrastingly, populations in Andalusia are composed of less introgressed individuals forming forests throughout the Algeciras mountains (Parque Natural de los Alcornocales) and in the mountains surrounding Malaga. In Portugal the species appears in Monchique region (Blanca et al., 2000), where hybridization was been showed to occur in this region between *Q. canariensis* and other species of oak trees (Vila-Viçosa et al., 2014).

Other *Quercus* species currently inhabit the Iberian Peninsula. This is the case of *Quercus robur* L., *Quercus petraea* (Matt.) Liebl., *Quercus pyrenaica* Willd, *Quercus pubescens* Willd, *Quercus faginea* Lam., *Quercus ilex* L., *Quercus suber*, *Q. coccifera* L., *Q. lusitanica* Lam., and *Q. rotundifolia* Lam. (Olalde et al., 2002; Vázquez et al., 2000). Several of these species co-occur and naturally hybridize with *Q. canariensis*.

*Q. canariensis* is a tree up to 30m tall with a wide and dense crown, or a sub-bush in unfavourable environments with trunk up to 1.5 m diameter (Arbolapp, 2017). The leaves are 10–15 cm long and 6 to 8 cm broad, with 6 to 12 pairs of shallow lobes, glabrous and green on the upper side, glaucous with hairs along midrib on the underside. Fruit in achene (acorn) yellowish brown, covered in its basal part with a dome, that have ovate-triangular scales (Blanca et al., 2000) (Fig. 1).

*Q. canariensis* characteristically prefers warmer environments and grows in siliceous soils (Pérez-Ramos et al., 2009). The species also prefers fresh and well-drained soils, that have low carbonate materials, mainly in valleys and watercourses in which microclimatic conditions are usually more humid where the overstorey canopy is denser. This makes this species more sensitive to local environmental changes (Urbieta et al., 2008b) caused by climate changes. Consequently, *Q. canariensis* is a vulnerable species with declining populations and at risk of a potential extinction, but despite all of this is still a largely overlooked species with much aspects of its ecology unknown. Currently, *Q. canariensis* is classified globally by the IUCN Red List of Threatened

Species as Data Deficient (IUCN, 2020). This evidences the limited knowledge currently accumulated about this species.

## 1.2 - Conservation importance

Global environment changes influenced by human activities are negatively affecting these forests where *Q. canariensis* can be found. Among the multiple factors

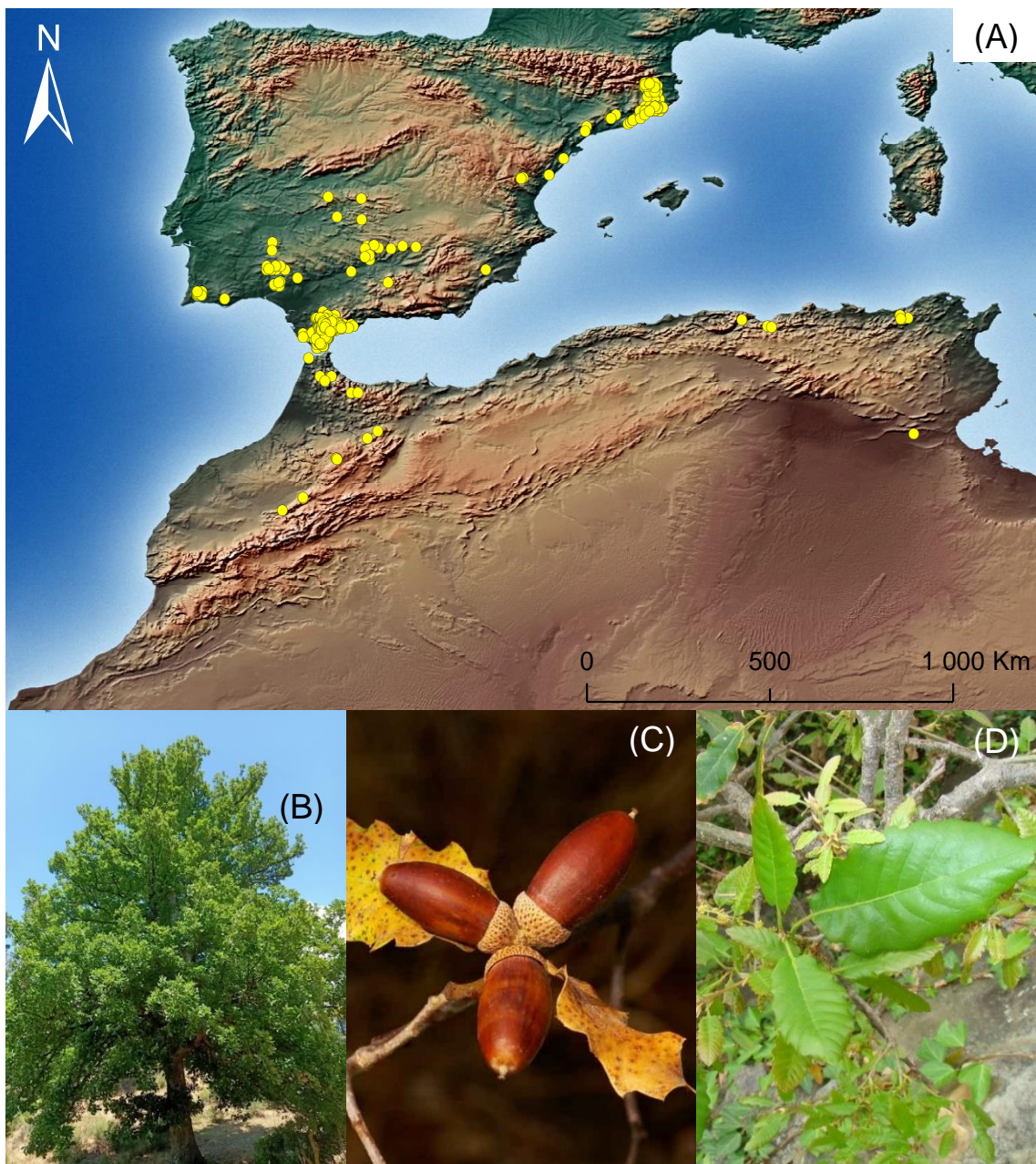


Fig. 1 - (A) - *Q. canariensis* species distribution (source: IUCN Red List of Endangered Species: <https://www.iucnredlist.org/fr/species/78809256/78809271>); (B) - Specimen of *Q. canariensis*; (C) - Achene (acorn) fruit of *Q. canariensis*; (D) - Leaf from *Q. canariensis*. Map Image from: <https://www.naturalearthdata.com/downloads/10m-raster-data/10m-cross-blend-hypso>

driving this decline, three should be highlighted: 1) changes of land usage for agricultural purposes that leads to the loss of suitable habitat areas (i.e. deforestation) and the degradation of the soil quality consequence of those activities (Oldeman, 1992), 2) The introduction of exotic pathogens consequence of increased global trade and 3) Climate change that cause an increase in temperature and changes in precipitation that often leads to droughts. (Adams et al., 2010).

Forests that are largely comprised by individuals of *Q. canariensis* and other species in the *Quercus* genus have ecological and economic importance. They provide a variety of services some like supply (wood, cork and cattle feed), of regulation (carbon sequestration and mitigation of climate change, formation and maintenance of the soil and improvement of air and water quality) and cultural (recreational, landscaping) (Marañón et al., 2016). Trees of the *Quercus* genus are crucial elements for the maintenance of biodiversity and the ecosystem functions they integrate (Marañón et al., 2014). Because socio-economic and ecological factors are key to prioritise conservation actions to certain important species to the ecosystem, *Q. canariensis* is a prime candidate for the conservation effort (Brehan et al., 2010). As such, *Q. canariensis* should be a priority species for any adopted conservation strategy.

### 1.3 - Climate Change

Recent rapid climate change is affecting a wide variety of organisms, and the irregular climate of the past century is already affecting the physiology, distribution, and phenology of some species (Hughes, 2000, McCarty, 2001).

The negative effects of climate change include but are not limited to reduced growth and increases in stress and mortality due to the combined impacts of climate change and climate-driven changes in the ecosystem dynamics, one of the many possible feedback to climate change is through the effects on water and carbon flux in forest ecosystems. (Ayres and Lombardero, 2000, Scholze et al., 2006).

In general, it is agreed that climate has influenced species distributions, often through physiological thresholds of temperature and precipitation tolerance specific to the species (Woodward, 1987). But also, the increase of the occurrences and durations of droughts because of the increasing temperature, which may lead to increased tree

mortality through different but strongly interdependent physiological mechanisms, including carbon starvation and hydraulic failure. This shows that climate change is an emerging factor of tree mortality (Allen et al., 2010, McDowell et al., 2011, Choat et al., 2012).

In the Iberian Peninsula, climate has become drier and warmer in recent decades (Sumner et al., 2003, Drobinski et al., 2020). Consequently, species are expected to follow the shifting climate and change their distributions northward and upward in elevation, within their dispersal capabilities and limited by resource availability (Walther et al., 2002).

Therefore, *Q. canariensis* tree populations must be able to endure changing climate, either by adapting to new conditions through selection on local genetic variation or migrating to new locations with favourable conditions (Aitken et al., 2008). However, this expected distribution shift is a slow process and can be hindered by the dispersal capabilities of the species or by the inaccessibility to suitable areas due to geographical barriers or because human activity.

Meanwhile, changes in climate get more extreme, the risks of extinction are expected to increase for every degree global temperature rises. The consequences of climate change induced extinctions will become gradually more apparent in the future, if we do not intervene (Urban, 2015).

Additionally, climate change will have an impact intraspecific genetic diversity, and will trigger responses from the affecting species, such as changes in the distribution of genetic variants, changes in levels of phenotypical plasticity of individuals and populations when adapting to the new habitat and finally evolutionary adaptation to changing environmental conditions (Sgro` et al., 2011, Pauls et al., 2013). These responses depend on the existence of genetic variation within the populations. Therefore, absence of genetic variation, will increase risk of extinction in wild populations (Spielman et al. 2004).

Thus, there is a need to include evolutionary processes that preserve genetic diversity and adaptive potential in protected area planning and management (Sgro` et al., 2011, Mace and Purvis, 2008). With new emerging genomic tools, like NGS, and an increased understanding of the genetic basis of adaptive responses to environmental change we can seriously integrate evolutionary processes in conservation efforts.



## 1.4 - Next Generation Sequencing (NGS)

The recent arrival of next-generation sequencing (NGS) technologies now allow sequencing of DNA strands with increased capabilities far beyond that of Sanger sequencing (Sanger et al., 1977), millions of DNA bases can be sequenced in one round at a fraction of the cost. NGS techniques have been commercially available since 2005, Solexa sequencing technology was the first technique. Ever since, several different sequencing methods have been developed and continually improved at an increasing rate. Leading to lower costs and improved capabilities of these technologies, although there is still a lot of potential in NGS technologies to be explored. With new fields of study being created, allowing the analysis of a variety of datasets, and finding the answers to questions not possible before.

Although, there is presently many NGS methods commercially available since Solexa, in this work all NGS was done with Illumina dye sequencing. Throughout the sequencing procedure DNA molecules and primers are first attached flow cell and amplified with DNA polymerase so that DNA clusters can be formed. To determine the sequence, four types of reversible terminator bases (RT-bases) are added and non-incorporated nucleotides are washed away. A camera takes images of the fluorescently labelled nucleotides. At that point the dye, along with the terminal 3' blocker, is chemically removed from the DNA before the next cycle to begins.

Before NGS the way to address large-scale genome-level questions was restricted to working on model organisms or close relative species, that are well studied and possessed a wider range of genomic resources (Cronn, 2012). Currently, because of NGS, researchers that need data from large numbers of individuals for a study and those working on non-model organisms are no longer limited to slow and costly gene-by-gene approaches (Cronn, 2012, da Fonseca et al., 2016).

As NGS technologies continue to improve, their scope and application will correspondingly expand within and across scientific disciplines. Plant biology has much to gain from increasing our technological capacity in genomics. In terms of comparative genomics, the increasing number of fully sequenced plant genomes will enable greater understanding of genetic, genomic, developmental, and evolutionary processes.

## 1.5 - Ecological Niche based models (ENMs)

Predicting species' distributions has become an important component of conservation planning in recent years, and a wide variety of modelling techniques have been developed for this purpose (Zurell et al., 2020).

These models generally use associations between environmental variables and known species' occurrence records to find suitable environmental conditions within which populations can be preserved, across a study area as well as to infer them in other time periods, assuming that the niche was the same over time. This has proven valuable for generating biogeographical information that can be applied across a broad range of fields, including conservation biology, ecology, and evolutionary biology (Pearson, 2007).

Ecological models can be divided in two types of models: mechanistic and correlative: Mechanistic models incorporate processes that limit distributions like physiological environmental constraints that effect their distribution and abundance and are strongly tied to the flow of mass and energy. Through such process climate change impacts biodiversity affecting multiple levels of organization such as species, populations, communities, and ecosystems.

Correlative models are widely used to predict the spatial distribution of species and impacts of climate change. They use the statistical association between spatial environmental data and species occurrence data to determine processes that limit the distribution of the species. They have practical advantages over mechanistic modelling methods due to the simplicity, flexibility and availability of the presence data and absence data they use (Kearney et al., 2010, Robertson et al., 2003).

Since mechanistic models often require more time, effort, resources and data to construct and validate, correlative models are preferred if the study does not require a more focused and detailed hypotheses through identification of a limiting process (Kearney and Porter, 2009, Robertson et al., 2003).

Although many species presence data sources exist, like in natural history museums and herbaria, absence data is much more limited, because of poor or lack of sampling and in general it is harder to obtain, especially in remote areas (Ponder et al., 2001).

ENMs gives an estimate of a species' ecological niche in the examined study area. A species' fundamental niche consists of all conditions that allow for its long-term survival, whereas its realized niche is a subset of the fundamental niche that species occupies. Consequently, this represents an approximation of the species' realized niche, in the study area. If the model portrays the species' full niche requirements, areas of predicted presence will generally be larger than the species' realized distribution, because of many possible factors like geographic barriers, biotic interactions and human presence, consequently species rarely occupy all areas that fulfil their niche requirements (Phillips et al., 2006). None the less, those predictions from an ENMs may give a general idea of the species' future distribution under climate change, a species' past distribution to assess evolutionary relationships, or the potential future distribution of an invasive species. Predictions of current habitat and suitable future habitat areas can be useful for management and conservation efforts.

## 1.6 - Objectives

This work aims to: (i) study the taxonomic status of *Q. canariensis*, including evolutionary relationships; (ii) quantify the current amount of genetic diversity within the species; and (iii) to evaluate whether the suitable habitat of *Q. canariensis* is expected to decline during the next decades as a consequence of climate change. The present study is focused in the three *Q. canariensis* distribution hotspots in the Iberian Peninsula (hereafter named Monchique, Andalusia and Catalonia). Genomic analyses are performed using Next Generation Sequencing technology and niche modelling includes climatic data from the Last Glacial Maximum to 2070.

By completing these objectives, we expected to know more about the impact that climate change had on the species and help in its conservation efforts by trying to predict how climate change is going to affect the distribution of the species in the future. Therefore, prioritising areas that in the future may be part of the distribution of species for the conservation efforts.

## 2. Materials and Methods

### 2.1 – Genomic analyses

#### 2.1.1- Sampling

Samples from 29 specimens of *Q. canariensis* and 1 each from *Q. faginea*, *Q. petraea* and *Q. pyrenaica* were obtained for genomic analyses (Appendix 1). The samples consisted of dry leaf tissues stored in Herbário (PO) do Museu de História Natural e da Ciência da Universidade do Porto (MHNC-UP). These samples were collected across the Iberian Peninsula, in three hotspots Monchique (Portugal), Andalusia (Spain) and Catalonia (Spain) by CIBIO researcher Carlos Vila-Viçosa and collaborators during various fieldwork excursions (Fig. 2).



Fig. 2 - Locations of the three hotspots areas in *Q. canariensis* (yellow circles). (Map Image source: <https://www.naturearthdata.com/downloads/10m-raster-data/10m-cross-blend-hypso>)

## 2.1.2 – DNA extraction

The protocol used for DNA extraction (Doyle and Doyle, 1987) was based on another one developed in 1984 for high molecular weight DNA isolation from small amount fresh leaf tissue of single plant using CTAB.

In this case the methodology was modified in the first step from the original protocol, namely 0.5 g to 1 g samples of young leaf tissues were, instead, grounded to a fine powder in liquid nitrogen. That powder was then placed in 1.5mL microtubes containing 700 µL 2% CTAB extraction buffer (2% CTAB, 1.4 M NaCl, 0.2% 2-mercaptoethanol, 20 mM EDTA, 100 mM Tris-HCl, pH 8.0). After this, all the other steps from the original protocol have been followed (Appendix 2).

Extraction success was then verified by electrophoresis on a 2% agarose gel died with GelRed™ once all DNA was extracted. Then we check the purity of the DNA with nanodrop and quantified the DNA using Quant-it PicoGreen from ThermoFisher. Samples exceeding the recommended concentration values for the library construction process were diluted.

## 2.1.3 - Genomic library construction and Sequencing

After successful DNA extraction we proceeded to library construction, required by the Illumina dye sequencing. This is done by transposases that randomly cut the DNA into 500 bp fragments and simultaneously add adaptors. For this process the protocol for use with NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® by New England Biolabs Inc. was followed (Appendix 3).

We created 14 individual libraries and three DNA pools, corresponding to the three hotspot regions, previously mention. This consisted in seven to eight samples of extracted DNA, from individuals originated from the same regions, being combined in equal parts in a tube and send along with the other samples for sequencing.

The DNA libraries created were sent to a commercial genomics company (NOVOGENE) to be sequenced in HiSeqX sequencing system.

#### 2.1.4 - Whole genome sequencing data analysis

Afterwards the raw data produced from the whole genome sequencing (WGS) process we use various software to extract the target sequences from the whole raw data. In this case taxonomic and genetic diversity analyses were done based on the complete chloroplast DNA (cpDNA) sequences and on Internal Transcribed Spacers (ITS) DNA sequences from all individual samples and from the three regional DNA pools sequenced.

To address the task of assembling the entire cpDNA we used NOVOPlasty, a *de novo* organelle assembler and heteroplasmy caller software (Dierckxsens et al., 2017). NOVOPlasty uses seed-and-extend algorithm to assemble organelle genomes from WGS data, starting from a related or distant single seed sequence.

For the ITS sequences we used a pipeline to extract them from WGS data. Firstly, Burrow-Wheeler Aligner, or bwa, version 0.7.17 (Li, 2013), to align the reads to a reference ITS sequence. The reference sequence was previously downloaded from GenBank (*Quercus petraea* subsp. *Iberica* accession number: KM588211). Then SAMtools version 1.10.2 software (Li et al., 2009) was used to generate a BAM file containing all the mapped reads. The BAM files were then imported into Geneious Prime 2020.2 (<https://www.geneious.com>) to generate a majority consensus sequence.

Because NOVOPlasty could not reconstruct entire cpDNA molecules in several individuals, we applied the same procedure described for ITS using the best quality cpDNA obtained (s102\_14) as a reference. These sequences were used to conduct the genetic analysis. CpDNA were visualized using CPGAVAS2 (Shi et al., 2019).

Since cpDNA is a circular molecule to perform sequence alignment we must define a position to act as a sort of starting point for the alignment. This was done for every cpDNA sequence used in the analysis, using a custom script built in R version 4.0.2 (R Core Team 2020). The cpDNA and the ITS sequences were viewed on BioEdit version 7.2 software (Hall, 1999). All sequence alignments were done with MAFFT version 7.471 software (Kato et al., 2002).

Sequenced samples we added sequences of both entire cpDNA and ITS from all the other species of *Quercus*, available in GenBank to establish a comparison to our samples. After comparing, our cpDNA sequences and the ones from GenBank we discovered an inversion in all Iberian species from our samples. Because the inversion

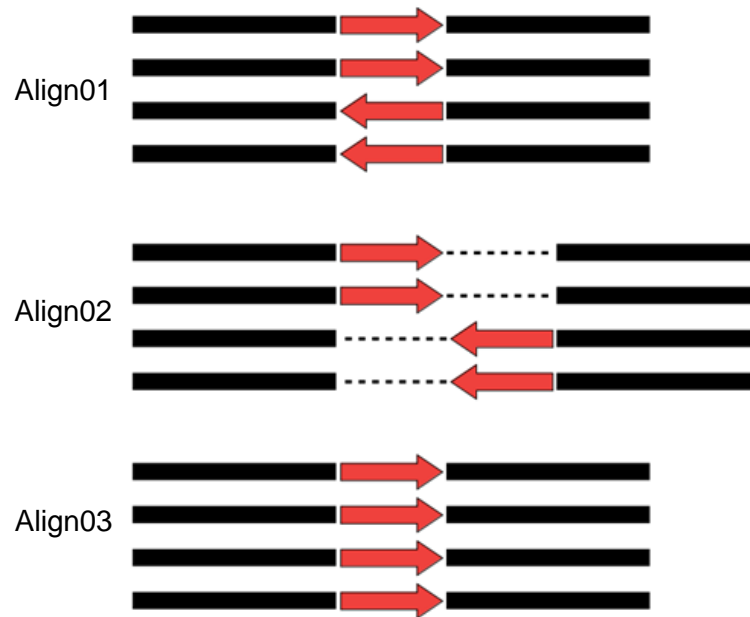


Fig. 3 – Schematic representation of the methods used to edit the sequence fragment with the inversion

created a region of artificially low similarity and because classic sequence-based phylogenetic methods are restricted to substitution events (Ashkenazy et al., 2014) we used BioEdit to edit the initial alignment (referred as “align01”) following two different methods to create two new alignments, one for each method (Fig. 3).

For the first method, the fragment of the sequences with the inversion and it is equivalent in the sequences without the inversion were separated by adding gaps to the alignment (referred as “align02”). For the second method we estimated the reverse-complement of the region with inversion, to match all the other sequences in the alignment (referred as “align03”, Fig. 3).

### 2.1.5 – Phylogenetic tree analysis

Phylogenetic tree construction was done for the cpDNA and the ITS datasets separately. The cpDNA dataset consists in 54 sequences, including our samples and sequences from other species of *Quercus* downloaded from GenBank. The ITS dataset is made up 160 sequence from different species of *Quercus*, downloaded from GenBank and our samples. Both datasets were tested for most suitable substitution model in CIPRES Science Gateway (Miller et al., 2010) with ModelTest-NG version 0.1.5 (Darriba et al., 2020), using the Bayesian Information Criterion (BIC), Akaike Information Criterion

(AIC) and Akaike Information Criterion corrected (AICc). Results show the best model for the cpDNA and the ITS datasets are TPM1uf + G + I and TIM3 + G + I, respectively.

We used the software RAxML-NG version 8.2.12 (Kozlov et al., 2019) in CIPRES Science Gateway to perform ML phylogenetic inference for the *Quercus* sequences. Analyses were run with the combined tree search and bootstrapping analysis function with a maximum of 1000 replicates and the selected substitution models. The trees were rooted through an outgroup of similar sequences from *Castanea seguinii* (accession number: MH998383). The resulting phylogenetic tree was visualized and edited with FigTree version 1.4.

### 2.1.6 – Principal Coordinate Analysis (PCoA)

Principal coordinates analysis is an isometric multi-dimensional scaling method that allows the calculation of geometric coordinates of data objects in an n-dimensional space given a matrix of similarities or distances between those objects (Gower, 2014).

This was done in R version 4.0.2 using the “ape” package (Paradis, E and Schliep, K, 2019). First we import the alignment data into R with the function “read.dna”. Next, we constructed a distance matrix using the “dist.dna” function for the various imported alignments data. The model chosen to do this was “logdet” (Lockhart et al., 1994), because is the most similar the model used in the phylogenetic analysis.

Subsequently, we used the function “pcoa” to do the analysis for the selected distant matrix. The results were plotted and edited with Microsoft Excel. This analysis was done for the ITS dataset and for “align03” (the cpDNA dataset edited to revert the inversion in our Iberian species).

### 2.1.7 – Nucleotide diversity

The nucleotide diversity is the total number of differences between pairs of sequences divided by the number of comparisons.

Nucleotide diversity was calculated in R version 4.0.2 with the “pegas” package (Paradis, E, 2010), using the function “nuc.div”. The variance of the estimated diversity uses formula (10.9) from Nei (1987). To increase sample size, we grouped together



individuals and pools of *Q. canariensis* by region of origin and then we calculated nucleotide diversity for these groups. This was done for the cpDNA and the ITS datasets.

Paralleled to the analysis in R, we used raw data for each of the regional pools and ran an analysis to calculate Tajima's  $\pi$ . The raw data was put through PoPoolation (Kofler et al., 2011) pipeline. This pipeline analyses pooled next generation sequencing data by using the sliding window method. We used a window size and step size of 160000 nucleotides for the cpDNA datasets to calculate the total value of nucleotide diversity. For the ITS we tried to do the same method with a window and step size of 500 nucleotides, but the values couldn't be calculated. Consequently, we decided to use smaller windows and step size of 60 and 30 nucleotides, respectively. Afterwards we calculated a mean value of nucleotide diversity with the results.

## 2.2 - Climatic modelling

For this work, we also aimed to determine suitable habitat areas for *Q. canariensis* for the future as well as to determine its possible past habitat areas. To do this we used modelling using climatic variables projected for the future and for past times. For this a dataset with a total of 169 register occurrences at 1 x 1 km (WGS 1984 datum)

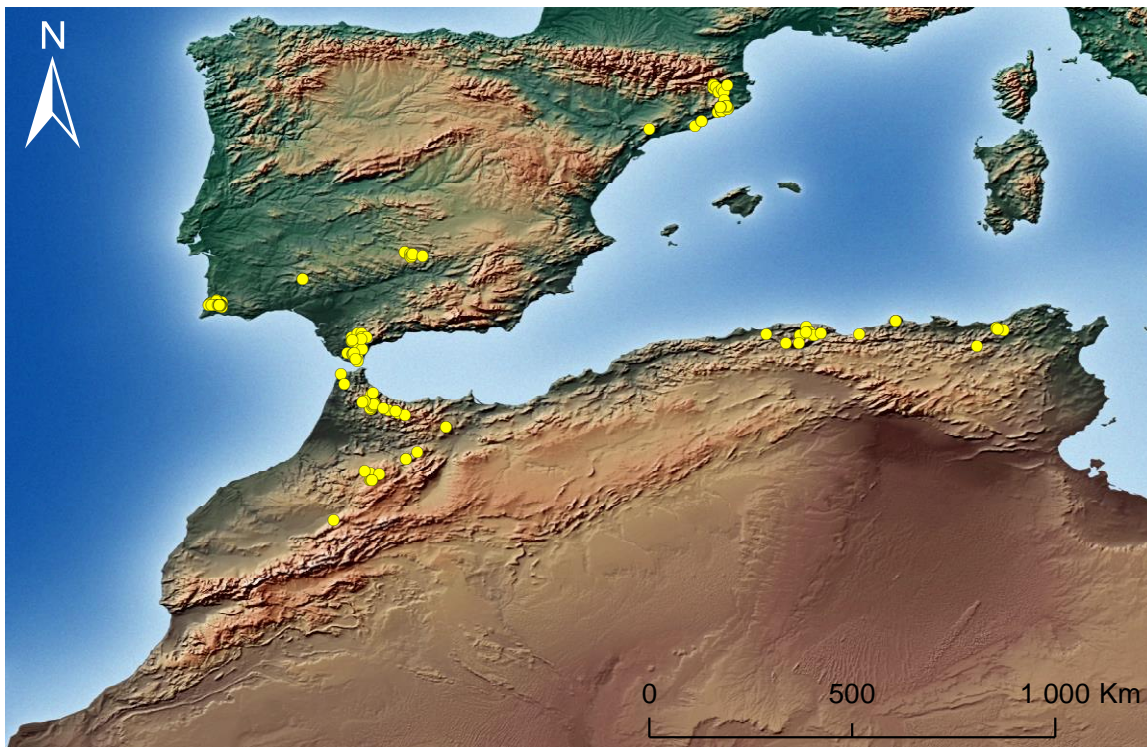


Fig. 4 – Occurrence records of *Q. canariensis* used to build the ENMs (pre-rarefaction). Occurrences represented by yellow dots. (Map Image source: <https://www.naturalearthdata.com/downloads/10m-raster-data/10m-cross-blend-hypso>)

was used for the distributional range *Q. canariensis* covering all Iberian Peninsula and part northwest Africa (including of Morocco, Algeria, and Tunisia (Fig. 4).

This dataset was acquired from CIBIO researcher Carlos Vila-Viçosa, and encompassed data collected from fieldwork expeditions, GBIF - Global Biodiversity Information Facility (<https://www.gbif.org/>), Flora On (<https://flora-on.pt/>), SIVIM: Sistema de información de la Vegetación Iberica y Macaronésica (<http://www.sivim.info/sivi/>), scientific articles and from herbaria collections (Muséum national d'Histoire naturelle, Paris (France), Herbario HSS del Instituto de Investigaciones Agrarias Finca La Orden-Valdesequera (CICYTEX).

Because our dataset was obtained using different sources, occurrence overrepresentation was observed. Sampling bias could affect the quality and reliability of the models (Kramer-Schadt et al, 2013) as such species occurrences were subject to a process of spatial rarefication. The rarefication process was done with SDM toolbox version 2.4 (Brown et al., 2017) in ArcGIS 10.5 (ESRI, 2016). This tool removes spatially autocorrelated occurrence points by reducing multiple occurrence records to a single record occurrence within the specified distance. The purpose is to avoid models to overfit and model performance values to be inflated. This left us with a new dataset of 55 occurrence records.

Nineteen bioclimatic variables were downloaded from WorldClim version 1.4 (Hijmans et al., 2005) at 5km<sup>2</sup> of resolution (2.5 minutes) for present conditions. Then, the correlation between variables was calculated, using Pearson's *r* method, in ArcGIS with the tool "Remove highly correlated variables", part of SDMtoolbox version 2.4. The correlation threshold was set to R=0.7. The results, presented in Table 1, show the low correlated variables (R<0.7) selected for use and with potential biological meaning to the species.

Table 1 - Bioclimatic variables used for building ENMs of *Q. canariensis*, with respective units

Bioclimatic variables	Units
BIO3 = Isothermality (BIO2/BIO7) (×100)	%
BIO6 = Min Temperature of Coldest Month	°C
BIO8 = Mean Temperature of Wettest Quarter	°C
BIO12 = Annual Precipitation	Mm
BIO15 = Precipitation Seasonality (Coefficient of Variation)	%

Climatic variables were used for five time periods: Last Glacial Maximum (LGM) about 22,000 years ago and Mid Holocene (MH) about 6,000 years ago, present (from 1960 to 1990), 2050 and 2070. Distinct Global Circulation Models (GCMs) were used to simulate the data for the paleoclimatic future variables. For LGM, MH, 2050 and 2070 three different GCMs were used: CCSM4 (Community Climate System Model, version 4, Gent et al., 2011), MIROC-ESM (Model for Interdisciplinary Research on Climate, Watanabe et al., 2011) and MPI-ESM (Max-Planck-Institute Earth System Model, Giorgetta et al., 2013).

All the variables were edited in ArcGIS 10.5 (ESRI, 2016). Initially they were clipped to the study area and converted to ASCII (.asc) format using the “Extract by Mask (Folder)” in the SDMtoolbox and them afterwards, the raster files were exported in ascii format to later use in MaxEnt version 3.4.1 (Phillips et al., 2006) for the construction of models.

Model calibration was performed with R package “ENMeval” (Muscarella et al., 2014) in R version 4.0.2 by doing modelling trials with different values for the regularization parameter and using distinct feature combinations. Result show the best evaluated model has a regularization parameter of 5.2 and linear, quadratic, product, threshold, and hinge features. Afterwards, all models were constructed using this the regulation parameter and features previously mentioned.

All models were created by using the Maximum Entropy approach in MaxEnt version 3.4.1. We selected maximum entropy modelling because it uses presence-only data and performs well with incomplete data and requires small sample size and allows gaps.

Firstly, a model was created using present the five selected bioclimatic variables, 10 replicates with bootstrap, random seed and 20% test percentage. Individual evaluation of model fit was done through the area under-the-curve (AUC) of the receiver-operating-characteristics (ROC) plot (Fielding and Bell, 1997, Merow et al., 2013). Posteriorly, other models were projected to future and past conditions. Other options selected for the modelling process on MaxEnt was “Do clamping” to reduce the probability of inflation of the predictable suitable areas, this takes variables outside the training range as if they were at the limit. These makes the values of the projected areas stay within the range of values of the study area (Phillips, S. J., 2005, Radosavljevic and Anderson, 2014). Mean predictions were built for each time in the ArcGIS version 10.5.

## 3. Results and Discussion

### 3.1 – cpDNA assembly and annotation

The *de novo* whole chloroplast genomes assembly of *Q. canariensis* samples, produced circular DNA molecules between 161.181 bp and 161.237 bp long, within the expected range for the *Quercus* genus (observed range in the GenBank *Quercus* complete cpDNA 160.415- 161.366). The chloroplast was viewed using CPGAVAS2 (Fig.5). Along with *Quercus dentata*, a species endemic to China, that represents the samples used without the inversion the cpDNA sequence.

The chloroplast genome exhibited the typical structure that many chloroplasts have, containing two inverted repeats (IRs), which separate a long single copy section (LSC) from a short single copy section (SSC). In our case the inversion was the entire SSC section of chloroplast (marked on Fig. 5 with red arrows).

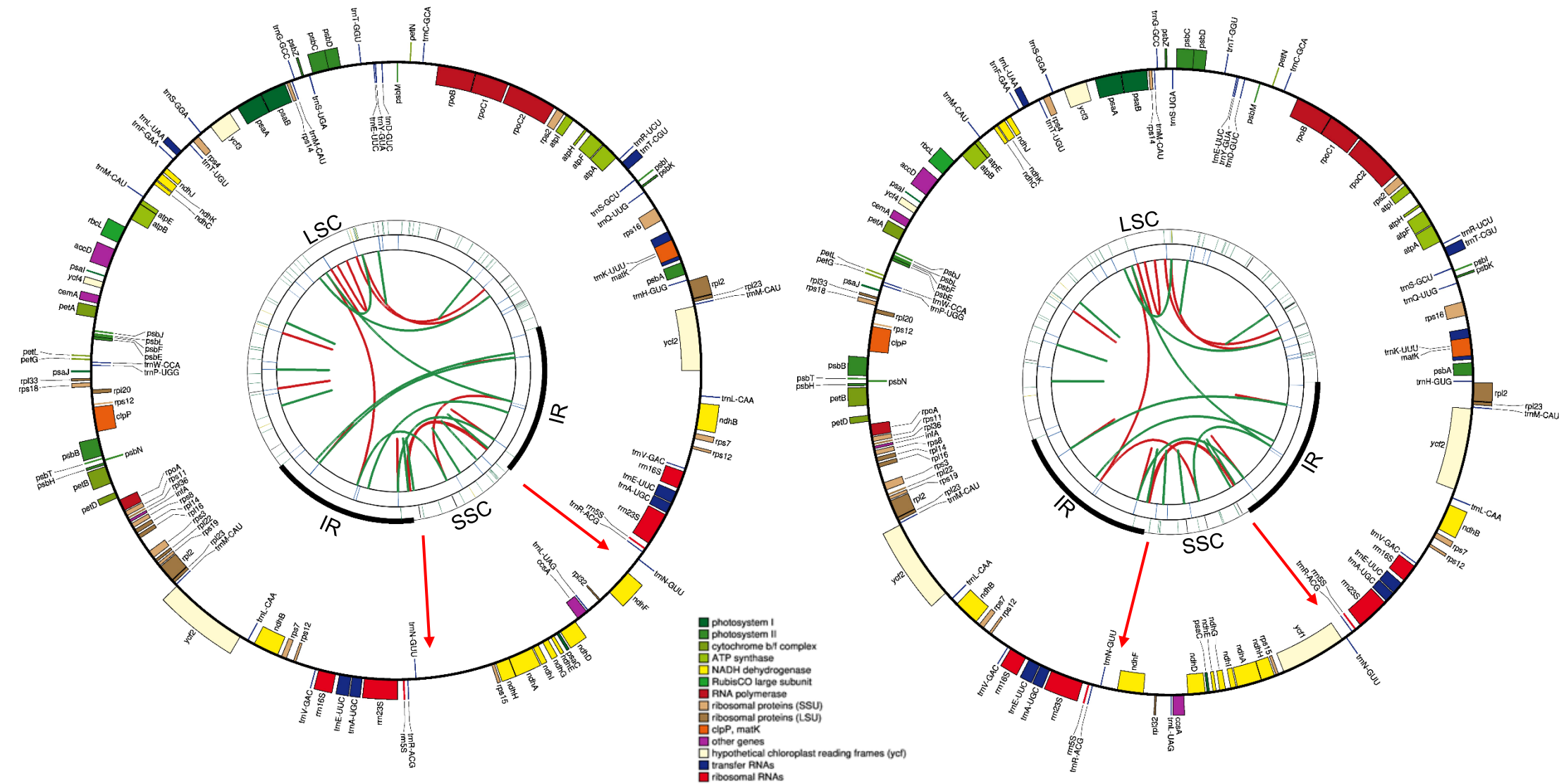


Fig. 5 - Schematic representation of the chloroplast features, produced in CPGAVAS2. Chloroplast of s102\_14 (161237bp), representative of individual with inversion (left) and *Q. dentata* (161250bp), representative of individuals without inversion. The inversion spans between *rps15* and *ndhF* genes (red arrows). Inverted repeats, long single copy section and short single copy section are marked with "IR", "SSC" and "LSC" respectively.

## 3.2. – Genetic Diversity analysis

### 3.2.1 - Phylogenetic analysis

We performed ML phylogenies using two edited alignments: one assuming lack of homology between sequences with and without inversion (“align02”) and the other estimating the reverse-complement of the inverted region (“align03”). The sequences that had the inversion are represented with green branches in Fig.6 and 7. Both trees have similar topology. Both trees have high bootstrap support (BS>90) for nodes in the phylogeny that separate species from different subgenus and Sections

In both trees there is an obvious decrease of BS values within our samples, since they are individuals of the same species, excluding sample s32A (*Q. pyrenaica*), s32H (*Q. petraea*) and s109\_9 (*Q. faginea*), and also samples from individuals from other European *Quercus* species this may cause decrease of the signal in the dataset (Soltis, P. S., and Soltis, D. E. 2003). However, the support BS values of the branches that represent the three hotspots of *Q. canariensis* have all values above 50.

Intraspecific substructure can be observed in both trees, the samples from each of the tree *Q. canariensis* hotspots are all grouped together (green) and with individual from other species that are also from the Iberian Peninsula. Individuals of *Q. canariensis* from Catalonia are clearly separated in a different branch from the individuals from Andalusia and Portugal. This is also supported in the PCoA and nucleotide diversity analysis of cpDNA (see below). Contrastingly, individuals from Andalusia and Portugal are intermingled in both the phylogeny and the PCoA (Fig. 10 and 12).

On broader side our individuals that are from the Iberian Peninsula are separated from other individuals that are from Asia (red) and from North America (blue) with high support values (BS>90) in both trees. Although we have some discrepancies in both trees, *Q. serrata* is grouped together with species from the *Cerris* subgenus instead of being with in the *Quercus* section (Asia), like *Q. serrata* var. *brevipetioleata* and *Q. spinosa* should be in the *Ilex* section.

The results obtained with the ITS dataset showed to be very different from the results obtained with the chloroplast dataset. ITS phylogenetic tree had very low resolution and BS support values. Because of the high entropy, it was discarded from the results (Fig.6).

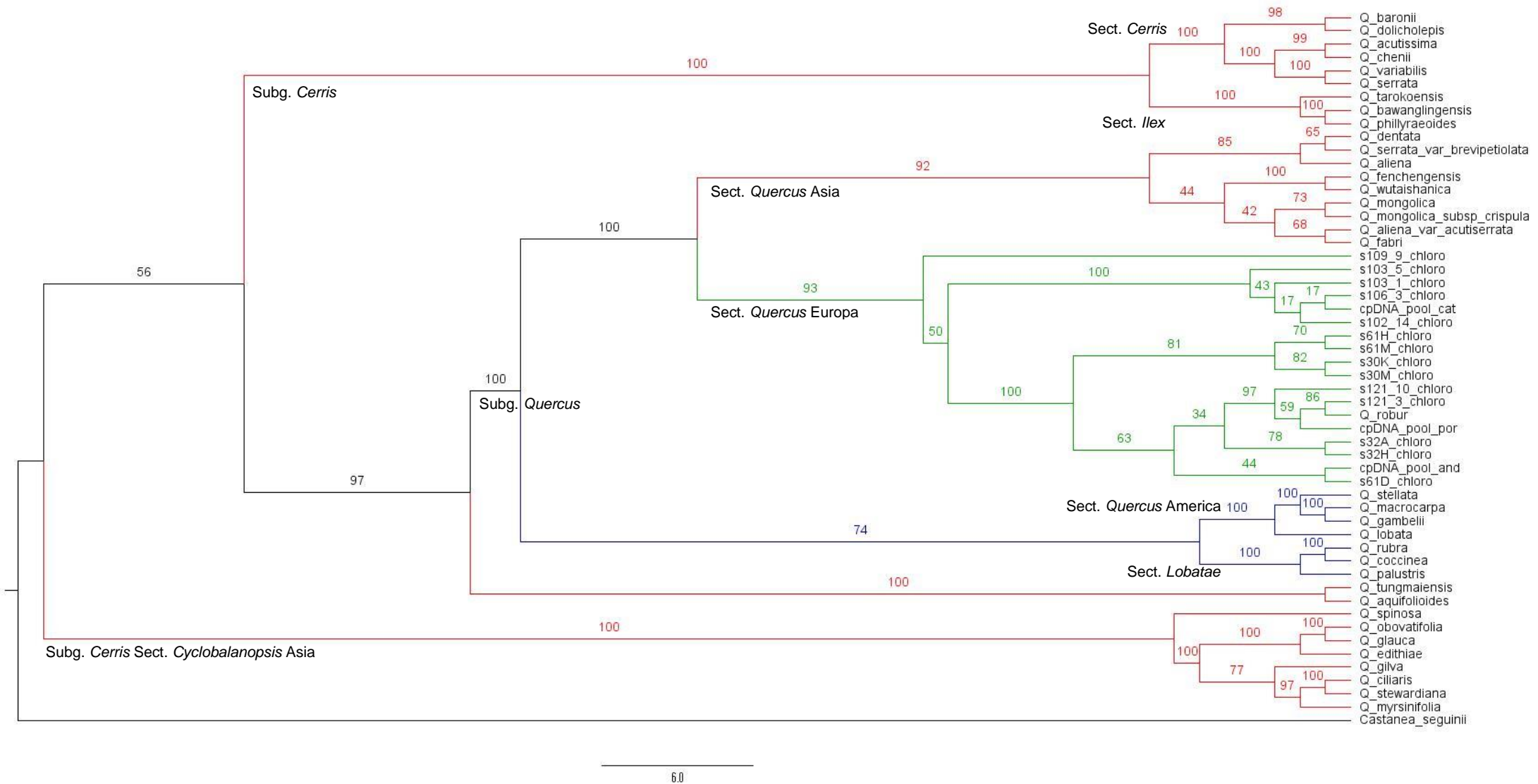


Fig. 6 – Maximum Likelihood phylogenetic tree for individuals of the *Quercus* cpDNA. Nodes with bootstrap support values (BS). Tree generated from “align02”.

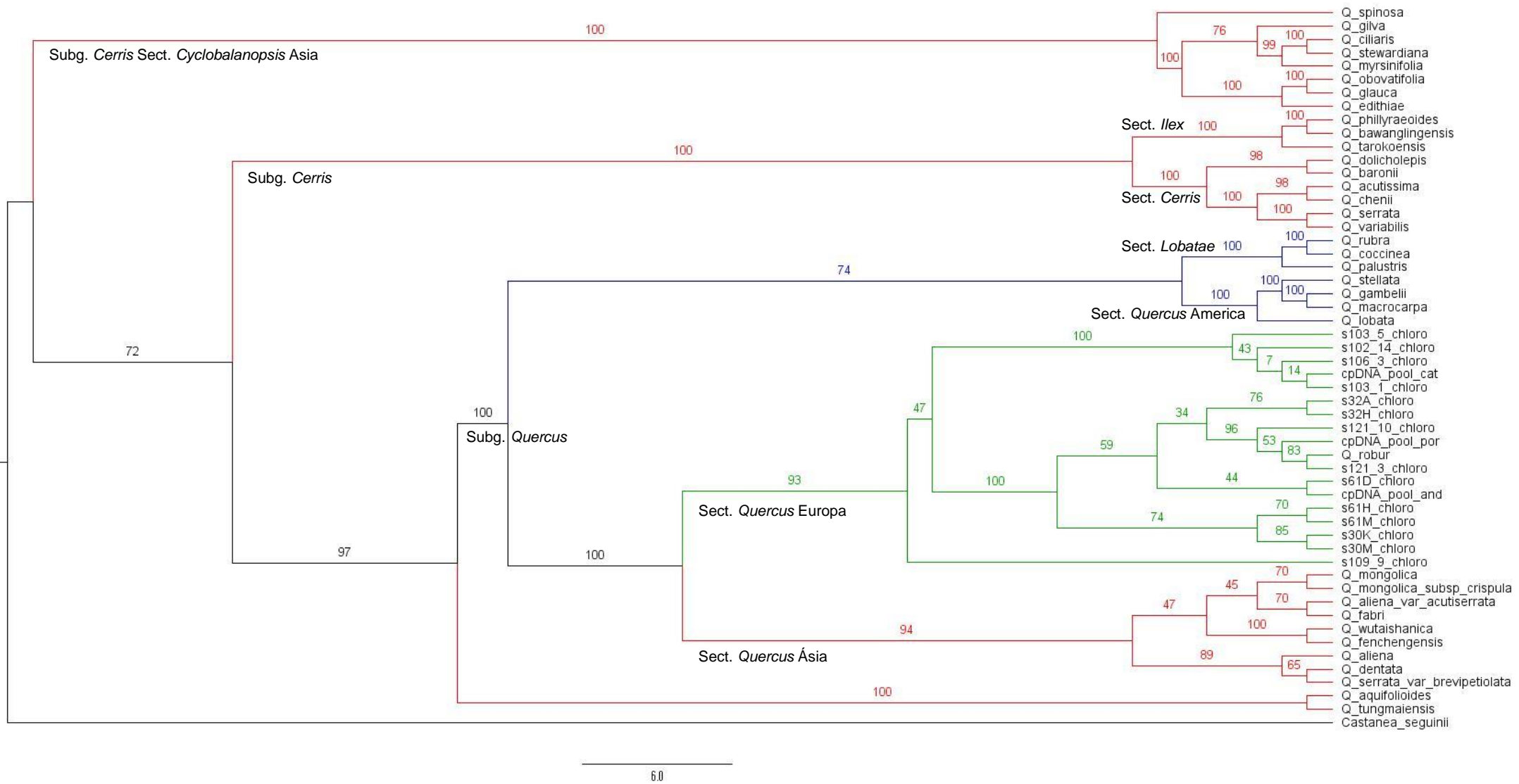


Fig. 7 - Maximum Likelihood phylogenetic tree for individuals of the *Quercus* cpDNA. Nodes with bootstrap support values (BS). Tree generated from "align03".



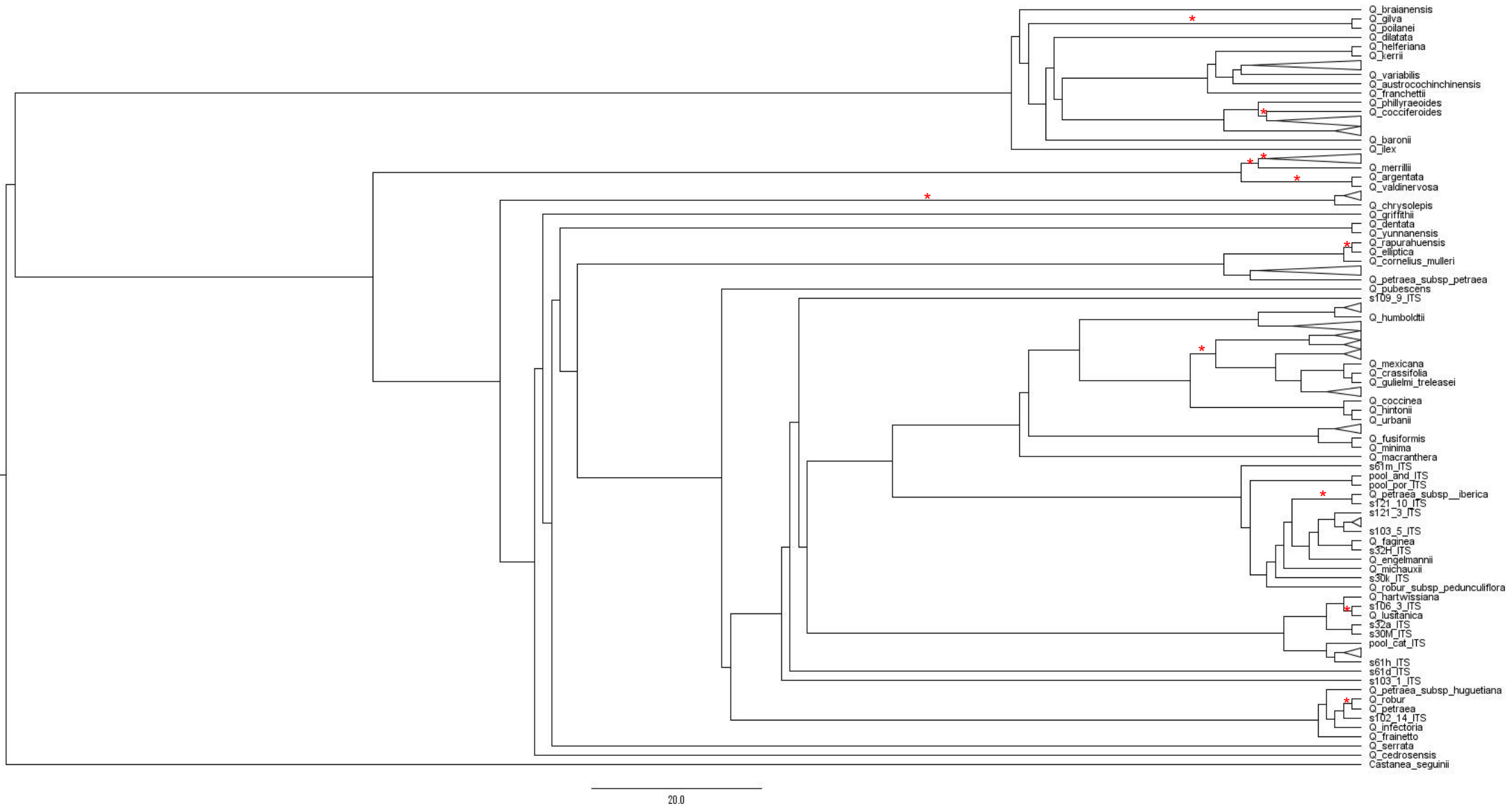


Fig. 8 - Maximum Likelihood phylogenetic tree for individuals of the *Quercus* ITS sequences, BS values higher than 75 marked with “\*”, (some of clades are collapsed to facilitate visualization)

### 3.2.2– PCoA (cpDNA)

The main results of the cpDNA PCoA is presented in Fig. 9 and 10. This analysis was done with “align03”. Points in the graph represent individuals in the alignment. The PCoA groups individuals by region and their respective subgenus and section of their distribution, forming tight groups of points.

The individuals from Asia case we see more than one group of points, this reflects the separation their subgenus and section. The closest to our samples (green) are individual from the *Quercus* section of *Quercus* subgenus (red in Fig. 9). When comparing with the phylogenetic tree (Fig. 7) we also see a *Quercus* Sect. clade from Asia that are much closer to the European species that the other species from Asia. And to the right, the blue group individuals from Asia are part of the *Cerris and Ilex* section of the *Cerris* subgenus, while the purple group are part of *Cyclobalanopsis* section of the same subgenus (Fig. 5).

To increase our samples resolution, we made a second PCoA analysis with a sub-dataset (black circle in Fig.5), that includes the group of individuals where our samples are part of. In Fig. 10 it is possible see our samples and individuals belonging to the *Quercus* section from Europe and Asia more in depth.

The results of new analysis (Fig.10) show that the individuals from Catalonia are the most different of the three. Furthermore, the individuals from Portugal and Andalusia have little to no differences. Also, in this graph we can see distance between the Asian and European individuals in the *Quercus* section of the *Quercus* Subgenus. Regarding this the individuals there is a clear distinction from individuals from Asia and from Europe.

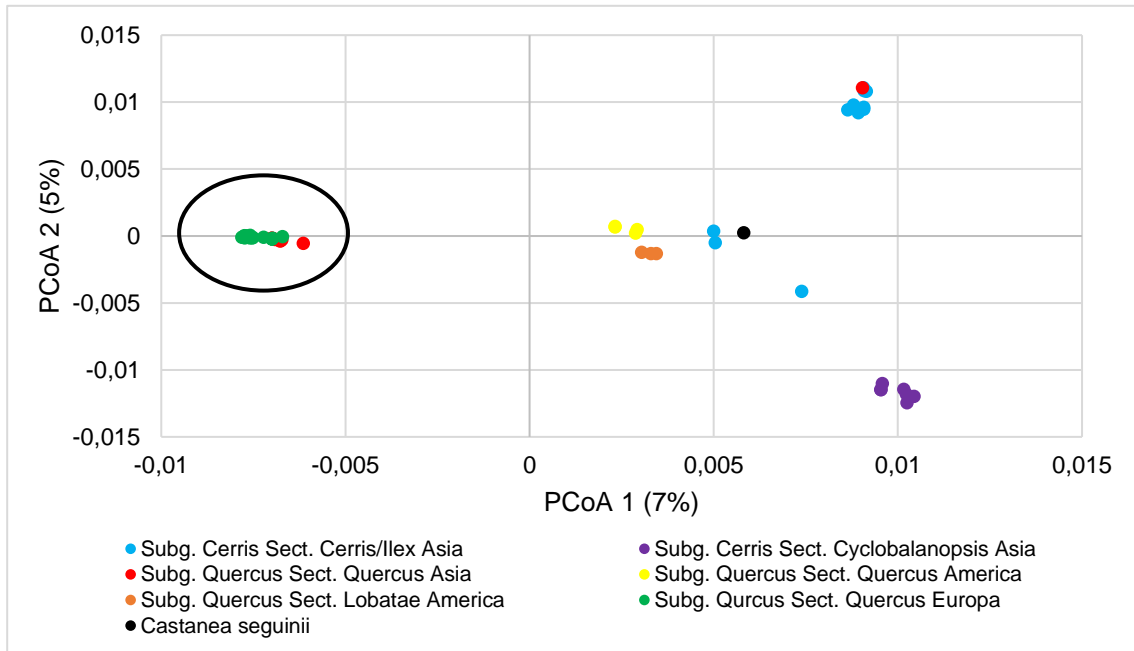


Fig. 9 - Principal coordinates analysis plot. Results of the cpDNA dataset showing the first two principal coordinates that, combined, explain 9% of the observed genetic variation.

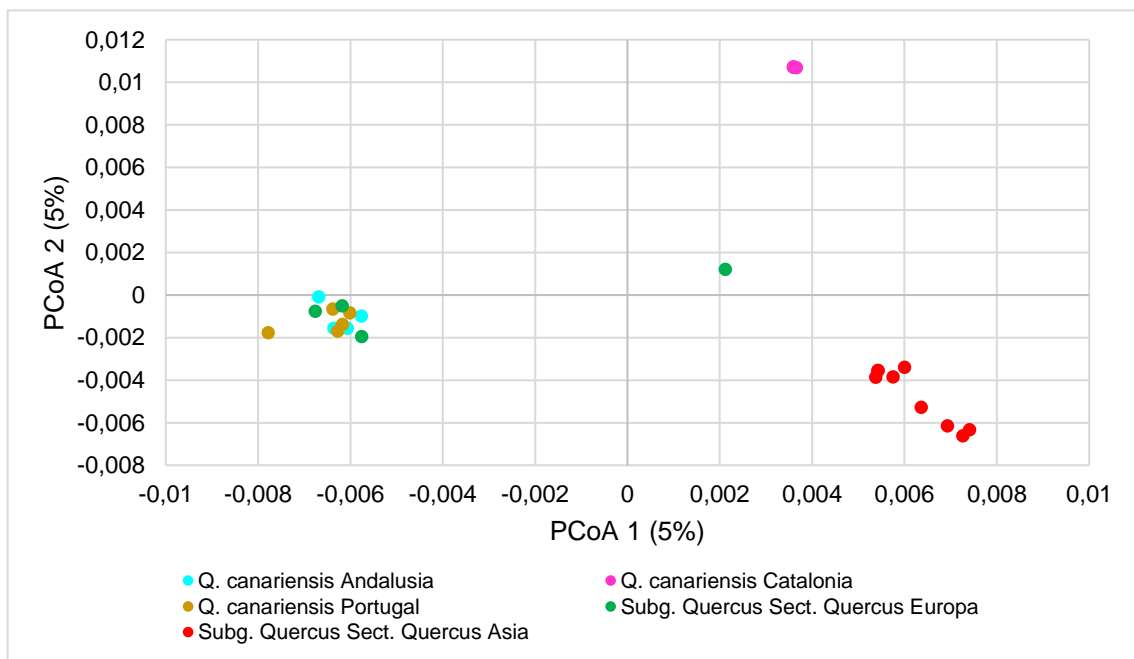


Fig. 10 - Principal coordinates analysis plot. Results of cpDNA sub-dataset showing the first two principal coordinates that, combined, explain 10% of the observed genetic variation.

### 3.2.3 - PCoA (ITS)

The results of the ITS PCoA are presented in graph Fig. 11 and 12. Like the results of cpDNA we did a second PCoA (Fig.12) with a second sub-dataset (black circle in Fig.11).

Like the cpDNA analysis the ITS do not show a clear separation of individuals by region. In turn we see groups of individuals in accordance with their subgenus and section. This is most likely due to the large number of individuals in the dataset than the cpDNA analysis, that includes species that are from the same taxonomy group but are from different continents.

Even though, we see that there are mostly well grouped and that, like in Fig.6, the closest group to our species is the *Quercus* subgenus. This makes sense as they are from the same subgenus.

In Fig.9 we see our samples with more resolution, unlike with the cpDNA the samples *Q. canariensis* don't show much differences between the three regions. And in general, there is no obvious distinction between species from America, Asia, or Europe.

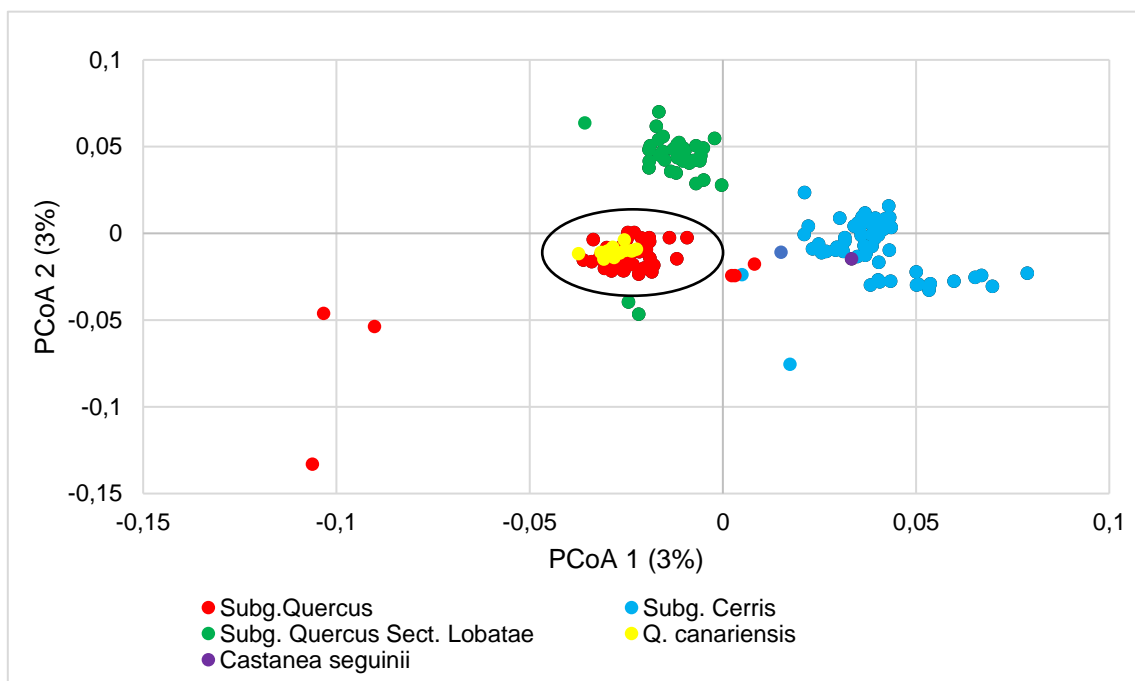


Fig. 11 - Principal coordinates analysis plot. Results of ITS dataset showing the first two principal coordinates that, combined, explain 6% of the observed genetic variation.

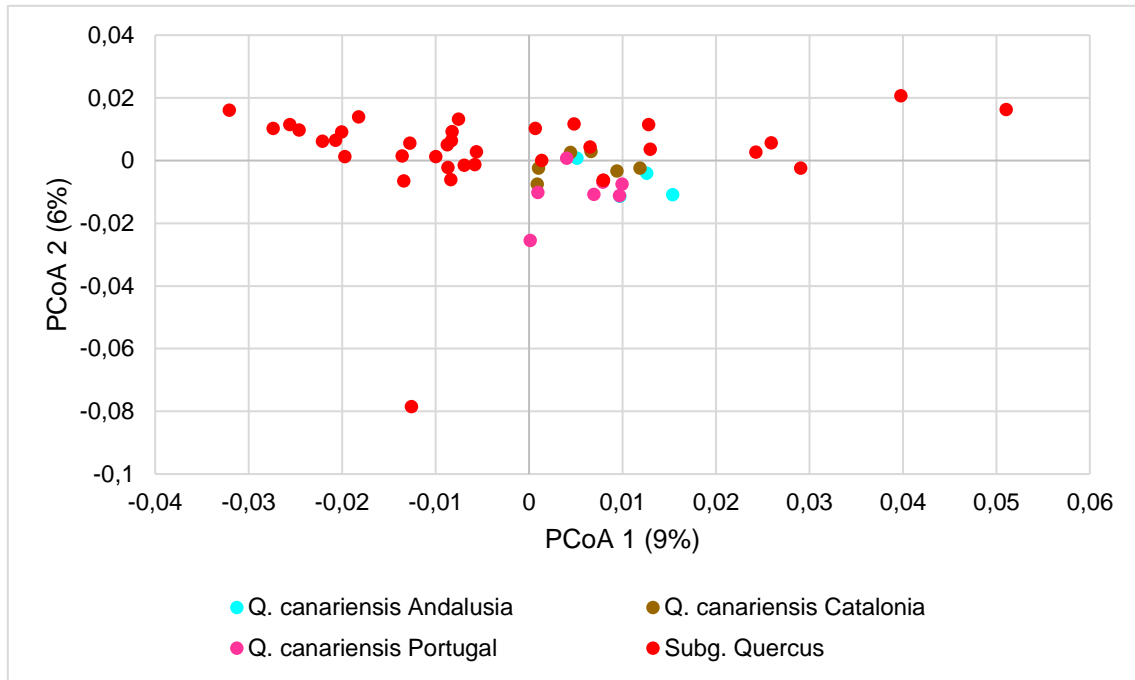


Fig. 12 - Principal coordinates analysis plot. Results of the ITS sub-dataset showing the first two principal coordinates that, combined, explain 15% of the observed genetic variation.

### 3.5.4- Nucleotide Diversity ( $\pi$ )

Nucleotide diversity indexes ( $\pi$ ) calculated in R for the regional groups of *Q. canariensis* for Andalusia (AND), Catalonia (CAT) and Portugal (POR) region for cpDNA and the ITS are shown in (Table 2, 3 and 4, 5 respectively).

The nucleotide diversity index for cpDNA is the lowest for CAT dataset, followed by AND dataset and finally POR dataset. When we compared the results of the two tables of results, we see that for the ITS (Table 4) the lowest nucleotide diversity belongs to AND dataset, instead of belonging to the CAT dataset. POR dataset maintains the highest value of nucleotide diversity in both cpDNA and ITS.

Nucleotide diversity calculated using PoPoolation (Table 3 and 5), support the results from R the lowest values are from Catalonia and there is slight difference between Andalusia and Portugal. Andalusia is the most diverse in this case, followed by Portugal and lastly Catalonia. The results of the nucleotide diversity from both R and PoPoolation, but also the phylogenetic tree and PCoA of cpDNA dataset are all conclusive in this matter. This is also supported by our ENM results, that show that in the past, as far back as the MH Portuguese and Andalusian hotspots were probably connected. Since this separation occurred recently, the two hotspots haven't accumulated enough differences to be observed in our analysis.

Furthermore, our ENM results also predicted that, as far back as the LGM the Catalonia populations has been isolated, this may be the one of the possible causes for differences in the Catalonia populations.

The ITS results from PoPoolation (Table 5), show that Portugal has the lowest nucleotide diversity followed by Andalusia and Catalonia. These results are completely different from the results from R. Which points to the inconclusive results from all the ITS dataset.

Table 2 - Results of nucleotide diversity ( $\pi$ ) calculated with “pegas” package in R, for the cpDNA

Region	Nucleotide diversity ( $\pi$ )	Dataset
Andalusia (AND)	$1,893 \times 10^{-4}$	Pool_and + s61D + s61H + s61M
Catalonia (CAT)	$2,481 \times 10^{-6}$	Pool_cat + s102_14 + s103_1 + s103_5 + s106_3
Portugal (POR)	$3,450 \times 10^{-4}$	Pool_por + s30K + s30M + s121_3 + s121_10

Table 3 – Results of nucleotide diversity ( $\pi$ ) calculated with PoPoolation, for the cpDNA

Region	Nucleotide diversity ( $\pi$ )	Dataset
Andalusia (AND)	$3,077 \times 10^{-4}$	Pool_and
Catalonia (CAT)	$1,746 \times 10^{-4}$	Pool_cat
Portugal (POR)	$2,938 \times 10^{-4}$	Pool_por

Table 4 - Results of nucleotide diversity ( $\pi$ ) calculated with “pegas” package in R, for the ITS

Region of origin	Nucleotide diversity ( $\pi$ )	Dataset
Andalusia	$5,005 \times 10^{-3}$	Pool_and + s61D + s61H + s61M
Catalonia	$9,619 \times 10^{-3}$	Pool_cat + s102_14 + s103_1 + s103_5 + s106_3
Portugal	$1,219 \times 10^{-2}$	Pool_por + s30K + s30M + s121_3 + s121_10

Table 5 - Results of nucleotide diversity ( $\pi$ ) calculated with PoPoolation, for the ITS

Region of origin	Nucleotide diversity ( $\pi$ )	Dataset
Andalusia	$1,015 \times 10^{-2}$	Pool_and
Catalonia	$1,176 \times 10^{-2}$	Pool_cat
Portugal	$6,854 \times 10^{-3}$	Pool_por

### 3.3 – Inference of distribution patterns through ENM

All the ENM models simulated with MaxEnt had high predictive capacities (AUC > 0.9), which suggested the reliability of the simulations based on the bioclimatic factors and occurrences of *Q. canariensis*.

The most important variables related to the species distributions were: Annual Precipitation (Bio 12) after this Precipitation Seasonality (Bio 15). But most of variable contribution came from Bio 12. The remaining variables little showed no percentage of contributions.

Table 6 - Average (standard deviation) percent contribution of each variable to the model replicates

Variable	Percent contribution	Standard Deviation
bio12	87.1	2.86
bio15	12.1	2.85
bio8	0.7	0.48
bio3	0.1	0.1
bio6	0	0

The models obtained using the present climatic conditions predicted potential areas of occurrence that overall fit the distribution of the occurrence dataset and overlaps with the distribution of the species (Fig.1, Fig. 10, Fig. 11, and Fig.12). The model identified areas with the highest climatic suitability in Tangier region of Morocco and in the Cádiz province in Andalusia, Spain. But also, along the Mediterranean coast of Algeria and Tunisia, this coincides with the distribution of the species. Although there are some areas in which the predicted climatic suitability is low even though occurrences are present, more specifically the southern part of the Monchique region in Portugal and the province of Barcelona.

For the projections to the past conditions, all three models used for the species showed an identical pattern of decrease of suitable habitat over time. With the case of the MIROC-ESM (Fig.12) having a decrease in suitable habitat in the Iberian Peninsula but an increase in Morocco going from Last Glacial Maximum (LGM) to Mid Holocene (MH).

In the LGM we have, in general, for all three models a great increase of suitable habitat area, in Portugal, Andalusia and the north Africa that encompass the Present distribution of the species. We also see that the habitat area from Portugal and Andalusia is connected according to the model. While in Catalonia there is a decrease in favourable conditions and habitat area in all the models for LGM.

Next in the MH there is a contraction in suitable habitat in some regions. Although there are some cases that the habitat has been moved, that is other less suitable areas have become more suitable. Like in Fig.11 we can see contraction in suitable area in the south of the Iberian Peninsula, while in the centre and north of Portugal we have an expansion. This is also true for Morocco in Fig.12 where see a decrease near the Mediterranean coast while in the interior there is an increase. In Catalonia we see an increase in favourable conditions in all three models. As well as a decrease in favourable conditions in the south of the Iberian Peninsula between Monchique and Andalusia, effectively separating the populations or decrease gene flow between them.

Projections into the future, for 2050 using different models coincide to show a decrease in suitable areas in Monchique Andalusia and north Africa, but also an increase of favourable conditions for the centre and north of Portugal and along the north coast of the Iberian Peninsula, The different models also support an increase of favourable conditions all the way across the Pyrenees mountain range, with the exception of the MPI-ESM model (Fig.11) that shows a decrease compared with the present.



In 2070 we got different results deepening on the model; in CCSM4 (Fig. 10) the suitable areas are mostly the same, with a decrease in Andalusia. This contrasts with MPI-ESM model (Fig. 11) that shows an increase of suitable area in the same region. In MIROC-ESM (Fig. 12) we see a continuous decrease when compared to 2050 in the south of Portugal, Andalusia, and North Africa. There is a consensus from all the 2070 models for the increased favourable conditions in the north of the Iberian Peninsula.

The results from ENM have a similar pattern in all three bioclimatic models used. We see a contraction of suitable habitat from the LGM to present time and, from the present into the future, except MPI-ESM model that predict an expansion in suitable areas by 2070. In the case of MIROC-ESM model we also see a contraction of some of the habitat areas an expansion of other areas from LGM to the MH before a contraction from of the same areas from MH to the present day. The comparison of LGM and present-day distributions shows the recent separation of Portuguese and Andalusian hotspots and the consistent isolation of the Catalonia hotspot mentioned above.

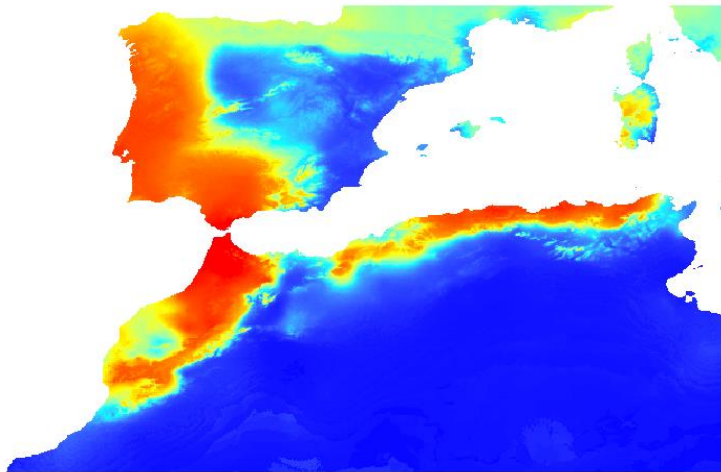
Regarding the current populations hotspot, we can see that Monchique in Portugal maintain the same level of suitability in future projections using the MPI-ESM, but not in the CCSM and MIROC-ESM, that show a variation in the level of suitability. This suggest that the species must move northward in Portugal to fulfil its niche requirements in the future. In Cadiz, Andalusia the projections into future conditions show that despite the decrease both potential habitat area and level suitability, it will remain a hotspot for the species. Contrarily, in Catalonia we can see consensus to the shift of suitable areas into higher altitude areas, as previously predicted by other authors (e.g. Penuelas and Boada, 2003).

In North Africa, more specific the Tangier region of Morocco, like Cadiz, will remain a hotspot for the species in the future. Despite the fluctuations in Atlas Mountains areas, it will remain as a suitable habitat into the future. The suitable areas in Algeria Mediterranean coast will also suffer a contraction, but the area will retain high levels of suitability. Furthermore, the results also show a high probability of habitat fragmentation in Algeria, with the current distribution being separated into isolated populations, thus affecting the gene pool of this hotshot.

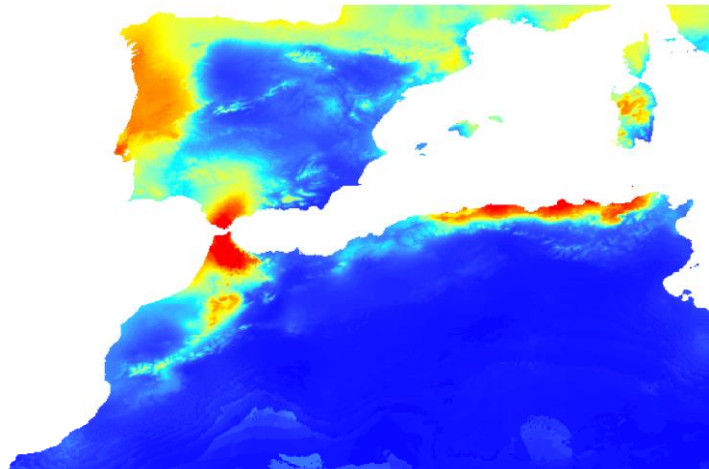
Additionally, models predict increased suitability in areas that fall outside the current distribution of *Q. canariensis*. Notably there is the Isle of Sardinia which has a high level of suitability. Despite this, its highly improbable that *Q. canariensis* will be able to reach this area. This is true for other areas like for example north Atlantic coast of the

Iberian Peninsula. Even though the high level that of suitable conditions our model does not consider the dispersion capabilities of *Q. canariensis*, nor other variables like geographical barriers our human activity that also impart the dispersion range.

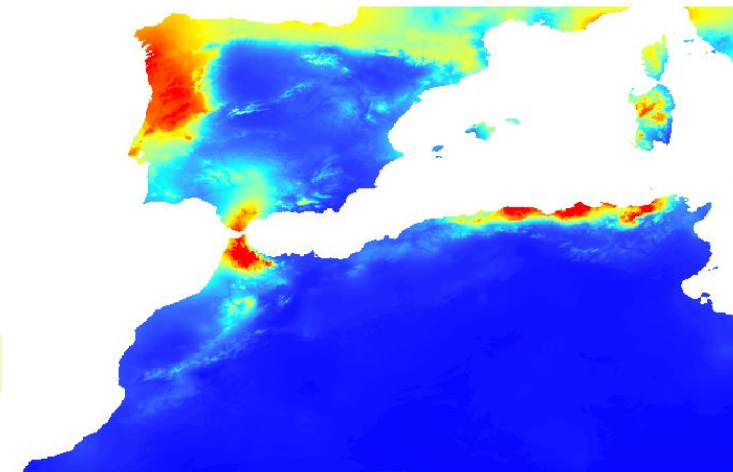
Last Glacial Maximum



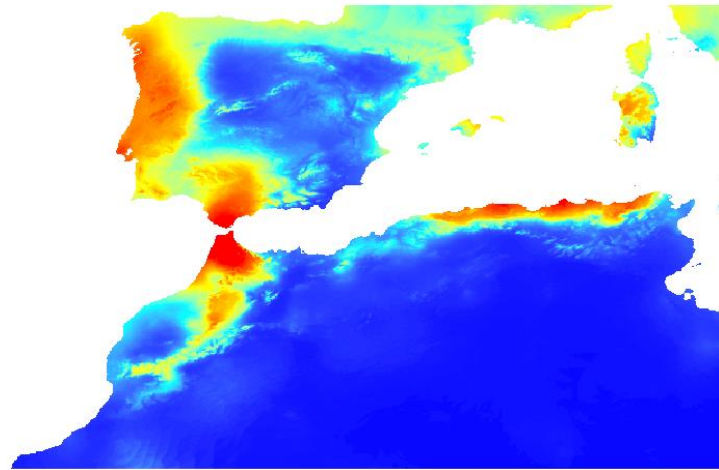
Present



2050



Mid Holocene



2070

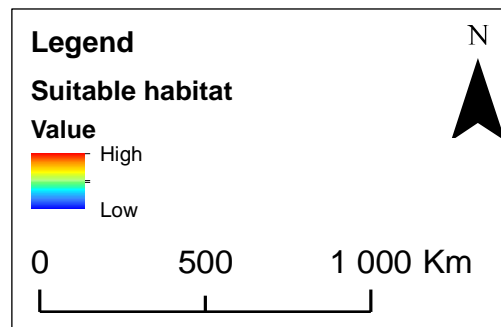
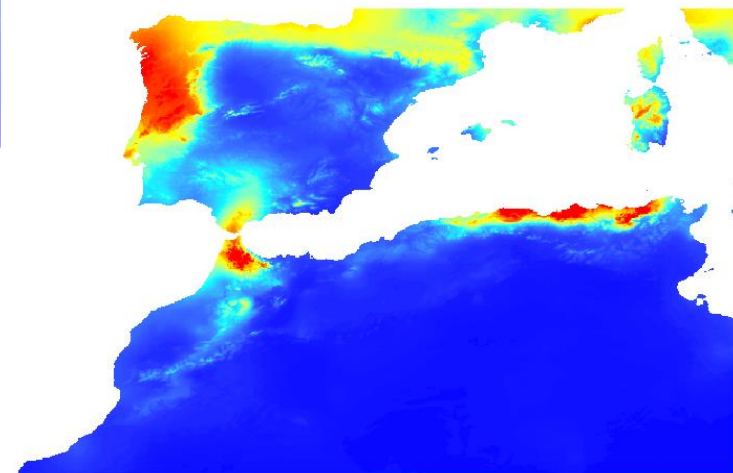
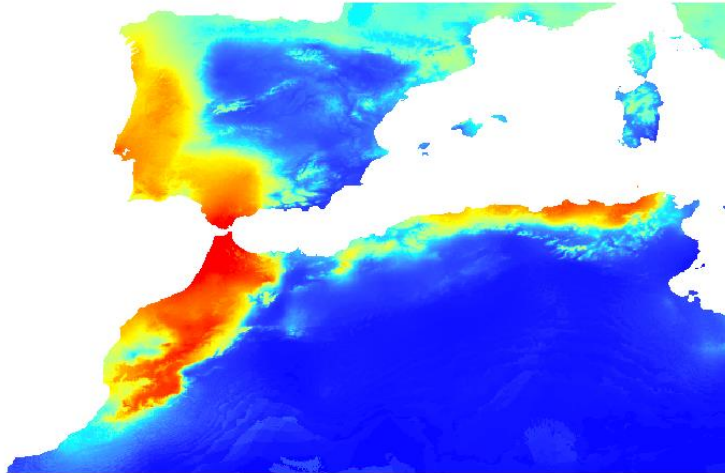
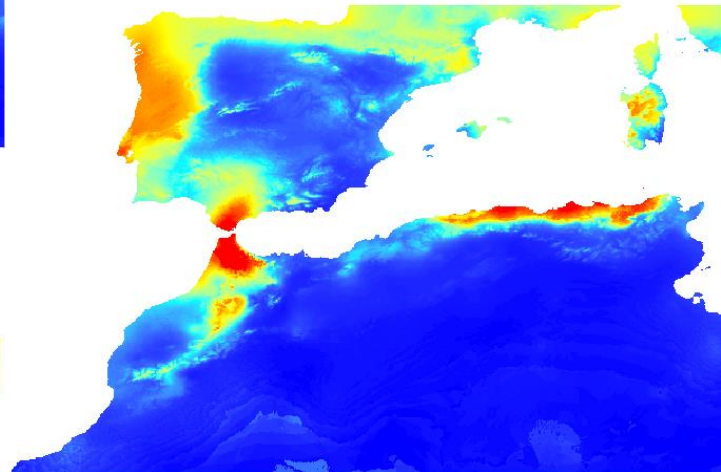


Fig. 13 - Predicted current (middle), past (left) and future (right) periods habitat for *Q. canariensis*, using CCSM4 model.

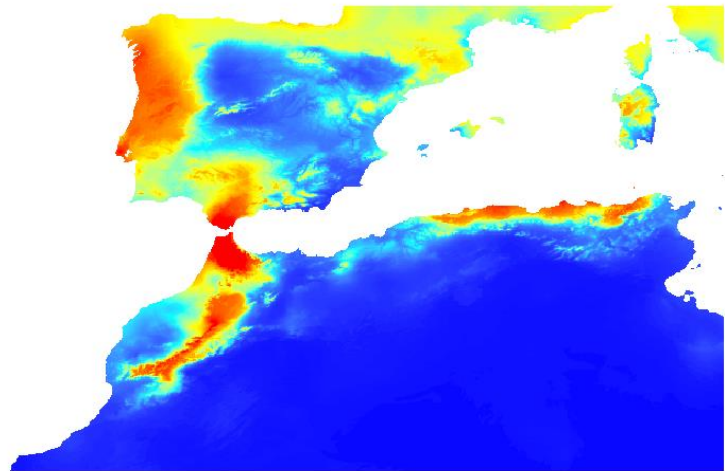
Last Glacial Maximum



Present



Mid Holocene



2070

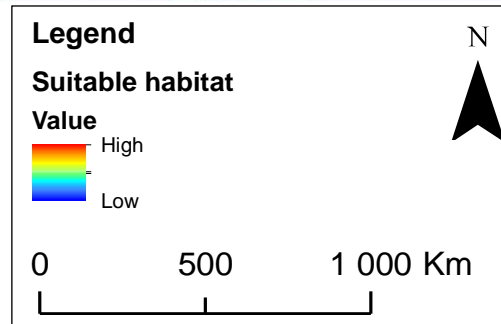
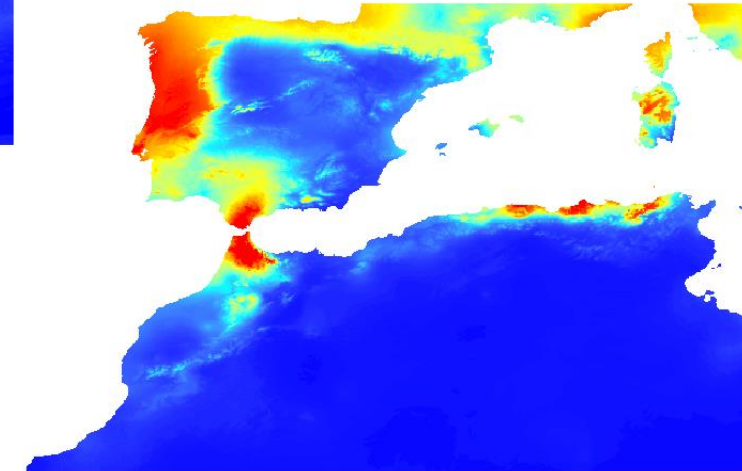
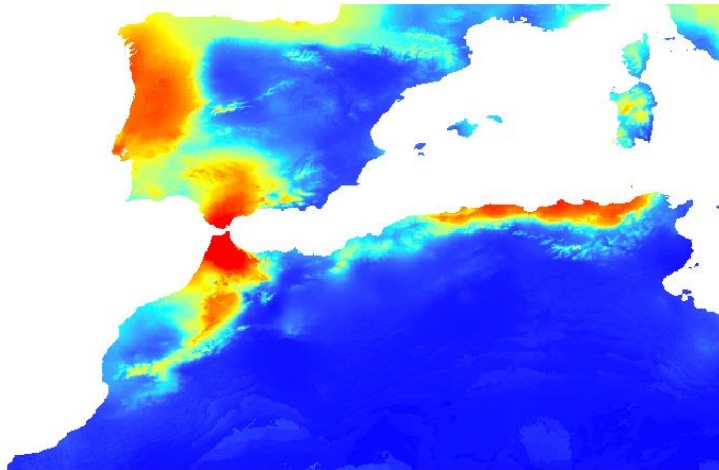
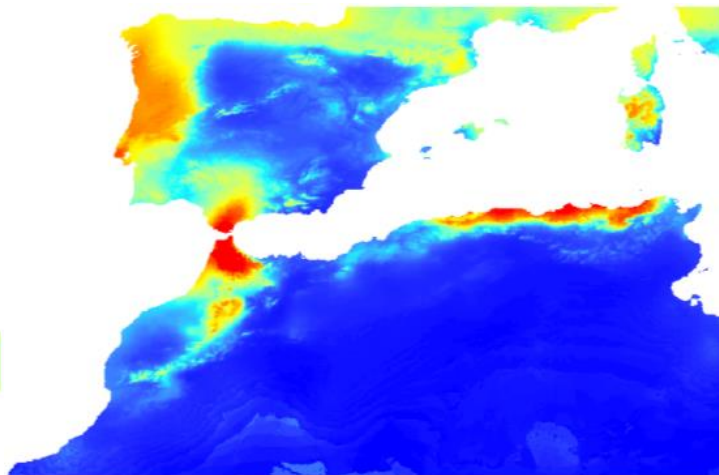


Fig. 14 - Predicted current (middle), past (left) and future (right) periods habitat for *Q. canariensis*, using MPI-ESM model.

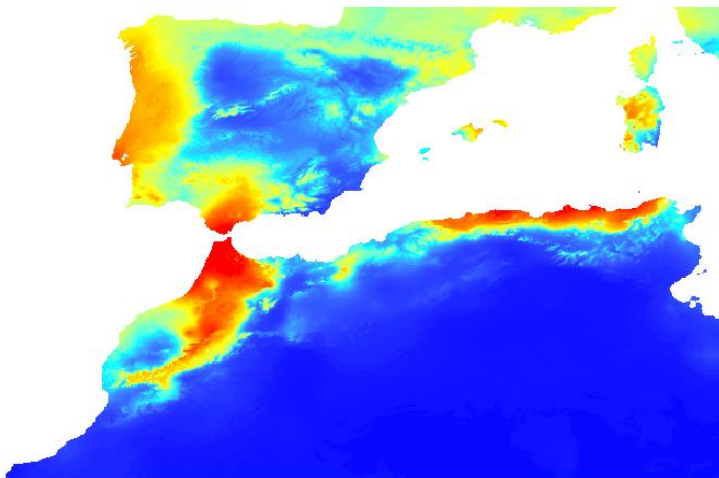
Last Glacial Maximum



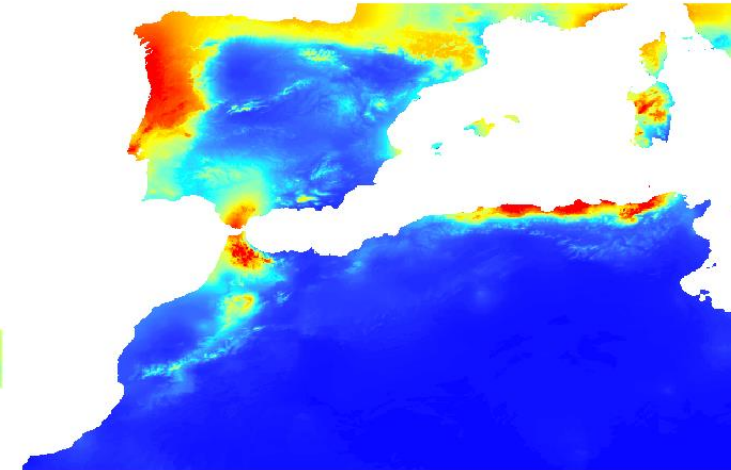
Present



Mid Holocene



2050



2070

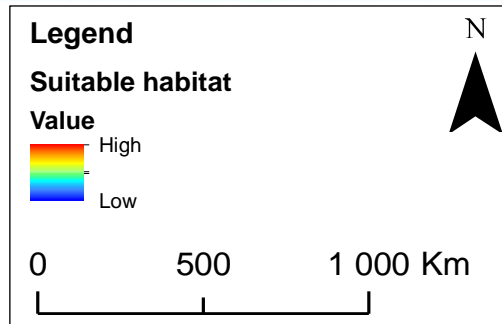
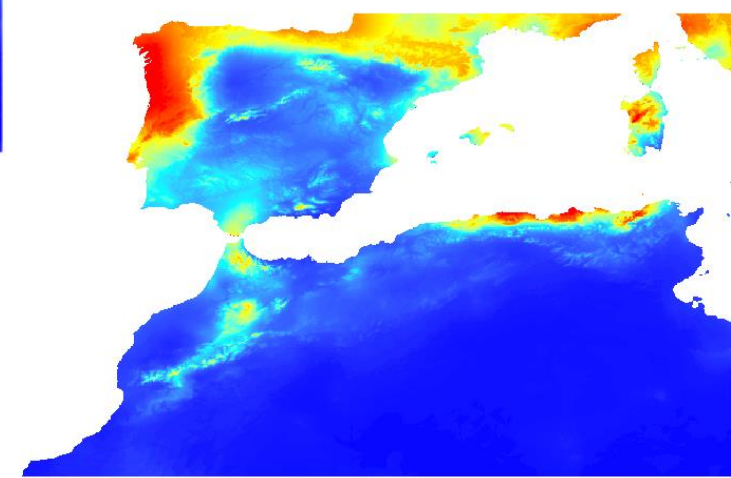


Fig. 15 - Predicted current (middle), past (left) and future (right) periods habitat for *Q. canariensis*, using MIROC-ESM model.

## 4 - Conclusion

In this work we analysed the complete chloroplast genome and ITS nuclear DNA of *Q. canariensis* an endemic species from both the Iberian Peninsula and North Africa. Additionally, we used ENM software to project the possible distribution of the *Q. canariensis* in the past and future.

Firstly, we must recognize a shortcoming of this work. The nuclear DNA marker used, the ITS's, have not produce conclusive results when compared with the cpDNA. It is clear in case that the ITS sequences were not sufficient to create a genetic diversity profile of *Q. canariensis* regarding the nuclear DNA.

Taking into accounts our projections into the future distribution of *Q. canariensis* we can see that there are areas that appears to be more stable, in terms of suitable conditions. Cadiz, Tangier and Algerian coast seem to be prime candidates for any conservation efforts. These areas offer a high level of suitability and area for *Q. canariensis*. Monchique area has some fluctuations depending on the model, so to prevents the disappearance of *Q. canariensis* from this area conservation action (including seed collection) should be taken now before climatic conditions start to deteriorate. In Catalonia since the tendency is to shift suitable areas to higher altitude, thus the areas around the Pyrenees mountains is the appropriate choice for this region.

## 5. Future work

Since this study into *Q. Canariensis* is currently still ongoing, we are currently studying additional nuclear DNA regions. For that, we are mapping reads generated for this work against the complete genome of *Quercus robur* (<https://www.ncbi.nlm.nih.gov/genome/10990>). We expect that this new analysis will produce in more conclusive information from the *Q. canariensis* nuclear DNA structure.

Additionally, we have also recently obtained individuals samples from North Africa that can also be included to supplement the analysis and to improve our knowledge on the species across its entire distribution range.

## References

- Adams, H. D., Macalady, A. K., Breshears, D. D., Allen, C. D., Stephenson, N. L., Saleska, S. R., Huxman, T. E., and McDowell, N. (2010), Climate-Induced Tree Mortality: Earth System Consequences, *Eos Trans. AGU*, 91( 17), 153– 154, doi:10.1029/2010EO170003.
- Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T., & Curtis-McLane, S. (2008). Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary applications*, 1(1), 95-111.
- Allen, C. D., Macalady, A. K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., ... & Gonzalez, P. (2010). A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *Forest ecology and management*, 259(4), 660-684.
- Arbolapp (CSIC/FECYT). 2017. *Quercus canariensis*. Available at: <http://www.arbolapp.es/en/species/info/quercus-canariensis/>.
- Ashkenazy, H., Cohen, O., Pupko, T., & Huchon, D. (2014). Indel reliability in indel-based phylogenetic inference. *Genome biology and evolution*, 6(12), 3199-3209.
- Ayres, M. P., & Lombardero, M. J. (2000). Assessing the consequences of global change for forest disturbance from herbivores and pathogens. *Science of the Total Environment*, 262(3), 263-286.
- Blanca, G., Cabezudo, B., Hernández-Bermejo, J.E., Herrera, C.M., Muñoz, J. and Valdés, B. (eds). 2000. Libro Rojo de la Flora Silvestre Amenazada de Andalucía. Tomo II. pp. 375. Consejería de Medio Ambiente, Junta de Andalucía, Sevilla.
- Brehm, J.M., Maxted, N., Martins-Loução, M.A. et al. New approaches for establishing conservation priorities for socio-economically important plant species. *Biodivers Conserv* 19, 2715–2740 (2010). <https://doi.org/10.1007/s10531-010-9871-4>.
- Brown JL, Bennett JR, French CM (2017). SDMtoolbox 2.0: the next generation Python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses. *PeerJ PDF*
- Choat, B., Jansen, S., Brodribb, T. J., Cochard, H., Delzon, S., Bhaskar, R., ... & Jacobsen, A. L. (2012). Global convergence in the vulnerability of forests to drought. *Nature*, 491(7426), 752-755.

- Cronn, R., Knaus, B. J., Liston, A., Maughan, P. J., Parks, M., Syring, J. V., & Udall, J. (2012). Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany*, 99(2), 291-311.
- da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., ... & Pereira, R. J. (2016). Next-generation biology: sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, 30, 3-13.
- Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic acids research*, 45(4), e18-e18.
- Doyle, J. J., & Doyle, J. L. (1987). *A rapid DNA isolation procedure for small quantities of fresh leaf tissue* (No. RESEARCH).
- Drobinski, P., Da Silva, N., Bastin, S., Mailler, S., Muller, C., Ahrens, B., ... & Lionello, P. (2020). How warmer and drier will the Mediterranean region be at the end of the twenty-first century?. *Regional Environmental Change*, 20(3), 1-12.
- Egan, A. N., Schlueter, J., & Spooner, D. M. (2012). Applications of next-generation sequencing in plant biology. *American journal of botany*, 99(2), 175-185.
- ESRI 2016. ArcGIS Desktop: Release 10.5. Redlands, CA: Environmental Systems Research Institute.
- Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation*, 38-49.
- Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation*, 38-49.
- García-López, Javier & Gonzalo Jimenez, Julian & Allué, Carmen. (2005). Fitoclimatología de *Quercus canariensis* Willd. en España. Potencialidades y adecuaciones fitoclimáticas. *Flora Montiberica*, ISSN 1138-5952, Nº. 29, 2005 (Ejemplar dedicado a: Homenaje a Antonio Segura Zubizarreta), pags. 14-29.
- Geneious Prime 2020.2 (<https://www.geneious.com>)



- Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., ... & Worley, P. H. (2011). The community climate system model version 4. *Journal of climate*, 24(19), 4973-4991.
- Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Böttinger, M., ... & Glushak, K. (2013). Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *Journal of Advances in Modeling Earth Systems*, 5(3), 572-597.
- Gorener, V., Harvey-Brown, Y. & Barstow, M. 2017. *Quercus canariensis*. The IUCN Red List of Threatened Species 2017: e.T78809256A78809271. <https://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T78809256A78809271.en>. Downloaded on 15 July 2020.
- Gower, J. C. (2014). *Principal coordinates analysis*. Wiley StatsRef: Statistics Reference Online.
- Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology letters*, 8(9), 993-1009.
- Hall, T. A. (1999, January). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In *Nucleic acids symposium series* (Vol. 41, No. 41, pp. 95-98). [London]: Information Retrieval Ltd., c1979-c2000..
- Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 25(15), 1965-1978.
- Hughes, L. (2000). Biological consequences of global warming: is the signal already apparent?. *Trends in ecology & evolution*, 15(2), 56-61.
- Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14), 3059-3066.
- Kearney, M. R., Wintle, B. A., & Porter, W. P. (2010). Correlative and mechanistic models of species distribution provide congruent forecasts under climate change. *Conservation letters*, 3(3), 203-213.

- Kearney, M., & Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology letters*, 12(4), 334-350.
- Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., ... & Schlötterer, C. (2011). PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PloS one*, 6(1), e15925.
- Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21), 4453-4455. Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Molecular biology and evolution*, 37(1), 291-294.
- Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., ... & Cheyne, S. M. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, 19(11), 1366-1379.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079.
- Lockhart, P. J., Steel, M. A., Hendy, M. D., & Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular biology and evolution*, 11(4), 605-612.
- Mace, G. M., & Purvis, A. (2008). Evolutionary biology and practical conservation: bridging a widening gap. *Molecular Ecology*, 17(1), 9-19.
- Marañón, T., Díaz, C. M. P., Ramos, I. M. P., & Villar, R. (2014). Tendencias en la investigación sobre ecología y gestión de las especies de *Quercus*. *Revista Ecosistemas*, 23(2), 124-129.
- Marañón, T., Ibáñez, B., Anaya-Romero, M., Muñoz-Rojas, M., & Pérez-Ramos, I. (2016). Oak trees and woodlands providing ecosystem services in Southern Spain. In *Trees beyond the wood conference proceedings* (pp. 291-299).
- McCarty, J. P. (2001). Ecological consequences of recent climate change. *Conservation biology*, 15(2), 320-331.

- McDowell, N., Pockman, W. T., Allen, C. D., Breshears, D. D., Cobb, N., Kolb, T., ... & Yezzer, E. A. (2008). Mechanisms of plant survival and mortality during drought: why do some plants survive while others succumb to drought?. *New phytologist*, 178(4), 719-739.
- Merow, C., Smith, M. J., & Silander Jr, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, 36(10), 1058-1069.
- Miller, M.A., Pfeiffer, W., and Schwartz, T. (2010) "Creating the CIPRES Science Gateway for inference of large phylogenetic trees" in Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010, New Orleans, LA pp 1 - 8.
- Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M., & Anderson, R. P. (2014). ENM eval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods in ecology and evolution*, 5(11), 1198-1205.
- Nei, M. (1987) *Molecular evolutionary genetics*. New York: Columbia University Press.
- Olalde, M., Herrán, A., Espinel, S., & Goicoechea, P. G. (2002). White oaks phylogeography in the Iberian Peninsula. *Forest Ecology and Management*, 156(1-3), 89-102.
- Oldeman, L. R. (1992). Global extent of soil degradation. In *Bi-Annual Report 1991-1992/ISRIC* (pp. 19-36). ISRIC.
- Paradis E (2010). "pegas: an R package for population genetics with an integrated-modular approach." *Bioinformatics*, **26**, 419-420.
- Paradis E, Schliep K (2019). "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R." *Bioinformatics*, **35**, 526-528.
- Pauls, S. U., Nowak, C., Bálint, M., & Pfenninger, M. (2013). The impact of global climate change on genetic diversity within populations and species. *Molecular ecology*, 22(4), 925-946.
- Pearson, R. G. (2007). Species' distribution modeling for conservation educators and practitioners. *Synthesis*. American Museum of Natural History, 50, 54-89.

- Penuelas, J., & Boada, M. (2003). A global change-induced biome shift in the Montseny mountains (NE Spain). *Global change biology*, 9(2), 131-140.
- Pérez-Ramos, I.M. and Marañón, T. 2009. 9240 Robledales ibéricos de *Quercus faginea* y *Quercus canariensis*. En: id: Dirección General de Medio Natural y Política Forestal, Ministerio de Medio Ambiente, y Medio Rural y Marino. In: VVAA (ed.), Bases ecológicas preliminares para la conservación de los tipos de hábitat de interés comunitario en España, pp. 56. Madrid. Rushton, B. S. (1993). Natural hybridization within the genus *Quercus* L. In *Annales des sciences forestières* (Vol. 50, No. Supplement, pp. 73s-90s). EDP Sciences.
- Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4), 231-259.
- Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4), 231-259.
- Ponder, W. F., Carter, G. A., Flemons, P., & Chapman, R. R. (2001). Evaluation of museum collection data for use in biodiversity assessment. *Conservation biology*, 15(3), 648-657.
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- Radosavljevic, A., & Anderson, R. P. (2014). Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of biogeography*, 41(4), 629-643.
- Robertson, M. P., Peter, C. I., Villet, M. H., & Ripley, B. S. (2003). Comparing models for predicting species' potential distributions: a case study using correlative and mechanistic predictive modelling techniques. *Ecological Modelling*, 164(2-3), 153-167.
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12), 5463-5467.
- Scholze, M., Knorr, W., Arnell, N. W., & Prentice, I. C. (2006). A climate-change risk analysis for world ecosystems. *Proceedings of the National Academy of Sciences*, 103(35), 13116-13120.

- Sgrò, C. M., Lowe, A. J., & Hoffmann, A. A. (2011). Building evolutionary resilience for conserving biodiversity under climate change. *Evolutionary applications*, 4(2), 326-337.
- Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X., Huang, L., & Liu, C. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic acids research*, 47(W1), W65-W73.
- Soltis, P. S., & Soltis, D. E. (2003). Applying the bootstrap in phylogeny reconstruction. *Statistical Science*, 256-267.
- Spielman, D., Brook, B. W., & Frankham, R. (2004). Most species are not driven to extinction before genetic factors impact them. *Proceedings of the National Academy of Sciences*, 101(42), 15261-15264.
- Sumner, G. N., Romero, R., Homar, V., Ramis, C., Alonso, S., & Zorita, E. (2003). An estimate of the effects of climate change on the rainfall of Mediterranean Spain by the late twenty first century. *Climate Dynamics*, 20(7-8), 789-805.
- Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1), 36-46.
- Urban, M. C. (2015). Accelerating extinction risk from climate change. *Science*, 348(6234), 571-573.
- Urbieto, T.I., Zavala, M.A. & Marañón, T. 2008b. Human and non-human determinants of forest composition in southern Spain: evidence of shifts toward cork oak dominance as a result of management over the past century. *Journal of Biogeography* 35: 1688–1700.
- Van, K., Kim, D. H., Shin, J. H., & Lee, S. H. (2011). Genomics of plant genetic resources: past, present and future. *Plant Genetic Resources*, 9(2), 155.
- Vázquez, F. M., Ramos, S., & Ruiz, T. (2000). Hybridisation processes in Mediterranean Oaks from South Spain. *J Intern Oak Soc*, 12, 108-117.
- Vila-Viçosa, C., Vázquez, F., Meireles, C., & Pinto-Gomes, C. (2014). Taxonomic peculiarities of marcescent oaks (*Quercus*, Fagaceae) in southern Portugal.
- Walther, G., Post, E., Convey, P. *et al.* Ecological responses to recent climate change. *Nature* **416**, 389–395 (2002). <https://doi.org/10.1038/416389a>

- Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., ... & Ise, T. (2011). MIROC-ESM 2010: Model description and basic results of CMIP5-20c3m experiments. *Geoscientific Model Development*, 4(4), 845.
- Woodward, F. I., & Williams, B. G. (1987). Climate and plant distribution at global and local scales. *Vegetatio*, 69(1-3), 189-197.
- Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., ... & Lahoz-Monfort, J. J. (2020). A standard protocol for reporting species distribution models. *Ecography*.

# Appendix

Appendix 1 - List of samples of *Q. canariensis* used in genetic diversity analyses

Sample ID	Place of Origin	Species
s30E	Monchique (Portugal)	<i>Q. canariensis</i>
s30G	Monchique (Portugal)	<i>Q. canariensis</i>
s30I	Monchique (Portugal)	<i>Q. canariensis</i>
s30J	Monchique (Portugal)	<i>Q. canariensis</i>
s30K	Monchique (Portugal)	<i>Q. canariensis</i>
s30M	Monchique (Portugal)	<i>Q. canariensis</i>
s32A	Monchique (Portugal)	<i>Q. pyrenaica</i>
s32H	Monchique (Portugal)	<i>Q. petraea</i>
s61B	Andalusia (Spain)	<i>Q. canariensis</i>
s61C	Andalusia (Spain)	<i>Q. canariensis</i>
s61D	Andalusia (Spain)	<i>Q. canariensis</i>
s61G	Andalusia (Spain)	<i>Q. canariensis</i>
s61H	Andalusia (Spain)	<i>Q. canariensis</i>
s61I	Andalusia (Spain)	<i>Q. canariensis</i>
s61K	Andalusia (Spain)	<i>Q. canariensis</i>
s61L	Andalusia (Spain)	<i>Q. canariensis</i>
s61M	Andalusia (Spain)	<i>Q. canariensis</i>
s61N	Andalusia (Spain)	<i>Q. canariensis</i>
s102_5	Catalonia (Spain)	<i>Q. canariensis</i>
s102_6	Catalonia (Spain)	<i>Q. canariensis</i>
s102_8	Catalonia (Spain)	<i>Q. canariensis</i>
s102_14	Catalonia (Spain)	<i>Q. canariensis</i>
s103_1	Catalonia (Spain)	<i>Q. canariensis</i>
s103_5	Catalonia (Spain)	<i>Q. canariensis</i>
s103_7	Catalonia (Spain)	<i>Q. canariensis</i>
s106_1	Catalonia (Spain)	<i>Q. canariensis</i>
s106_3	Catalonia (Spain)	<i>Q. canariensis</i>
s106_4	Catalonia (Spain)	<i>Q. canariensis</i>
s109_9	Catalonia (Spain)	<i>Q. faginea</i>
s121_3	Monchique (Portugal)	<i>Q. canariensis</i>
s121_10	Monchique (Portugal)	<i>Q. canariensis</i>

Pool_and	Andalusia (Spain)	s61B, s61C, s61G, s61I, s61K, s61L s61N
Pool_cat	Catalonia (Spain)	s103_1, s103_7, s102_5, s102_6, s102_8, s106_1, s106_3, s106_4
Pool_por	Monchique (Portugal)	s30E, s30G, s30I, s30J, s30K, s121_3, s121_10

Appendix 2 – DNA extraction protocol

DNA extraction protocol for leaf tissue samples (modified from Doyle and Doyle, 1987).

1. Preheat 5ml to 7.5 ml of CTAB isolation buffer (2% hexadecyltrimethylammonium bromide (CTAB), 1.4 M NaCl, 0.2% 2-mercaptoethanol, 20 mM EDTA, 100 mM Tris-HCl, pH 8.0) in a 30 ml glass centrifuge tube to 60°C in a water bath.
2. Grind 0.5g to 1.0 g samples of young leaf tissues to a fine powder in liquid nitrogen
3. Incubate sample at 60°C for 30 (15-60) minutes with optional occasional gentle swirling.
4. Extract once with chloroform-isoamyl alcohol (24:1), mixing gently but thoroughly.

This produces two phases, an upper aqueous phase which contains the DNA, and a lower chloroform phase that contains some degraded proteins, lipids, and many secondary compounds. The interface between these two phases contains most of the "junk"--cell debris, many degraded proteins, etc.

5. Spin in clinical centrifuge (swinging bucket rotor) at room temperature to concentrate phases. We use setting 7 on our IEC clinical (around 6,000 x g) for 10 min.

This is mainly to get rid of the junk that is suspended in the aqueous phase.

Generally, the aqueous phase will be clear, though often coloured, following centrifugation, but this is not always the case.

6. Remove aqueous phase with wide bore pipet, transfer to clean glass centrifuge tube, add 2/3 volumes cold isopropanol, and mix gently to precipitate nucleic acids.

A wide bore pipet is used because DNA in solution is a long, skinny molecule that is easily broken (sheared) when it passes through a narrow opening. Gentleness also improves the quality (length) of DNA.

In some cases, this stage yields large strands of nucleic acids that can be spooled out with a glass hook for subsequent preparation. In most cases, this is not the case, however, and the sample is either flocculent, merely cloudy-looking, or, in some instances, clear. If no evidence of precipitation is observed at this stage, the sample may be left at room temperature for several hours to overnight. This is one convenient



stopping place, in fact, when many samples are to be prepped. In nearly all cases, there is evidence of precipitation after the sample has been allowed to settle out in this manner.

7. If possible, spool out nucleic acids with a glass hook and transfer to 10-20 ml of wash buffer (76% EtOH, 10 mM ammonium acetate).

a. Preferred alternative: Spin in clinical centrifuge (e.g. setting 3 on IEC) for 1-2 min. Gently pour off as much of the supernatant as possible without losing the precipitate, which will be a diffuse and very loose pellet. Add wash buffer directly to pellet and swirl gently to resuspend nucleic acids.

b. Last resort: Longer spins at higher speeds may be unavoidable if no precipitate is seen at all. This will result, generally, in a hard pellet (or, with small amounts, a film on the bottom of the tube) that does not wash well and may contain more impurities. Such pellets are difficult to wash, and in some cases, we tear them with a glass rod to promote washing at which point they often appear flaky.

Nucleic acids generally become much whiter when washed, though some colour may still remain.

8. Spin down (or spool out) nucleic acids (setting 7 IEC, 10 min) after a minimum of 20 min of washing. The wash step is another convenient stopping point, as samples can be left at room temperature in wash buffer for at least two days without noticeable problems.

9. Pour off supernatant carefully (some pellets are still loose even after this longer spin) and allow to air dry briefly at room temperature.

10. Resuspend nucleic acid pellet in 1 ml TE (10 mM Tris-HCl, 1 mM EDTA, pH 7.4).

Although we commonly continue through additional purification steps, DNA obtained at this point is generally suitable for restriction digestion and amplification, so we'll stop here.

If DNA is to be used at this stage, pellets should be more thoroughly dried than indicated above.

Gel electrophoresis of nucleic acids at this step often reveals the presence of visible bands of ribosomal RNAs as well as high molecular weight DNA.

11. Add RNAase A to a final concentration of 10  $\mu$ g/ml and incubate 30 min at 37°C.

12. Dilute sample with 2 volumes of distilled water or TE, add ammonium acetate (7.5 M stock, pH 7.7) to a final concentration of 2.5 M, mix, add 2.5 volumes of cold EtOH, and gently mix to precipitate DNA.

DNA at this stage usually appears cleaner than in the previous precipitation. Dilution with water or TE is helpful, as we have found that precipitation from 1 ml total volume often produces a gelatinous precipitate that is difficult to spin down and dry adequately.




13. Spin down DNA at high speed (10,000 x g for 10 min in refrigerated centrifuge or setting 7 in clinical for 10 min).

14. Air dry sample and resuspend in appropriate amount of TE.

Appendix 3 - DNA Library Protocol

Protocol for use with NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (E7645, E7103)



Symbols

	This caution sign signifies a step in the protocol that has multiple paths leading to the same end point but is dependent on a user variable, like the amount of input DNA.
	Colored bullets indicate the cap color of the reagent to be added to a reaction.
	Stopping points in the protocol.

Starting Material: 500 pg–1 µg fragmented DNA. We recommend that DNA be sheared in 1X TE. If the DNA volume post shearing is less than 50 µl, add 1X TE to a final volume of 50 µl. Alternatively, samples can be diluted with 10 mM Tris-HCl, pH 8.0 or 0.1X TE.

1. NEBNext End Prep

1.1. Add the following components to a sterile nuclease-free tube:


 (green) NEBNext Ultra II End Prep Enzyme Mix	3 µl
 (green) NEBNext Ultra II End Prep Reaction Buffer	7 µl
Fragmented DNA	50 µl
Total Volume	60 µl

1.2. Set a 100 µl or 200 µl pipette to 50 µl and then gently pipette the entire volume up and down at least 10 times to mix thoroughly. Perform a quick spin to collect all liquid from the sides of the tube.

Note: It is important to mix well. The presence of a small amount of bubbles will not interfere with performance.

1.3. Place in a thermocycler, with the heated lid set to ≥ 75°C, and run the following program:

- 30 minutes @ 20°C
- 30 minutes @ 65°C
- Hold at 4°C

	If necessary, samples can be stored at -20°C; however, a slight loss in yield (~20%) may be observed. We recommend continuing with adaptor ligation before stopping.
---	--

2. Adaptor Ligation

2.1. Determine whether adaptor dilution is necessary.


	If DNA input is ≤ 100 ng, dilute the NEBNext Adaptor for Illumina in a solution of 10 mM Tris-HCl containing 10mM NaCl, pH 7.5-8.0.
---	---

Table 2.1: Adaptor Dilution

INPUT	ADAPTOR (VOLUME OF ADAPTOR: TOTAL VOLUME)	DILUTION	WORKING ADAPTOR CONCENTRATION
1 µg–101 ng	No Dilution		15 µM
100 ng–5 ng	10-Fold (1:10)		1.5 µM
less than 5 ng	25-Fold (1:25)		0.6 µM

Note: The appropriate adaptor dilution for your sample input and type may need to be optimized experimentally. The dilutions provided here are a general starting point. Excess adaptor should be removed prior to PCR enrichment.

2.2. Add the following components directly to the End Prep Reaction Mixture:

End Prep Reaction Mixture (Step 1.3 in Section 1)	60 µl
● (red) NEBNext Ultra II Ligation Master Mix*	30 µl
● (red) NEBNext Ligation Enhancer	1 µl
● (red) NEBNext Adaptor for Illumina **	2.5 µl
Total volume	93.5 µl

\* Mix the Ultra II Ligation Master Mix by pipetting up and down several times prior to adding to the reaction.

\*\* The NEBNext adaptor is provided in NEBNext Singleplex (NEB #E7350) or Multiplex (NEB #E7335, #E7500, #E7710, #E7730, #E7600, #E7535, and #E6609) Oligos for Illumina.

Note: The Ligation Master Mix and Ligation Enhancer can be mixed ahead of time and is stable for at least 8 hours @ 4°C. We do not recommend adding adaptor to a premix in the Adaptor Ligation Step.

2.3. Set a 100 µl or 200 µl pipette to 80 µl and then pipette the entire volume up and down at least 10 times to mix thoroughly. Perform a quick spin to collect all liquid from the sides of the tube.


(Caution: The NEBNext Ultra II Ligation Master Mix is very viscous. Care should be taken to ensure adequate mixing of the ligation reaction, as incomplete mixing will result in reduced ligation efficiency. The presence of a small amount of bubbles will not interfere with performance).

2.4. Incubate at 20°C for 15 minutes in a thermocycler with the heated lid off.


2.5. Add 3 µl of ● (red) USER™ Enzyme to the ligation mixture from Step 2.3.

Note: Steps 2.5 and 2.6 are only required for use with NEBNext Adaptors. USER enzyme can be found in the NEBNext Singleplex (NEB #E7350) or Multiplex (NEB #E7335, #E7500, #E7710, #E7730, #E7600, and #E6609) Oligos for Illumina.


2.6. Mix well and incubate at 37°C for 15 minutes with the heated lid set to ≥ 47°C.


	Samples can be stored overnight at -20°C.
---	---

### 3. Size Selection or Cleanup of Adaptor-ligated DNA

 If the starting material is greater than 50 ng, follow the protocol for size selection in Section 3A. For input less than or equal to 50 ng, size selection is not recommended to maintain library complexity. Follow the protocol for cleanup without size selection in Section 3B.

#### 3.A Size Selection of Adaptor-ligated DNA

 Note: The following section is for cleanup of the ligation reaction. The volumes of SPRIselect or NEBNext Sample Purification Beads provided here are for use with the sample contained in the exact buffer at this step. AMPure XP Beads can be used as well. If using AMPure XP Beads, allow the beads to warm to room temperature for at least 30 minutes before use. These bead volumes may not work properly for a cleanup at a different step in the workflow, or if this is a second cleanup at this step. For cleanups of samples contained in different buffer conditions, the volumes may need to be experimentally determined.

 The following size selection protocol is for libraries with 200 bp inserts only. For libraries with different size fragment inserts, refer to the table below for the appropriate volumes of beads to be added. The size selection protocol is based on starting volume of 96.5 µl.

To select a different insert size than 200 bp, please use the volumes in this table:

Table 3.1: Recommended conditions for bead-based size selection.

	APPROXIMATE INSERT SIZE	150 bp	200 bp	250 bp	300-400 bp	400-500 bp	500-700 bp
LIBRARY PARAMETERS	Approx. Final Library Size (Insert+Adaptor+primers)	270 bp	320 bp	370 bp	480 bp	600 bp	750-800 bp
VOLUME TO BE ADDED (µl)	1st Bead Selection	50	40	30	25	20	15
	2nd Bead Selection	25	20	15	10	10	10

3A.1. Vortex SPRIselect or NEBNext Sample Purification Beads to resuspend.

3A.2. Add 40 µl (~ 0.4X) of resuspended beads to the 96.5 µl ligation reaction. Mix well by pipetting up and down at least 10 times. Be careful to expel all of the liquid out of the tip during the last mix. Vortexing for 3-5 seconds on high can also be used. If centrifuging samples after mixing, be sure to stop the centrifugation before the beads start to settle out.

3A.3. Incubate samples on bench top for at least 5 minutes at room temperature.

3A.4. Place the tube/plate on an appropriate magnetic stand to separate the beads from

the supernatant. If necessary, quickly spin the sample to collect the liquid from the sides of the tube or plate wells before placing on the magnetic stand.

3A.5. After 5 minutes (or when the solution is clear), carefully transfer the supernatant containing your DNA to a new tube (Caution: do not discard the supernatant). Discard the beads that contain the unwanted large fragments.

3A.6. Add 20  $\mu$ l (0.2X) resuspended SPRIselect or NEBNext Sample Purification Beads to the supernatant and mix at least 10 times. Be careful to expel all of the liquid from the tip during the last mix. Then incubate samples on the bench top for at least 5 minutes at room temperature.

3A.7. Place the tube/plate on an appropriate magnetic stand to separate the beads from the supernatant. If necessary, quickly spin the sample to collect the liquid from the sides of the tube or plate wells before placing on the magnetic stand.

3A.8. After 5 minutes (or when the solution is clear), carefully remove and discard the supernatant that contains unwanted DNA. Be careful not to disturb the beads that contain the desired DNA targets (Caution: do not discard beads).

3A.9. Add 200  $\mu$ l of 80% freshly prepared ethanol to the tube/plate while in the magnetic stand. Incubate at room temperature for 30 seconds, and then carefully remove and discard the supernatant. Be careful not to disturb the beads that contain DNA targets.

3A.10. Repeat Step 3.A.9 once. Be sure to remove all visible liquid after the second wash. If necessary, briefly spin the tube/plate, place back on the magnet and remove traces of ethanol with a p10 pipette tip.

3A.11. Air dry the beads for up to 5 minutes while the tube/plate is on the magnetic stand with the lid open.

Caution: Do not overdry the beads. This may result in lower recovery of DNA target. Elute the samples when the beads are still dark brown and glossy looking, but when all visible liquid has evaporated. When the beads turn lighter brown and start to crack, they are too dry.

3A.12. Remove the tube/plate from the magnetic stand. Elute the DNA target from the beads into 17  $\mu$ l of 10 mM Tris-HCl or 0.1X TE.

3A.13. Mix well on a vortex mixer or by pipetting up and down 10 times. Incubate for at least 2 minutes at room temperature. If necessary, quickly spin the sample to collect the liquid from the sides of the tube or plate wells before placing back on the magnetic stand.

3A.14. Place the tube/plate on a magnetic stand. After 5 minutes (or when the solution is clear), transfer 15  $\mu$ l to a new PCR tube for (amplification).



Samples can be stored at -20°C.

3.B Cleanup of Adaptor-ligated DNA without Size Selection (for input  $\leq$  50 ng)

The following section is for cleanup of the ligation reaction. If your input DNA is  $>$  100 ng, follow the size selection protocol in section 3.A.

Note: The volumes of SPRIselect or NEBNext Sample Purification Beads provided here are for use with the sample contained in the exact buffer at this step. AMPure XP Beads can be used as well. If using AMPure XP Beads, allow the beads to warm to room temperature for at least 30 minutes before use. These bead volumes may not work properly for a cleanup at a different step in the workflow, or if this is a second cleanup at this step. For cleanups of samples contained in different buffer conditions, the volumes may need to be experimentally determined.

3B.1. Vortex SPRIselect or NEBNext Sample Purification Beads to resuspend.

3B.2. Add 87  $\mu$ l (0.9X) resuspended beads to the Adaptor Ligation reaction. Mix well by pipetting up and down at least 10 times. Be careful to expel all of the liquid out of the tip during the last mix. Vortexing for 3-5 seconds on high can also be used. If centrifuging samples after mixing, be sure to stop the centrifugation before the beads start to settle out.

3B.3. Incubate samples on bench top for at least 5 minutes at room temperature.

3B.4. Place the tube/plate on an appropriate magnetic stand to separate the beads from the supernatant. If necessary, quickly spin the sample to collect the liquid from the sides of the tube or plate wells before placing on the magnetic stand.

3B.5. After 5 minutes (or when the solution is clear), carefully remove and discard the supernatant. Be careful not to disturb the beads that contain DNA targets (Caution: do not discard the beads).

3B.6. Add 200  $\mu$ l of 80% freshly prepared ethanol to the tube/ plate while in the magnetic stand. Incubate at room temperature for 30 seconds, and then carefully remove and discard the supernatant. Be careful not to disturb the beads that contain DNA targets.

3B.7. Repeat Step 3.B.6 once for a total of two washes. Be sure to remove all visible liquid after the second wash. If necessary, briefly spin the tube/plate, place back on the magnet and remove traces of ethanol with a p10 pipette tip.

3B.8. Air dry the beads for up to 5 minutes while the tube/plate is on the magnetic stand with the lid open.

Caution: Do not over-dry the beads. This may result in lower recovery of DNA target. Elute the samples when the beads are still dark brown and glossy looking, but when all visible liquid has evaporated. When the beads turn lighter brown and start to crack they are too dry.

3B.9. Remove the tube/plate from the magnetic stand. Elute the DNA target from the beads by adding 17  $\mu$ l of 10 mM Tris-HCl or 0.1X TE.


3B.10. Mix well by pipetting up and down 10 times, or on a vortex mixer. Incubate for at least 2 minutes at room temperature. If necessary, quickly spin the sample to collect the liquid from the sides of the tube or plate wells before placing back on the magnetic stand.


3B.11. Place the tube/plate on the magnetic stand. After 5 minutes (or when the solution is clear), transfer 15  $\mu$ l to a new PCR tube.



Samples can be stored at -20°C.

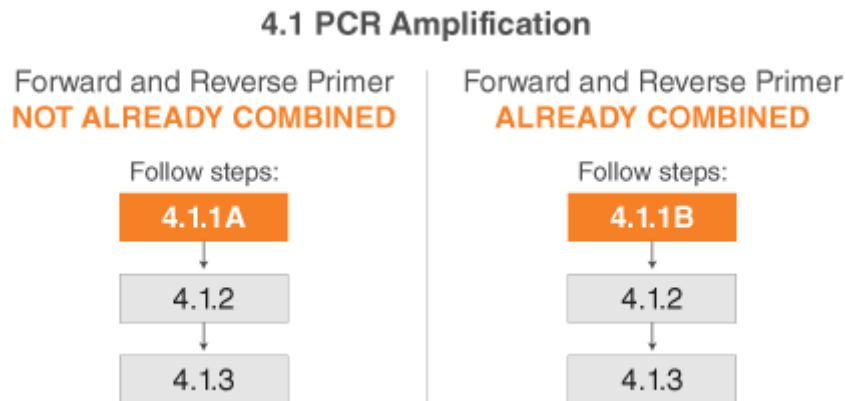
#### 4. PCR Enrichment of Adaptor-ligated DNA

 Note: Check and verify that the concentration of your oligos is 10 µM.

 Follow Section 4.1.1A if you are using the following oligos (10 µM primer):  
 NEBNext Singleplex Oligos for Illumina (NEB #E7350)  
 NEBNext Multiplex Oligos for Illumina (Set 1, NEB #E7335)  
 NEBNext Multiplex Oligos for Illumina (Set 2, NEB #E7500)  
 NEBNext Multiplex Oligos for Illumina (Set 3, NEB #E7710)  
 NEBNext Multiplex Oligos for Illumina (Set 4, NEB #E7730)  
 NEBNext Multiplex Oligos for Illumina (Dual Index Primers, NEB #E7600)




Follow Section 4.1.1B if you are using NEBNext Multiplex Oligos for Illumina (96 Index Primers, NEB #E6609)

#### 4.1 PCR Amplification



##### 4.1.1A. Forward and Reverse Primer not already combined

Add the following components to a sterile strip tube:

Adaptor Ligated DNA Fragments (Step 3.1.14 or 3.2.11)	15 µl
 (blue) NEBNext Ultra II Q5 Master Mix	25 µl
 (blue) Index Primer/i7 Primer*, **	5 µl
 (blue) Universal PCR Primer/i5 Primer*, ***	5 µl
Total volume	50 µl

\* The primers are provided in NEBNext Singleplex (NEB #E7350) or Multiplex (NEB #E7335, #E7500, #E7710, #E7730, #E7600) Oligos for Illumina. For use with Dual Index Primers (NEB #E7600), look at the NEB #E7600 manual for valid barcode combinations and tips for setting up PCR reactions.

\*\* For use with NEBNext Multiplex Oligos (NEB #E7335 or #E7500, #E7710, #E7730) use only one index primer per PCR reaction. For use with Dual Index Primers (NEB #E7600) use only one i7 primer per reaction.

\*\*\* For use with Dual Index Primers (NEB #E7600) use only one i5 Primer per reaction.

4.1.2. Set a 100 µl or 200 µl pipette to 40 µl and then pipette the entire volume up and

down at least 10 times to mix thoroughly. Perform a quick spin to collect all liquid from the sides of the tube.

4.1.3. Place the tube on a thermocycler and perform PCR amplification using the following PCR cycling conditions:

CYCLE STEP	TEMPERATURE	TIME	CYCLES
Initial Denaturation	98°C	30 seconds	1
Denaturation	98°C	10 seconds	3–15*
Annealing/Extension	65°C	75 seconds	
Final Extension	65°C	5 minutes	1
Hold	4°C	∞	

\*The number of PCR cycles should be chosen based on input amount and sample type. Thus, samples prepared with a different method prior to library prep may require re-optimization of the number of PCR cycles. The number of cycles should be high enough to provide sufficient library fragments for a successful sequencing run, but low enough to avoid PCR artifacts and over-cycling (high molecular weight fragments on Bioanalyzer). The number of PCR cycles recommended in Table 4.1 are to be seen as a starting point to determine the number of PCR cycles best for standard library prep samples. Use Table 4.2 for applications requiring high library yields (~1 µg) such as target enrichment.

4.1.1B. Forward and Reverse Primer already combined

Add the following components to a sterile strip tube:

Adaptor Ligated DNA Fragments (Step 3.1.14 or 3.2.11)	15 µl
●(blue) NEBNext Ultra II Q5 Master Mix	25 µl
●(blue) Index/Universal Primer****	10 µl
Total volume	50 µl

\*\*\*\* The primers are provided in NEBNext Multiplex Oligos for Illumina (NEB #E6609). Please refer to the NEB #E6609 manual for valid barcode combinations and tips for setting up PCR reactions.

4.1.2. Set a 100 µl or 200 µl pipette to 40 µl and then pipette the entire volume up and down at least 10 times to mix thoroughly. Perform a quick spin to collect all liquid from the sides of the tube.

4.1.3. Place the tube on a thermocycler and perform PCR amplification using the following PCR cycling conditions:

CYCLE STEP	TEMPERATURE	TIME	CYCLES
Initial Denaturation	98°C	30 seconds	1
Denaturation	98°C	10 seconds	3–15*
Annealing/Extension	65°C	75 seconds	
Final Extension	65°C	5 minutes	1
Hold	4°C	∞	

\*The number of PCR cycles should be chosen based on input amount and sample type. Thus, samples prepared with a different method prior to library prep may require re-optimization of the number of PCR cycles. The number of cycles should be high enough to provide sufficient library fragments for a successful sequencing run, but low enough to avoid PCR artifacts and over-cycling (high molecular weight fragments on Bioanalyzer). The number of PCR cycles recommended in Table 4.1 are to be seen as a starting point to determine the number of PCR cycles best for standard library prep



samples. Use Table 4.2 for applications requiring high library yields (~1 µg) such as target enrichment.

Table 4.1.

INPUT DNA IN THE END PREP REACTION	# OF CYCLES REQUIRED FOR STANDARD LIBRARY PREP ~100 ng (30-100 nM):
1 µg*	3**
500 ng*	3**
100 ng*	3
50 ng	3-4
10 ng	6-7
5 ng	7-8
1 ng	9-10
0.5 ng	10-11

\* These input ranges will work best when size selection is done  
 \*\* NEBNext adaptors contain a unique truncated design. Libraries constructed with NEBNext adaptors require a minimum of 3 amplification cycles to add the complete adaptor sequences for downstream processes.

Table 4.2.

INPUT DNA IN THE END PREP REACTION	# OF CYCLES REQUIRED FOR TARGET ENRICHMENT LIBRARY PREP (~1 µg):
1 µg*	3-4*,**
500 ng*	4-5*
100 ng*	6-7*
50 ng	7-8
10 ng	9-10
5 ng	10-11
1 ng	12-13
0.5 ng	14-15

\* Cycle number was determined for size selected libraries.  
 \*\* NEBNext adaptors contain a unique truncated design. Libraries constructed with NEBNext adaptors require a minimum of 3 amplification cycles to add the complete adaptor sequences for downstream processes.

#### 4.1.4. Proceed to Cleanup of PCR Amplification in Section 5.

### 5. Cleanup of PCR Reaction

Note: The volumes of SPRIselect or NEBNext Sample Purification Beads provided here are for use with the sample contained in the exact buffer at this step. AMPure XP Beads can be used as well. If using AMPure XP Beads, allow the beads to warm to room temperature for at least 30 minutes before use. These volumes may not work properly

for a cleanup at a different step in the workflow. For cleanups of samples contained in different buffer conditions, the volumes may need to be experimentally determined.

5.1. Vortex SPRIselect or NEBNext Sample Purification Beads to resuspend.

5.2. Add 45 µl (0.9X) resuspended beads to the PCR reaction. Mix well by pipetting up and down at least 10 times. Be careful to expel all of the liquid out of the tip during the last mix. Vortexing for 3-5 seconds on high can also be used. If centrifuging samples after mixing, be sure to stop the centrifugation before the beads start to settle out.

5.3. Incubate samples on bench top for at least 5 minutes at room temperature.

5.4. Place the tube/plate on an appropriate magnetic stand to separate the beads from the supernatant. If necessary, quickly spin the sample to collect the liquid from the sides of the tube or plate wells before placing on the magnetic stand.

5.5. After 5 minutes (or when the solution is clear), carefully remove and discard the supernatant. Be careful not to disturb the beads that contain DNA targets (Caution: do not discard the beads).

5.6. Add 200 µl of 80% freshly prepared ethanol to the tube/ plate while in the magnetic stand. Incubate at room temperature for 30 seconds, and then carefully remove and discard the supernatant. Be careful not to disturb the beads that contain DNA targets.

5.7. Repeat Step 5.6 once for a total of two washes. Be sure to remove all visible liquid after the second wash. If necessary, briefly spin the tube/ plate, place back on the magnet and remove traces of ethanol with a p10 pipette tip.

5.8. Air dry the beads for up to 5 minutes while the tube/plate is on the magnetic stand with the lid open.

Caution: Do not over-dry the beads. This may result in lower recovery of DNA target. Elute the samples when the beads are still dark brown and glossy looking, but when all visible liquid has evaporated. When the beads turn lighter brown and start to crack they are too dry.

5.9. Remove the tube/plate from the magnetic stand. Elute the DNA target from the beads by adding 33 µl of 0.1X TE.

5.10. Mix well by pipetting up and down 10 times, or on a vortex mixer. Incubate for at least 2 minutes at room temperature. If necessary, quickly spin the sample to collect the liquid from the sides of the tube or plate wells before placing back on the magnetic stand.

5.11. Place the tube/plate on the magnetic stand. After 5 minutes (or when the solution is clear), transfer 30 µl to a new PCR tube for and store at -20°C.

5.12. Check the size distribution on an Agilent Bioanalyzer High Sensitivity DNA chip. The sample may need to be diluted before loading