Title: New insights on adaptation and population structure of cork oak using genotyping by sequencing

Running head: New insights on cork oak adaptation using GBS

Authors: Pina-Martins, F.[1], Baptista, J.[2], Pappas Jr G.[3], & Paulo, O. S.[1]

1 Computational Biology and Population Genomics Group, Centre for Ecology, Evolution and Environmental Changes, Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal

2 Department of Biology, CESAM, University of Aveiro, Aveiro, Portugal

3 Department of Cell Biology, University of Brasilia, Brazil

Corresponding author: Francisco Pina-Martins – f.pinamartins@gmail.com

Keywords: Genotyping by sequencing, West Mediterranean, local adaptation, risk of non-adaptedness, association study, natural selection effects, *Quercus suber*.

Paper type: Primary Research Article

Accepted for publication in *Global Change Biology*. The final citation is:

Pina-Martins, F., Baptista, J., Pappas, G., & Paulo, O. S. (2019). New insights into adaptation and population structure of cork oak using genotyping by sequencing. *Global Change Biology*, *25*(1), 337–350. doi: 10.1111/gcb.14497

1 Abstract

Species respond to global climatic changes in a local context. Understanding this process, including its speed and intensity is paramount due to the pace at which such changes are currently occurring. Tree species are particularly interesting to study in this regard due to their long generation times, sedentarism, and ecological and economic importance. *Quercus suber* L. is an evergreen forest tree species of the Fagaceae family with an essentially Western Mediterranean distribution. Despite frequent assessments of the species' evolutionary history, large-scale genetic studies have mostly relied on plastidial markers, whereas nuclear markers have been used on studies with locally focused sampling strategies. In this work, "Genotyping by Sequencing" (GBS) is used to derive 1,996 SNP markers to assess the species' evolutionary history from a nuclear DNA perspective, gain insights on how local adaptation is shaping the species' genetic background, and to forecast how *Q. suber* may respond to global climatic changes from a genetic perspective. Results reveal (1) an essentially unstructured species, where (2) a balance between gene flow and local adaptation keeps the species' gene pool somewhat homogeneous across its distribution, but still allowing (3) variation clines for the individuals to cope with local conditions. "Risk of Non-Adaptedness" (RONA) analyses, suggest that for the considered variables and most sampled locations, (4) the cork oak should not require large shifts in allele frequencies to survive the predicted climatic changes. Future directions include integrating these results with ecological niche modelling perspectives, improving the RONA methodology and expanding its use to other species. With the implementation presented in this work, the RONA can now also be easily assessed for other organisms.

2 Introduction

Understanding how and at which rate species respond to global climatic change in their environmental context is becoming an increasingly important question due to the pace at which these are taking place (Kremer et al., 2012; Primack et al., 2009). To avoid obliteration, species may respond to such changes by either altering their distribution range, or by adapting to the new conditions. The latter can occur "instantly", due to phenotypic plasticity, or across several generations, by local adaptation (Aitken, Yeaman, Holliday, Wang, & Curtis-McLane, 2008). The kind of response species can provide is known to depend on factors like location, distribution range, and/or genetic background (Gienapp, Teplitsky, Alho, Mills, & Merilä, 2008; Ohlemuller, Gritti, Sykes, & Thomas, 2006).

Tree species are characterized by sedentarism and long lifespan and generation times, allied with generally large distribution ranges and capacity for long distance dispersal through pollen and seeds (Kremer et al., 2012). These traits make them interesting subjects to study regarding their response to global climatic changes (Thuiller et al., 2008).

In this work, we address the case of the cork oak (*Quercus suber* L.). With a distribution ranging most of the West Mediterranean region (Figure 1), this oak species is the most selective evergreen oak of the Mediterranean basin in terms of precipitation and temperature conditions (Vessella, López-Tirado, Simeone, Schirone, & Hidalgo, 2017). European oaks in particular, are known to have endured past climatic alterations, but how they can cope with the current, rapidly occurring changes is not yet fully understood (Kremer, Potts, & Delzon, 2014; Kremer et al., 2012). Despite this tree's ecological and economic importance, there is yet much to learn regarding the consequences of global climatic change on its future (Benito Garzón, Sánchez de Dios, & Sainz Ollero, 2008).

Some recent works have attempted to answer this very question, but focusing on range expansion and contraction with the assumption of a genetically homogeneous species and niche conservationism (Correia, Bugalho, Franco, & Palmeirim, 2017; Vessella et al., 2017). Both these studies also highlight the need for a genetic study regarding the adaptation potential of *Q. suber*. Unlike what happen in other oak species (Rellstab et al., 2016), studies integrating genetic information and response to climatic alterations of *Q. suber* (eg. (Modesto et al., 2014)) are rare and of small scale (Jose Alberto Ramírez-Valiente, Valladares, Huertas, Granados, & Aranda, 2011). Even though this study made the important assessement that some cork oak traits can be associated to genetic variants, its local geographic scope combined with the relatively low number if used markers, limits its utility in a distribution wide perspective. Large scale information regarding *Q. suber*'s gene flow patterns and local adaptation dynamics is paramount to understanding the species' potential to endure rapid climatic changes through adaptation (Savolainen, Lascoux, & Merilä, 2013).

In general terms, to predict a species' response to change (Kremer et al., 2012), it is fundamental to know both its genetic architecture of adaptive traits (Alberto et al., 2013) and evolutionary history (Kremer et al., 2014). However, the very nature of genetic and genomic data hampers the distinction of selection signals from other processes (McVean & Spencer, 2006), especially demographic events (Bazin, Dawson, & Beaumont, 2010). In order to disentangle population structure (mostly shaped by gene flow, inbreeding, and genetic drift) and selection (Foll, Gaggiotti, Daub, Vatsiou, & Excoffier, 2014), recent methods incorporate population structure information to detect adaptation (Gautier, 2015; Günther & Coop, 2013). Likewise, methods to accurately estimate population structure should be performed without loci known to be under selection (De Kort et al., 2014).

In non-model organisms like the cork oak, loci of adaptive value can potentially be identified by two kinds of methods – outlier analyses and environmental association analyses. While the former identify loci that depart from the expected allele frequencies as under selection (Foll & Gaggiotti, 2008; Vitalis, Gautier, Dawson, & Beaumont, 2014), they do not indicate what which loci is responding to (Gautier, 2015). The latter, while being able to associate the markers to an external covariate, are limited to detecting linear relations, and cannot assert wether or not the identified correlations are of causative nature (Gautier, 2015).

The evolutionary history of *Q. suber* has been studied in the past using multiple methodologies and in different geographic ranges. The most recent large-scale studies on the subject suggest that cork oak is divided into four strictly defined lineages (Magri et al., 2007; Simeone et al., 2009). Two of these lineages range from the south-east of France, to Morocco, including the Iberian peninsula and the Balearic Islands, a third lineage ranges from the Monaco region to Algeria and Tunisia, including the islands of Corsica and Sardinia. The fourth lineage spans the entire Italic peninsula, including Sicilia. Based only on plastidial markers, these lineages have been shown to hardly share any haplotypes (Magri et al., 2007). Notwithstanding, later works based on nuclear DNA have hinted at a different scenario, where the species is not as strictly divided (Costa et al., 2011; J. A. Ramírez-Valiente, Valladares, & Aranda, 2014). These works are, however, limited in either geographic scope or number of markers to confidently conclude that such segregation is only present in plastidial markers.

Genomic resources represent a new way to study the genetic mechanisms responsible for local adaptation (Rellstab, Gugerli, Eckert, Hancock, & Holderegger, 2015) through the use of environmental association analyses, which correlate environmental data with genetic markers, thus highlighting loci putatively involved in the adaptation process (Rellstab et al., 2016). The same methods, can thus, in principle, be used to assess the degree of maladaptation to

predicted future local conditions (Rellstab et al., 2016). The Risk of Non-Adaptedness (RONA) method was developed with this very goal (Rellstab et al., 2016). In short, for every significant association between a SNP and an environmental variable, the RONA method plots each location's individuals' allele frequencies vs. the respective environmental variable. This is done for both the current value and the future prediction. A correlation between allele frequencies and the current variable values is then calculated and the corresponding best fit line is inferred. The distance between the fitted line and the two coordinates is then compared per location and its normalized difference is considered the RONA value for each association and location (which can vary between 0 and 1). In theory, the higher the difference in conditions between the current values and the prediction, the more the studied species should have to shift its allele frequencies to survive in the location under the new conditions. Despite the innovation and importance of the method for the general scientific community, in the original paper, RONA is applied only for the work's case study (calculating RONA values for several Swiss species of Quercus based on candidate genes), and no public implementation is provided. Applying this kind of methodology to *Q. suber* would fill the gap mentioned in (Correia et al., 2017; Vessella et al., 2017), that multidisciplinary approaches are required to more accurately provide sound recommendations for the conservation of forests.

In the present work, a panel of Single Nucleotide Polymorphism (SNP) markers derived from the Genotyping by Sequencing (GBS) technique (Elshire et al., 2011) was developed to accomplish the following goals: (1) attempt to infer the species' genetic structure and evolutionary history, (2) detect signatures of natural selection, and (3) investigate the adaptation potential of *Q. suber* based on the RONA method developed and presented on (Rellstab et al., 2016).

3 Material & Methods

3.1 Sample and environmental data collection

In order to provide a comprehensive view of the species genetic background, samples were collected from 17 locations spanning most of *Q. suber*'s distribution. Fresh leaves were collected from six individuals from, *Bulgaria*, *Corsica*, *Kenitra*, *Monchique*, *Puglia*, *Sardinia*, *Sicilia*, *Tuscany*, *Tunisia* and *Var*, and from five individuals from *Algeria*, *Catalonia*, *Haza de Lino*, *Landes*, *Sintra*, *Taza* and *Toledo* for a total of 95 individuals (Table 1, Figure 1). It is worth noting that trees from Bulgaria are not of natural origin, but rather the result of human introduction from Iberian locations (Borelli & Varela, 2000; Petrov & Genov, 2004).

Most samples were collected from an international provenance trial (FAIR I CT 95 0202) established at "Monte Fava", Alentejo, Portugal (38°00' N; 8°7' W) (Varela, 2000), except Portuguese and Bulgarian samples, which were collected directly from their native locations. The collected plant material was stored at –80°C until DNA extraction.

Altitude, latitude and longitude spatial variables (Varela, 2000) were recorded for each of the native sampling sites. Nineteen Bioclimatic (BIO) variables, BIO1 to BIO19 were collected from the WorldClim database (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005) at 30 arc-seconds (~ 1 km) resolution for both "Current conditions ~1960-1990" and "Future" predictions for 2070, using two different *Representative Concentration Pathways* (RCPs), *rcp26* and *rcp85* for the following "Global Climate Models" (GCMs): BCC-CSM1-1, CCSM4, GFDL-CM3, GISS-E2-R, HadGEM2-ES, IPSL-CM5A-LR, MRI-CGCM3, MPI-ESM-LR and NorESM1-M (IPCC, 2014) as these are available under permissive licenses and calculated for both *rcp26* and *rcp85*. Instead of using the GCMs directly, an average of the values was obtained for each coordinate, and merged into a single dataset, for both used RCPs (Supporting Table 1 and 2 respectively). Data was extracted from the GeoTiff files using a

python script, *layer_data_extractor.py* (https://github.com/StuntsPT/Misc_GIS_scripts) as of commit "bd36320".

Correlations between present Bioclimatic variables were assessed using Pearson's correlation coefficient as implemented in the R script *eliminate_correlated_variables.R* (https://github.com/JulianBaur/R-scripts) as of commit "43e6553", which resulted in the exclusion of six variables due to high correlation (r>0.95). Each sampling location was thus characterized by three spatial variables and 13 environmental variables (Supporting Table 3).

3.2 Library preparation and sequencing

Genomic DNA was extracted from liquid nitrogen grounded leaves of all samples collected for this work using the kit "innuPREP Plant DNA Kit" (Analytik Jena AG), according to the manufacturer's protocol.

The total amount of extracted DNA was quantified by spectrophotometry using a Nanodrop 1000 (Thermo Scientific) and integrity verified on Agarose gel (0.8 %). DNA samples were then diluted to a concentration of ~100 ng/µl and plated for genotyping.

DNA samples were then outsourced to "Genomic Diversity Facility", at Cornell University" for genotyping using the "Genotyping by sequencing" (GBS) technique as described in (Elshire et al., 2011). Samples were shipped in a single 96 well plate with one "blank" well for negative control. Sequencing was performed according to the standard protocol on a single Illumina HiSeq 2000 flowcell using the low frequency cutter enzyme "EcoT22I", due to the large size of *Q. suber*'s genome.

3.3 Genomic data analyses

The raw GBS data was analysed using the program *ipyrad* v0.7.24, which is based on *pyrad* (Eaton, 2014), using an "*anaconda*" environment containing - *MUSCLE* v3.8.31 (Edgar,

2004) and *VSEARCH* v2.7.0 (Rognes, Flouri, Nichols, Quince, & Mahé, 2016). A *denovo* sequence assembly was performed, but mtDNA and cpDNA reads were "baited" out by *ipyrad*'s mode "denovo-reference" using the complete mitochondiral genomes of *Populus davidiana* (KY216145.1) (Choi et al., 2017) , *Pyrus pyrifolia* (KY563267.1) (Chung, Lee, Kim, & Kim, 2017) and *Rosa chinensis* (CM009589.1) (Raymond et al., 2018), and chloroplastidial genomes of *Quercus rubra* (JX970937.1) (Alexander & Woeste, 2014), *Quercus aliena* (KU240007.1) and *Quercus variabilis* (KU240009.1) (Yang et al., 2016). This ensured that mtDNA and cpDNA reads were filtered from downstream analyses. Parameters included GBS as *datatype*, *clustering threshold* of 0.85, *mindepth* of 8 and *maximum barcode mismatch* of 0. Each sampling site had to be represented by at least three individuals for a SNP to be called, except the locations of *Kenitra* and *Taza*, where only one individual was required due to the lower representation of these sampling sites. Full parameters can be found in Supporting Datafile 1. The demultiplexed "fastq" files were submitted to NCBI's Sequence Read Archive (SRA) as "BioProject" PRJNA413625.

Downstream analyses were automated using "*GNU Make*". This file, containing every detail of every step of the analyses for easier reproducibility can be found in gitlab (https://gitlab.com/StuntsPT/Qsuber_GBS_data_analyses, tag "v03"). For improved reproducibility, a docker image with all the software, configuration files, parameters and the *Makefile*, ready to use is also provided (https://hub.docker.com/r/stunts/q.suber_gbs_data_analyses/, tag "v03"). The intent is not to allow the analyses process to be treated as a "black box", but rather to provide a full environment that can be reproduced, studied and modified by the scientific community.

Processed data from *ipyrad* was then filtered using *VCFtools* v0.1.14 (Danecek et al., 2011) with the following criteria: each sample has to be represented in at least 40 % of the SNPs,

and after this each SNP has to be represented in at least 80 % of the individuals. Furthermore, due to the relatively small sample size, the minimum allele frequency (MAF) of each SNP has to be at least 0.03 for it to be retained.

In order to minimize the effects of linkage disequilibrium, downstream analyses were performed using only one SNP per locus, by discarding all but the SNP closest to the centre of the sequence in each locus. This sub dataset was obtained using the python script *vcf_parser.py* (https://github.com/CoBiG2/RAD_Tools/blob/master/vcf_parser.py) as of commit "0893296".

All file format conversions were performed using *PGDSpider* v2.1.0.0 (Lischer & Excoffier, 2012), except for the *BayPass* and *SelEstim* formats, where the scripts ***geste2baypass.py*** (https://github.com/CoBiG2/RAD_Tools/blob/master/geste2baypass.py) and *gest2selestim.sh* (https://github.com/Telpidus/omics_tools) as of commit "b99636e" and "f74f66b" respectively were used, since the used version of *PGDSpider* does not handle either of these formats.

Descriptive statistics, such as Hardy-Weinberg Equilibrium (HWE), F_{ST} and F_{IS} were calculated using *Genepop* v4.6 (Rousset, 2008). The same software was further used to perform Mantel tests to determine an eventual effect of Isolation by Distance (IBD) by correlating "'F/(1-F)'-like with common denominator" with "Ln(distance)" following on 1,000,000 permutations. This test was performed excluding individuals sampled from *Bulgaria* due to their introduced origin.

3.4 Outlier detection and environmental associations

Outlier detection was performed using two programs: *SelEstim* v1.1.4 (Vitalis et al., 2014) (50 pilot runs of length 1,000 followed by a main run of length 10^6, with a burnin of 1,000, a

thinning interval of 20, and a detection threshold of 0.01) and *BayeScan* v2.1 (Foll & Gaggiotti, 2008) (20 pilot runs of length 5,000 followed by a main run of 500,000 iterations, a burnin of 50,000, a thinning interval of 10, and a detection threshold of 0.05) (full commands and parameters are available in Supporting Datafile 2), since these methods show the lowest rate of false positives (Narum & Hess, 2011; Vitalis et al., 2014). Only SNPs indicated as outliers by both programs were considered outliers for the purpose of this work. This was done to further reduce the chance of false positives, which is a known issue in this type of analyses (Gautier, 2015; Vitalis et al., 2014).

The software *BayPass* v2.1 (Gautier, 2015) wrapped under the script *Baypass_workflow.R* (https://gitlab.com/StuntsPT/pyRona/blob/master/pyRona/R/Baypass_workflow.R) from *pyRona* v0.1.3 was used to assess associations of SNPs to environmental variables using the "AUX" model (20 pilot runs of length 1,000, followed by a main run of length 500,000 with a burnin of 5,000 and a thinning interval of 25). Any association with a Bayes Factor (BF) above 15 was considered significant. Association analyses were performed excluding individuals from *Bulgaria* sampling site for the same reasons as in the Mantel tests.

Sequences containing outlier loci or SNPs associated to an environmental variable were queried against the genome of *Q. lobata* (Sork et al., 2016) v1.0 using BLAST v2.2.28+ (Altschul et al., 1997) with an e-value threshold of 0.00001.

3.5 Population Structure

Two distinct methods were used for clustering the individuals in order to understand the general pattern of individual or population grouping, namely, Principal Components Analysis (PCA) and *MavericK* (Verity & Nichols, 2016), which is based on STRUCTURE (Pritchard, Stephens, & Donnelly, 2000).

The PCA was performed with *snp_pca_static.R* (https://github.com/CoBiG2/RAD_Tools/blob/master/snp_pca_static.R) as of commit "bb2fc45".

In order to correctly interpret clustering analyses results, it is important to estimate the value of "K", which represents how many *demes* the data can be clustered into. The software *MavericK* is especially interesting for cluster estimation due to its innovative method for estimating "K", called "Thermodynamic Integration" (TI), which has shown superior performance in this task relative to other methods (Verity & Nichols, 2016). Analysis was divided in two stages: an initial single "pilot" stage which ran for 5,000 iterations, with a *burnin* of 500 using an admixture model, a free *alpha* parameter of "1" and "thermodynamic integration" (TI) turned off. This stage was used to infer tuned *alpha* and *alphaPropSD* values which were used in the subsequent "tuned" stage as parameters for the admixture model. This stage was comprised of five runs of 10,000 iterations (10 % burnin), with TI turned on and set to 20 rungs of 10,000 samples with 20 % burnin. *MavericK* was wrapped under *Structure_threader* v 1.2.2 (Pina-Martins, Silva, Fino, & Paulo, 2016) and was run for values of "K" between 1 and 8. The most suitable value of "K" was estimated using the TI method. Full parameter files are available as Supporting Datafile 2.

The same methodology was used on two more datasets derived from the original data. On one, only SNPs considered "neutral" were used, in order to obtain an unbiased population structure (De Kort et al., 2014). On the other one, only SNPs considered "non-neutral" were used, which should not be interpreted as population structure, but rather as an indication of wether local adaptation is responsible for the observed pattern.

3.6 Risk of non-adaptedness

The software *pyRona* was developed in this work as the first public implementation of the method described in (Rellstab et al., 2016) called "Risk of non-adaptedness" (RONA). This method provides a way to represent the theoretical average change in allele frequency at loci associated with environmental variables required for any given population to cope with changes in that variable. The program source code is hosted on public repositories, under a GPLv3 license, and can be downloaded free of charge at https://gitlab.com/StuntsPT/pyRona.

PyRona has a complete user manual, with installation instructions, usage patterns, and a graphical method description.

The RONA method as implemented in *pyRona*, however, is slightly different from the original method description (Supporting Datafile 3). Namely, instead of ranking environmental factors by *p*-value of the difference test between present and future values like the original description, *pyRona* will rank the environmental factors by the number of associations. Furthermore, the average RONA value provided by *pyRona* is weighted by the R^2 value of each involved correlation, unlike the original, which uses unweighted means.

In this work, two alternative climate prediction models were used to calculate a RONA value for each location in *pyRona* v0.1.3: a low emission scenario (RCP26) and a high emission scenario (RCP85) (IPCC, 2014) in order to account for uncertainties in the models' assumptions. Any associations flagged by *Baypass* with a BF above 15 were considered relevant and included in the RONA analysis. The three non-geospatial environmental variables most frequently associated with SNPs, were selected for determining generic RONA values.

4 Results

Genotyping by sequencing (Elshire et al., 2011), a technique based on restriction enzyme genomic complexity reduction followed by short-read sequencing, was employed to discover SNP markers from a total of 95 *Q. suber* individuals sampled from 17 geographical locations (Table 1).

A total of 225,214,094 reads (100 bp) generated by the GBS assay was processed by *ipyrad* (Eaton, 2014) computational pipeline. The first analytical step consisted in the assembly of raw reads into 4,548 distinct contiguous sequence fragments (genomic loci), from which an initial set of 8,978 SNPs were flagged. Twelve *Q. suber* samples were discarded due to low sequence representation during the assembly process, resulting in the retention of 83 individuals. After filtering according to the criteria presented in the methods section 3.3, 1,996 SNPs remained, which were used for all further analyses. This filtering process additionally removed two samples which were not represented for more than 55 % of the markers, and therefore, only 81 samples were used in the analyses (Supporting Table 4).

The calculated F_{IS} values for each sampling site are available in Supporting Table 4. These range from -0.0262 (*Var*) to 0.1145 (*Puglia*) with an average value of 0.0666. Pairwise F_{ST} values are available in Supporting Table 5. These range from 0.0028 between *Sardinia* and *Tuscany* to 0.1216 between *Landes* and *Var* (average F_{ST} of 0.0541).

When looking at HWE results per marker, of the 1,996 SNPs, 172 (~9 %) reveal a heterozygote deficit, whereas 88 (~4 %) reveal a deficit of homozygotes. Individual sampling sites are comprised of two few individuals to achieve biologically meaningful results. The performed Mantel test revealed no evidence of IBD among *Q. suber* individuals.

4.1 Outlier detection and environmental association

Population differentiation and ecological association approaches (François, Martins, Caye, & Schoville, 2016) were employed aiming at the identification of loci targeted by selection. In the first strategy, highly differentiated loci among populations, measured as outliers in F_{ST} distribution, were detected by the software *BayeScan* and *SelEstim* uncovering 29 and 17 outlier SNPs respectively (Supporting Table 6). All of the loci considered under outliers by *SelEstim* were also present in the set of loci flagged as outlier by *BayeScan*. This set of 17 common markers was considered as being putatively under the effect of natural selection.

For a functional characterization of these loci, the draft genome sequence of *Q. lobata* was used as a proxy for similarity searches. None of the 17 sequences revealed significant matches to *Q. lobata*'s genome scaffolds.

The ecological association approach was carried out using the software *BayPass* and yielded 274 associations between 249 SNPs and 12 of the 16 tested environmental variables (no associations were found with "Altitude", "Temperature Annual Range", "Precipitation of Wettest Month" or "Precipitation Seasonality"). These associations can be found in Supporting Table 7. Despite this relatively high number of associations, it is important to note that 70 of these associations were between a SNP and a geospatial variable: 12 associations with "Latitude" and 58 with "Longitude". Of all environmental variables, the one with most markers associated is "Precipitation of Driest Month" with 71 associations, followed by "Isothermality" with 35 associations, and "Mean Temperature of Driest Quarter" with 29 associations.

Sequences containing 22 of the 249 markers associated with environmental variables were matched to entries in the *Q. lobata* genome, however, of these only 10 were annotated (Table 2).

The union of the outlier loci set and the set of loci associated with at least one environmental variable resulted in a dataset of 259 SNPs which were deemed "non-neutral" (7 SNPs were common to both loci sets). The remaining 1737 SNPs were grouped in another sub-dataset, deemed "neutral".

4.2 Population structure

Clustering analyses were used to infer the current population structure of *Q. suber* in the West Mediterranean. The *TI* method implemented in the software *MavericK* determined the best "K" value to be "1" on all datasets. Despite this assessment, the presented plots are always with K=2 (Figure 2), but with strong evidence that the data does not support structuring of any kind. Q-plots for values of K above 2 were always either reduced to two clusters, or to every individual being roughly equally divided into fractions of all clusters (Supporting Figure 1).

The Q-matrix plot showing the relatedness of each genotype to each considered deme of *MavericK*'s results produced using all loci (Figure 2a) can be interpreted as a rough split between western individuals (from locations *Sintra*, *Monchique*, *Kenitra*, *Toledo*, *Landes*, *Taza*, *Haza de Lino* and *Catalonia*), which are mostly, but not completely, assigned to cluster "1" and eastern ones (from locations *Var*, *Algeria*, *Sardinia*, *Corsica*, *Tunisia*, *Tuscany*, *Sicilia* and *Puglia*), which are mostly assigned to cluster "2". Individuals from *Bulgaria* are a notable exception, since individual genotypes are mostly assigned to cluster "1" similar to those of individuals from western locations, likely due to the species' introduced origin (Varela, 2000). However, this West – East split is somewhat fuzzy, as individuals' genomes are never completely attributed to a single cluster. In fact, most individuals have a considerable part of their genome attributed to both cluster "1" and "2". Furthermore, individuals from some eastern locations have their genomes almost completely attributed to cluster "1" (*Var 21*,

Corsica 3, Corsica 11, Corsica 14 and *Puglia 5*), and all individuals from *Tunisia* and *Algeria* are almost equally split between both clusters.

The Q-plot obtained using the "neutral" loci subset (Figure 2b) is nearly identical to the one with all the loci, but with individual genomes from eastern locations being slightly more assigned to cluster "1" than in Figure 2a, and can be interpreted in the same way.

The Q-plot produced using only the 259 (12.9 %) "non-neutral" loci (Figure 2c), however, does bear a different clustering pattern from the previous ones. In this case, the East – West split is more evident, as eastern individual genomes' attribution to each cluster is not as evenly split, but rather displays a more pronounced attribution to cluster "2" than in Figure 2a. The opposite is also true for western individuals, but to a lesser extent.

The PCA clustering method (largest eigenvector values of 0.0405 and 0.0299) is essentially concordant with the previous methods, revealing two loosely defined groupings (Supporting Figure 2).

4.3 Risk of non-adaptedness (RONA)

A summary of the RONA analyses for both low (RCP26) and a high (RCP85) emission scenario predictions can be found in Figure 3 and Supporting Table 8. The most represented environmental variables are "Precipitation of Driest Month" (71 SNPs, mean R^2=0.1570), "Isothermality" (35 SNPs, mean R^2=0.2143) and "Mean Temperature of Driest Quarter" (29 SNPs, mean R^2=0.1501). The values of RONA per sampling site are always higher for RCP85 than for RCP26, except for "Precipitation of Driest Month" in *Tunisia* where RCP85 has a lower RONA than RCP26, and in *Kenitra* where they are the same (the "Precipitation of Driest Month" variable in *Kenitra* is not predicted to change from current conditions of 0 mm^2 regardless of the model).

Under the RCP26 predictions, the highest RONA values for "Precipitation of Driest Month" is *Landes* (0.0369), for "Isothermality" is *Puglia* (0.0461), and for "Mean Temperature of Driest Quarter" is *Catalonia* (0.1281). Under the RCP85 predictions *Landes* presents the highest RONA for "Precipitation of Driest Month" (0.1115) and *Catalonia* presents the highest values of RONA for "Mean Temperature of Driest Quarter" (0.3888) and "Isothermality" 0.0686). It is important to note that the high RONA values of *Catalonia* are approximately twice as high as the second highest RONA value on the RCP26 prediction and close to three times as high for RCP85, marking this location as the most likely to become deprived of cork oak individuals in the future.

5 Discussion

In this study, *Quercus suber* individuals were sampled across the species' distribution range to assess population structure, impact of local adaptation and provide an estimate of the RONA value of each sampled location.

Due to the relatively large size of *Q. suber*'s genome (Zoldos, Papes, Brown, Panaud, & Siljak-Yakovlev, 1998) a genome reduction technique, GBS, was used to discover SNPs for this species. There is no "standard" parameter set to call SNPs on GBS datasets, since this will ultimately depend on the organism being studied. The stringent approach used in this study was, however, deemed preferable to alternatives that could result in more SNPs being called at the cost of lowering confidence in the called variants, eventually biasing analyses results. In fact, since no biological replicates were performed for this study, a conservative approach was always preferred as to minimize biases in the results.

After stringent quality filtering, a set of 1,996 SNPs was used in this study. This number is lower than that of some studies with similar data (Berthouly-Salazar et al., 2016), which

obtained ~22k SNPs (albeit using a more frequent cutting enzyme), but still more than (De Kort et al., 2014), which obtained 1630 SNPs, very close to that of (Escudero, Eaton, Hahn, & Hipp, 2014) and (Pais, Whetten, & Xiang, 2017). Even though this number may seem small, in the universe of *Q. suber*'s genome of ~750 Mbp, this is to date the largest number of molecular markers available for this species and represents a step forward to increase the power of population genetics studies.

5.1 Population genetic structure

Past studies (Magri et al., 2007) have characterized *Q. suber* as a highly structured species, with an evolutionary history shaped by large effect events, such as plate tectonics. These were, however, mostly based on plastidial DNA data, which is known to not always provide a comprehensive view on a species' evolutionary history (Kirk & Freeland, 2011). The nuclear markers developed for this work provide a somewhat different perspective.

Hardy & Weinberg Equilibrium analysis revealed that few individual markers deviated from expectations. Only ~9 % reveal a heterozygote deficit, and only ~4 % reveal a deficit of homozygotes. These values do not indicate the presence of assembly bias.

The obtained values of F_{IS} are higher than those of unstructured European oaks when analysed with the same type of markers, such as *Quercus robur* or *Quercus petraea* (Guichoux et al., 2013), but are nonetheless relatively low in general, which is compatible with low levels of population structuring.

Similar to what is observed with F_{IS}, F_{ST} values are on average (0.0541) higher than on the above mentioned unstructured oak species (0.0125) (Guichoux et al., 2013), but lower than other well structured trees such as eucalypti (0.095) (Cappa et al., 2013). These results corroborate what the clustering analyses reveal: an incomplete segregation of the species in

two clusters, as seen on Figure 2. Although clustering analyses using all loci do not provide a clear structuring signal (and the "TI" method clearly favours a scenario of a single large panmictic population), the produced *Q. suber* Q-plots do show some degree of segregation between western and eastern individuals. This can be derived both from Figure 2a and Figure 2b, which are, very similar, and can be interpreted in the same way – as incomplete segregation between individuals from eastern and western locations.

Figure 2c, where the Q-plot was produced using only loci putatively under selection, should not be used to infer population structure, but can be be compared to the Q-plot obtained using only "neutral" loci to interpret the role of local adaptation in shaping *Q. suber*'s genetic background. In Figure 2c, the division between western and eastern individuals is clearer than in Figure 2a and B. Furthermore, the generally observed difference pattern is similar to what cen be seen in the locations of "Monchique" and "Sardinia": individual attributions to the "dominant" cluster in the "neutral" Q-plot, become even more pronounced in the "non-neutral" Q-plot. This is expected if local adaptation is responsible for these differences (otherwise, the differences between "neutral" and "non-neutral" Q-plots should be more random). This evidence, combined with the relatively low pairwise F_{ST} and F_{IS} values, suggests a balance between local adaptation and gene flow. Whereas the former is responsible for maintaining the species' standing genetic variation across the species range and the latter for the species's response to local environmental differences. Intense gene flow would also explain the relatively low proportion of outlier SNPs, which may be counteracting reactions to weak selective pressures. At the same time, this balance may provide the species a relatively large genetic variability to respond to strong selection (De Kort et al., 2014; Kremer et al., 2012).

Data from this work does not seem to support the four lineages hypothesis proposed in (Magri et al., 2007), however, it is also not incompatible with it, if it is assumed that nuDNA and cpDNA can have different evolutionary histories. In fact, it has been argued that for other tree species plastidial lineages exist due to population contractions and expansions from glacial refugia, but high gene flow erases any evidence of their existence in the nuclear genome (Eidesen et al., 2007).

Two hypotheses can thus be proposed to explain the currently observed genetic structure:

1. Balance between gene flow and local adaptation is responsible for both creating and maintaining the current level of nuclear divergence. Whereas local adaptation tends to cause divergence between contrasting regions, this effect is countered by species wide gene flow. Population contractions in refugia locations during glacial periods explain the occurrence of plastidial lineages, which are absent in the nuclear genome due to very intense gene flow.

2. Differential hybridization of *Q. suber* with *Q. cerris* in the East (Bagnoli et al., 2016) and with *Q. ilex s.l.* in the West (Burgarella et al., 2009) is responsible for the observed nuDNA structuring pattern and balance between gene flow and local adaptation is responsible for maintaining it. Combination of these phenomena can thus be considered the cause for the observed levels of East-West differentiation. Since *Q. suber* always acts as a pollen donor in these hybridization events (Boavida, Silva, & Feijó, 2001). Under this hypothesis, *Q. suber* would maintain a high nuclear population effective, even during glacial periods, but restrict plastidial lineages' geographic scope as suggested in (López de Heredia, Carrión, Jiménez, Collada, & Gil, 2007), which is further supported by the different dispersal capabilities of pollen and acorns (Sork, 1984). This scenario would result in large effective population size

differences between nuDNA and cpDNA, which can be an alternative explanation for cpDNA lineages to simple population contractions to glacial refugia.

The proposed hypotheses are supported by the SNP data presented here, but further studies are needed to confirm them. As such, the issue will remain open for investigation.

5.2 Outlier detection and environmental association analyses

The method used to detect outlier loci flagged ~0.9 % of the total SNPs, which is in line with what was found on other similar studies (Berdan, Mazzoni, Waurick, Roehr, & Mayer, 2015; Chen et al., 2012). Of the 17 outlier markers found, none could be matched to an annotated location in *Q. lobata*'s genome. This is likely due to a combination of factors, such as the distance between *Q. suber* and *Q. lobata*, and the incomplete annotation of *Q. lobata*'s genome. On the other hand, it emphasizes the need for more genomic resources in this area, which can potentially provide important functional information of these SNPs in *Q. suber*'s genome, that will at least for now remain unknown.

The environmental association analyses (EAA) served two purposes in this work. On one hand, the reported associations work as a proxy for detecting local adaptation, and on the other hand, allow the attribution of a RONA score to each sampling site. *Q. suber* is known to be very sensitive to precipitation and temperature conditions (Vessella et al., 2017), and as such, it was expected beforehand that some of the markers obtained in this study were to be associated with some of these conditions (Rellstab et al., 2016). In order to understand how important the found associations are for the local adaptation process, it is necessary to understand the putative function of the genomic region where each SNP was found. Querying the available sequences against *Q. lobata*'s genome annotations, has provided insights regarding some of the markers' sequences putative function. The proportion of sequences that

were a match to an annotated region, however, is rather small – only ~4.4 % of the queried sequences could be matched to such regions.

Of the 10 SNPs associated with an environmental variable that returned hits to annotated regions of *Q. lobata*'s genome, two were matched to regions annotated as close to animal genes, and one matched a region annotated as a chloroplastidial region, leaving 7 SNPs as interesting to explore for downstream analyses. While all these associations are potentially interesting to explore, doing so falls outside the grander scope of this work.

Of these markers, it is interesting to remark, that SNP 158, associated with the variable "Mean Temperature of Driest Quarter", for example is located in a region annotated as "Similar to TRE1: Trehalase", which is known to play a role in drought stress (Houtte et al., 2013). Likewise, SNP 168, associated with the variable "Precipitation of Driest Month", is located in a region matching the annotation of "Similar to PER47: Peroxidase 47", which is known to play a role in drought response (Li et al., 2017).

Like these two examples, more of the SNPs found have associations to environmental variables which are putatively located in genes involved in functions which are important in responding to the very variables they are associated with. This fact flags these markers as particularly useful to focus on in downstream studies.

5.3 Risk of non-adaptedness

Although the RONA method is a greatly simplified model (its limitations are described in Rellstab et al., 2016), it provides an initial estimate of how affected *Q. suber* is likely to be by environmental changes (at least as far as the tested variables are concerned). Furthermore, it is important to remark that due to *Baypass* being limited to a univariate method, the same

525 constraint also applies to the RONA analysis, meaning that multi-loci associations are not
526 considered.

527 The implementation developed for this work, named *pyRONA* suffers from most of the same
528 limitations as the original application, even though it is based on an arguably superior
529 association detection method (Gautier, 2015), (although the original LFMM (Frichot,
530 Schoville, Bouchard, & François, 2013) method is also available to use in *pyRona* since
531 version 0.3.0) and introduces a correction to the average values based on the R^2 of each
532 marker association by using weighted means. The automation achieved by using this new
533 implementation, easily allows two different emission scenarios (RCP26 and RCP85) to be
534 tested and compared.

535 With the exception of *Catalonia*, which seems to have an exceptionally high highest RONA
536 value under both prediction models, the other locations present relatively low RONA values
537 for the tested variables. The variable "Mean Temperature of Driest Quarter" appears to be the
538 tested variable that requires the greatest changes in allele frequencies to ensure adaptation of
539 the species to the local projected changes. These RONA values, are nevertheless smaller than
540 those presented in (Rellstab et al., 2016). This might be due to various factors, such as the
541 different variables tested, the geographic scope of the study, the species' respective tolerance
542 to environmental ranges, the differences between species' standing genetic variation, the
543 association detection method, or, more likely, a combination of several of these factors.

544 Notwithstanding, the obtained results seem to indicate that *Q. suber* is generally well
545 genetically equipped to handle climatic change in most of its current distribution (with the
546 notable exception of *Catalonia*). Despite cork oak's long generation time, it seems reasonable
547 that during the considered time frame current populations are able to shift their allele
548 frequencies (2 % to 12 % on average, depending on the predictive model) due to a relatively

549 high standing genetic variation, which according to (Kremer et al., 2012) should really work
550 in the species' favour in the presence of strong selective pressures.

551 This study, however, is limited to the considered environmental variables. Other factors that
552 were not included in this work may have a larger effect on *Q. suber*'s RONA. Inferring future
553 adaptive potential of species is not yet commonplace practice (Jordan, Hoffmann, Dillon, &
554 Prober, 2017; Rellstab et al., 2016), however, combining this type of study with ecological
555 niche modelling approaches has the potential to greatly improve the accuracy of both kinds of
556 predictions.

557 5.4 Final remarks

558 In this study, new nuclear markers were developed to shed new light on *Q. suber*'s
559 evolutionary history, which is important to understand, in order to attempt to predict the
560 species response to future environmental pressures (Kremer et al., 2014).

561 Despite the relatively large geographic distances involved, the nuclear markers used in this
562 work indicate a lesser genetic structuring than previously thought from cpDNA markers,
563 which clearly segregated the species in several well defined demes (Magri et al., 2007). The
564 SNP data from this work can thus be used to propose two new hypotheses to replace the
565 current view of a deep genetic structure as evidenced by cpDNA. The observed genetic
566 structure can be explained either by balance between gene flow and local adaptation, or
567 alternatively, differential hybridization of *Q. suber* with *Q. ilex s.l.* in the West and *Q. cerris*
568 in the East being responsible for geographic differences' origin, which are then maintained by
569 the mentioned balance between gene flow and local adaptation (albeit more research is
570 required to confirm this second hypothesis).

Despite the genetic structure homogeneity, outlier and association analyses hint at the existence of local adaptation. The RONA analyses suggest that this balance, between local adaptation and gene flow, may be key in *Q. suber*'s response to climatic change. It is also worth considering that despite the species' likely capability to shift its allele frequencies for survival in the short term, the effects of such changes in the long term can be quite unpredictable (Feder, Egan, & Nosil, 2012; Lenormand, 2002), and only very recently have they began to be understood (Aguilée, Raoul, Rousset, & Ronce, 2016).

This study starts by providing a new perspective into the population genetics of *Q. suber*, and, based on this data, suggests an initial conjecture on the species' future, despite the used technique's limitations. Even though studies regarding *Q. suber*'s response to climatic change are not new (Correia et al., 2017; Vessella et al., 2017), this is the first work where this response is investigated from an adaptive perspective.

The software, *pyRona*, was developed and is provided in hopes that the method is adopted by the larger scientific community to estimate the Risk of non-Adaptedness for other species, and eventually, make an impact in determining species conservation strategies. In this regard, the RONA results can be used in informing assisted migration projects (Aitken & Bemmels, 2016). In the specific case of the cork oak, European commercial stocks can be expected to benefit from the introduction of trees (and therefore alleles) adapted to more extreme temperature and precipitation conditions. As for which ones, should be further studied, but the genes that were functionally explored in this work, should provide a good starting point.

In the near future, it is expected that improvements are made to the RONA method. In particular, using more sophisticated association testing (including the use of multivariate methods) and combining this approach with ecological niche modelling should yield much improved insights into species' response to climatic change. These changes should be

supported by expanding the use of the method to other species, which have both genetic and climatic data available.

6 Data Archiving Statement

Raw GBS data is available on NCBI's Sequence Read Archive (SRA) as "BioProject" PRJNA413625.

A docker image containing the analysis process, software and "assembled" data is available in https://hub.docker.com/r/stunts/q.suber_gbs_data_analyses/.

The software pyRona is available in gitlab, and mirrored on github.

Acknowledgements

We would like to thank R. Nunes, A. S. Rodrigues, C. Ribeiro and I. Modesto, for their help during sample collection. We would further like to thank the two anonymous reviewers for the very through feedback they have provided.

Field and laboratory work, and bioinformatics platform were supported by Fundação para a Ciência e Tecnologia (FCT) – Portugal [grant numbers SOBREIRO/0036/2009 (under the framework of the Cork Oak ESTs Consortium) and UID/BIA/00329/2013 (2015-2018)]. F. Pina-Martins was funded by FCT [grant number SFRH/BD/51411/2011 (under the PhD program "Biology and Ecology of Global Changes", Univ. Aveiro & Univ. Lisbon, Portugal)].

7 Literature cited

Aguilée, R., Raoul, G., Rousset, F., & Ronce, O. (2016). Pollen dispersal slows geographical range shift and accelerates ecological niche shift under climate change. *Proceedings of the National Academy of Sciences*, *113*(39), E5741–E5748. https://doi.org/10.1073/pnas.1607612113

Aitken, S. N., & Bemmels, J. B. (2016). Time to get moving: assisted gene flow of forest trees. *Evolutionary Applications*, *9*(1), 271–290. https://doi.org/10.1111/eva.12293

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. https://doi.org/10.1111/j.1752-4571.2007.00013.x

Alberto, F. J., Aitken, S. N., Alia, R., Gonzalez-Martinez, S. C., Hanninen, H., Kremer, A., … Savolainen, O. (2013). Potential for evolutionary responses to climate change - evidence from tree populations. *Global Change Biology*, *19*(6), 1645–1661. https://doi.org/10.1111/gcb.12181

Alexander, L. W., & Woeste, K. E. (2014). Pyrosequencing of the northern red oak (Quercus rubra L.) chloroplast genome reveals high quality polymorphisms for population management. *Tree Genetics & Genomes*, *10*(4), 803–812. https://doi.org/10.1007/s11295-013-0681-1

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Bagnoli, F., Tsuda, Y., Fineschi, S., Bruschi, P., Magri, D., Zhelev, P., … Vendramin, G. G. (2016). Combining molecular and fossil data to infer demographic history of Quercus cerris: insights on European eastern glacial refugia. *Journal of Biogeography*, *43*(4), 679–690. https://doi.org/10.1111/jbi.12673

Bazin, E., Dawson, K. J., & Beaumont, M. A. (2010). Likelihood-Free Inference of Population Structure and Local Adaptation in a Bayesian Hierarchical Model. *Genetics*, *185*(2), 587–602. https://doi.org/10.1534/genetics.109.112391

Benito Garzón, M., Sánchez de Dios, R., & Sainz Ollero, H. (2008). Effects of climate change on the distribution of Iberian tree species. *Applied Vegetation Science*, *11*(2), 169–178. https://doi.org/10.3170/2008-7-18348

Berdan, E. L., Mazzoni, C. J., Waurick, I., Roehr, J. T., & Mayer, F. (2015). A population genomic scan in Chorthippus grasshoppers unveils previously unknown phenotypic divergence. *Molecular Ecology*, *24*(15), 3918–3930. https://doi.org/10.1111/mec.13276

Berthouly-Salazar, C., Mariac, C., Couderc, M., Pouzadoux, J., Floc'h, J.-B., & Vigouroux, Y. (2016). Genotyping-by-Sequencing SNP Identification for Crops without a Reference Genome: Using Transcriptome Based Mapping as an Alternative Strategy. *Frontiers in Plant Science*, *7*, 777. https://doi.org/10.3389/fpls.2016.00777

Boavida, L. C., Silva, J. P., & Feijó, J. A. (2001). Sexual reproduction in the cork oak (<Emphasis Type="Italic">Quercus suber</Emphasis> L). II. Crossing intra- and interspecific barriers. *Sexual Plant Reproduction*, *14*(3), 143–152. https://doi.org/10.1007/s004970100100

Borelli, S., & Varela, M. C. (2000). Mediterranean Oaks Network: Report of the first meeting. In *EUFORGEN Mediterranean Oaks Network: First meeting* (p. 74). Antalya, Turkey: EUFORGEN. Retrieved from http://www.euforgen.org/publications/publication/mediterranean-oaks-network-report-of-the-first-meeting/

Burgarella, C., Lorenzo, Z., Jabbour-Zahab, R., Lumaret, R., Guichoux, E., Petit, R. J., … Gil, L. (2009). Detection of hybrids in nature: application to oaks (Quercus suber and Q. ilex). *Heredity*, *102*(5), 442–452. https://doi.org/10.1038/hdy.2009.8

Cappa, E. P., El-Kassaby, Y. A., Garcia, M. N., Acuña, C., Borralho, N. M. G., Grattapaglia, D., & Marcucci Poltri, S. N. (2013). Impacts of Population Structure and Analytical Models in Genome-Wide Association Studies of Complex Traits in Forest Trees: A Case Study in Eucalyptus globulus. *PLoS ONE*, *8*(11), e81267. https://doi.org/10.1371/journal.pone.0081267

Chen, J., Källman, T., Ma, X., Gyllenstrand, N., Zaina, G., Morgante, M., … Lascoux, M. (2012). Disentangling the Roles of History and Local Selection in Shaping Clinal Variation of Allele Frequencies and Gene Expression in Norway Spruce (Picea abies). *Genetics*, *191*(3), 865–881. https://doi.org/10.1534/genetics.112.140749

Choi, M. N., Han, M., Lee, H., Park, H.-S., Kim, M.-Y., Kim, J.-S., … Park, E.-J. (2017). The complete mitochondrial genome sequence of Populus davidiana Dode. *Mitochondrial DNA Part B*, *2*(1), 113–114. https://doi.org/10.1080/23802359.2017.1289346

Chung, H. Y., Lee, T.-H., Kim, Y.-K., & Kim, J. S. (2017). Complete chloroplast genome sequences of Wonwhang (Pyrus pyrifolia) and its phylogenetic analysis. *Mitochondrial DNA Part B*, *2*(1), 325–326. https://doi.org/10.1080/23802359.2017.1331328

Correia, R. A., Bugalho, M. N., Franco, A. M. A., & Palmeirim, J. M. (2017). Contribution of spatially explicit models to climate change adaptation and mitigation plans for a priority forest habitat. *Mitigation and Adaptation Strategies for Global Change*, 1–16. https://doi.org/10.1007/s11027-017-9738-z

Costa, J., Miguel, C., Almeida, H., Oliveira, M. M., Matos, J. A., Simões, F., … Batista, D. (2011). Genetic divergence in Cork Oak based on cpDNA sequence data. *BMC Proceedings*, *5*(Suppl 7), P13. https://doi.org/10.1186/1753-6561-5-S7-P13

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Group, 1000 Genomes Project Analysis. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

De Kort, H., Vandepitte, K., Bruun, H. H., Closset-Kopp, D., Honnay, O., & Mergeay, J. (2014). Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species Alnus glutinosa. *Molecular Ecology*, *23*(19), 4709–4721. https://doi.org/10.1111/mec.12813

Eaton, D. A. R. (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, *30*(13), 1844–1849. https://doi.org/10.1093/bioinformatics/btu121

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Eidesen, P. B., Alsos, I. G., Popp, M., Stensrud, Ø., Suda, J., & Brochmann, C. (2007). Nuclear vs. plastid data: complex Pleistocene history of a circumpolar key species. *Molecular Ecology*, *16*(18), 3902–3925. https://doi.org/10.1111/j.1365-294X.2007.03425.x

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Escudero, M., Eaton, D. A. R., Hahn, M., & Hipp, A. L. (2014). Genotyping-by-sequencing as a tool to infer phylogeny and ancestral hybridization: A case study in Carex (Cyperaceae). *Molecular Phylogenetics and Evolution*, *79*, 359–367. https://doi.org/10.1016/j.ympev.2014.06.026

Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics*, *28*(7), 342–350. https://doi.org/10.1016/j.tig.2012.03.009

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, *180*(2), 977–993. https://doi.org/10.1534/genetics.108.092221

Foll, M., Gaggiotti, O. E., Daub, J. T., Vatsiou, A., & Excoffier, L. (2014). Widespread Signals of Convergent Adaptation to High Altitude in Asia and America. *The American Journal of Human Genetics*, *0*(0). https://doi.org/10.1016/j.ajhg.2014.09.002

François, O., Martins, H., Caye, K., & Schoville, S. D. (2016). Controlling false discoveries in genome scans for selection. *Molecular Ecology*, *25*(2), 454–469. https://doi.org/10.1111/mec.13513

Frichot, E., Schoville, S. D., Bouchard, G., & François, O. (2013). Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution*, *30*(7), 1687–1699. https://doi.org/10.1093/molbev/mst063

Gautier, M. (2015). Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics*, genetics.115.181453. https://doi.org/10.1534/genetics.115.181453

Gienapp, P., Teplitsky, C., Alho, J. S., Mills, J. A., & Merilä, J. (2008). Climate change and evolution: disentangling environmental and genetic responses. *Molecular Ecology*, *17*(1), 167–178. https://doi.org/10.1111/j.1365-294X.2007.03413.x

Guichoux, E., Garnier-Géré, P., Lagache, L., Lang, T., Boury, C., & Petit, R. J. (2013). Outlier loci highlight the direction of introgression in oaks. *Molecular Ecology*, *22*(2), 450–462. https://doi.org/10.1111/mec.12125

Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, *195*(1), 205–220. https://doi.org/10.1534/genetics.113.152462

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*, *25*(15), 1965–1978. https://doi.org/10.1002/joc.1276

Houtte, H. V., Vandesteene, L., Lopez-Galvis, L., Lemmens, L., Kissel, E., Carpentier, S., … Dijck, P. V. (2013). Over-expression of the trehalase gene AtTRE1 leads to increased drought stress tolerance in Arabidopsis and is involved in ABA-induced stomatal closure. *Plant Physiology*, pp.112.211391. https://doi.org/10.1104/pp.112.211391

IPCC. (2014). Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. *IPCC AR5 Synthesis Report Website*, 151 pp.

Jordan, R., Hoffmann, A. A., Dillon, S. K., & Prober, S. M. (2017). Evidence of genomic adaptation to climate in Eucalyptus microcarpa: Implications for adaptive potential to projected climate change. *Molecular Ecology*, *26*(21), 6002–6020. https://doi.org/10.1111/mec.14341

Kirk, H., & Freeland, J. R. (2011). Applications and Implications of Neutral versus Non-neutral Markers in Molecular Ecology. *International Journal of Molecular Sciences*, *12*(6), 3966–3988. https://doi.org/10.3390/ijms12063966

Kremer, A., Potts, B. M., & Delzon, S. (2014). Genetic divergence in forest trees: understanding the consequences of climate change. *Functional Ecology*, *28*(1), 22–36. https://doi.org/10.1111/1365-2435.12169

Kremer, A., Ronce, O., Robledo-Arnuncio, J. J., Guillaume, F., Bohrer, G., Nathan, R., … Schueler, S. (2012). Long-distance gene flow and adaptation of forest trees to rapid climate change. *Ecology Letters*, *15*(4), 378–392. https://doi.org/10.1111/j.1461-0248.2012.01746.x

Lenormand, T. (2002). Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, *17*(4), 183–189. https://doi.org/10.1016/S0169-5347(02)02497-7

Li, T., Yang, H., Zhang, W., Xu, D., Dong, Q., Wang, F., … Li, C. (2017). Comparative transcriptome analysis of root hairs proliferation induced by water deficiency in maize. *Journal of Plant Biology*, *60*(1), 26–34. https://doi.org/10.1007/s12374-016-0412-x

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298–299. https://doi.org/10.1093/bioinformatics/btr642

López de Heredia, U., Carrión, J. S., Jiménez, P., Collada, C., & Gil, L. (2007). Molecular and palaeoecological evidence for multiple glacial refugia for evergreen oaks on the Iberian Peninsula. *Journal of Biogeography*, *34*(9), 1505–1517. https://doi.org/10.1111/j.1365-2699.2007.01715.x

Magri, D., Fineschi, S., Bellarosa, R., Buonamici, A., Sebastiani, F., Schirone, B., … Vendramin, G. G. (2007). The distribution of Quercus suber chloroplast haplotypes matches the palaeogeographical history of the western Mediterranean. *Molecular Ecology*, *16*(24), 5259–5266. https://doi.org/10.1111/j.1365-294X.2007.03587.x

McVean, G., & Spencer, C. C. (2006). Scanning the human genome for signals of selection. *Current Opinion in Genetics & Development*, *16*(6), 624–629. https://doi.org/10.1016/j.gde.2006.09.004

Modesto, I. S., Miguel, C., Pina-Martins, F., Glushkova, M., Veloso, M., Paulo, O. S., & Batista, D. (2014). Identifying signatures of natural selection in cork oak (Quercus suber L.) genes through SNP analysis. *Tree Genetics & Genomes*, *10*(6), 1645–1660. https://doi.org/10.1007/s11295-014-0786-1

Narum, S. R., & Hess, J. E. (2011). Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources*, *11*, 184–194. https://doi.org/10.1111/j.1755-0998.2011.02987.x

Ohlemuller, R., Gritti, E. S., Sykes, M. T., & Thomas, C. D. (2006). Quantifying components of risk for European woody species under climate change. *Global Change Biology*, *12*(9), 1788–1799. https://doi.org/10.1111/j.1365-2486.2006.01231.x

Pais, A. L., Whetten, R. W., & Xiang, Q.-Y. (Jenny). (2017). Ecological genomics of local adaptation in Cornus florida L. by genotyping by sequencing. *Ecology and Evolution*, *7*(1), 441–465. https://doi.org/10.1002/ece3.2623

Petrov, M., & Genov, K. (2004). 50 Years of cork oak (Quercus suber L.) in Bulgaria. *Forest Science*, *3*, 93–101.

Pina-Martins, F., Silva, D., Fino, J., & Paulo, O. S. (2016). Structure_threader. *Zenodo*. https://doi.org/10.5281/zenodo.57262

Primack, R. B., Ibáñez, I., Higuchi, H., Lee, S. D., Miller-Rushing, A. J., Wilson, A. M., & Silander, J. A. (2009). Spatial and interspecific variability in phenological responses to warming temperatures. *Biological Conservation*, *142*(11), 2569–2577. https://doi.org/10.1016/j.biocon.2009.06.003

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Ramírez-Valiente, J. A., Valladares, F., & Aranda, I. (2014). Exploring the impact of neutral evolution on intrapopulation genetic differentiation in functional traits in a long-lived plant. *Tree Genetics & Genomes*, *10*(5), 1181–1190. https://doi.org/10.1007/s11295-014-0752-y

Ramírez-Valiente, Jose Alberto, Valladares, F., Huertas, A. D., Granados, S., & Aranda, I. (2011). Factors affecting cork oak growth under dry conditions: local adaptation and contrasting additive genetic variance within populations. *Tree Genetics & Genomes*, *7*(2), 285–295. https://doi.org/10.1007/s11295-010-0331-9

Raymond, O., Gouzy, J., Just, J., Badouin, H., Verdenaud, M., Lemainque, A., … Bendahmane, M. (2018). The Rosa genome provides new insights into the domestication of modern roses. *Nature Genetics*, *50*(6), 772–777. https://doi.org/10.1038/s41588-018-0110-3

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, *24*(17), 4348–4370. https://doi.org/10.1111/mec.13322

Rellstab, C., Zoller, S., Walthert, L., Lesur, I., Pluess, A. R., Graf, R., … Gugerli, F. (2016). Signatures of local adaptation in candidate genes of oaks (*Quercus* spp.) in respect to present and future climatic conditions. *Molecular Ecology*. https://doi.org/10.1111/mec.13889

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Rousset, F. (2008). genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, *8*(1), 103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x

Savolainen, O., Lascoux, M., & Merilä, J. (2013). Ecological genomics of local adaptation. *Nature Reviews Genetics*, *14*(11), 807–820. https://doi.org/10.1038/nrg3522

Simeone, Cosimo, M., Papini, A., Vessella, F., Bellarosa, R., Spada, F., & Schirone, B. (2009). Multiple genome relationships and a complex biogeographic history in the eastern range of Quercus suber L. (Fagaceae) implied by nuclear and chloroplast DNA variation. *Caryologia*, *62*(3), 236–252.

Sork, V. L. (1984). Examination of Seed Dispersal and Survival in Red Oak, Quercus Rubra (Fagaceae), Using Metal-Tagged Acorns. *Ecology*, *65*(3), 1020–1022. https://doi.org/10.2307/1938075

Sork, V. L., Fitz-Gibbon, S. T., Puiu, D., Crepeau, M., Gugger, P. F., Sherman, R., … Salzberg, S. L. (2016). First Draft Assembly and Annotation of the Genome of a California Endemic Oak *Quercus lobata* Née (Fagaceae). *G3: Genes, Genomes, Genetics*, *6*(11), 3485–3495. https://doi.org/10.1534/g3.116.030411

Thuiller, W., Albert, C., Araújo, M. B., Berry, P. M., Cabeza, M., Guisan, A., … Zimmermann, N. E. (2008). Predicting global change impacts on plant species' distributions: Future challenges. *Perspectives in Plant Ecology, Evolution and Systematics*, *9*(3–4), 137–152. https://doi.org/10.1016/j.ppees.2007.09.004

Varela, M. C. (2000). *Evaluation of genetic resources of cork oak for appropriate use in breeding and gene conservation strategies*. EC FAIR Programme.

Verity, R., & Nichols, R. A. (2016). Estimating the Number of Subpopulations (K) in Structured Populations. *Genetics*, *203*(4), 1827–1839. https://doi.org/10.1534/genetics.115.180992

Vessella, F., López-Tirado, J., Simeone, M. C., Schirone, B., & Hidalgo, P. J. (2017). A tree species range in the face of climate change: cork oak as a study case for the Mediterranean biome. *European Journal of Forest Research*, 1–15. https://doi.org/10.1007/s10342-017-1055-2

Vitalis, R., Gautier, M., Dawson, K. J., & Beaumont, M. A. (2014). Detecting and Measuring Selection from Gene Frequency Data. *Genetics*, *196*(3), 799–817. https://doi.org/10.1534/genetics.113.152991

Yang, Y., Zhou, T., Duan, D., Yang, J., Feng, L., & Zhao, G. (2016). Comparative Analysis of the Complete Chloroplast Genomes of Five Quercus Species. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.00959

Zoldos, V., Papes, D., Brown, S. C., Panaud, O., & Siljak-Yakovlev, S. (1998). Genome size and base composition of seven *Quercus* species: inter- and intra-population variation. *Genome*, *41*(2), 162–168. https://doi.org/10.1139/g98-006

8 Tables

Table 1: Coordinates and number of sampled individuals for every sampling site.

Sample site	Latitude (decimal deg.)	Longitude (decimal deg.)	Number of sampled individuals
Algeria	36.5400	7.1500	5
Bulgaria	41.43	23.17	6
Catalonia	41.8500	2.5333	5
Corsica	41.6167	8.9667	6
Haza de Lino	36.8333	-3.3000	5
Kenitra	34.0833	-6.5833	6
Landes	43.7500	-1.3333	5
Monchique	37.3167	-8.5667	6
Puglia	40.5667	17.6667	6
Sardinia	39.0833	8.8500	6
Sicilia	37.1167	14.5000	6
Sintra	38.7500	-9.4167	5
Taza	34.2000	-4.2500	5
Toledo	39.3667	-5.3500	5
Tunisia	36.9500	8.8500	6
Tuscany	42.4167	11.9500	6
Var	43.1333	6.2500	6
Total:	**-**	**-**	**95**

Table 2: Summary of BLAST hits for loci with SNPs associated to one or more environmental variables. “MTDQ” and “MTWQ” stand for “Mean Temperature of Driest Quarter” and “Mean Temperature of Wettest Quarter” respectively.

SNP name	Note (Similar to)	Associations
SNP 158	TRE1: Trehalase (Arabidopsis thaliana)	Mean Temperature of Driest Quarter
SNP 168	PER47: Peroxidase 47 (Arabidopsis thaliana)	Precipitation of Driest Month
SNP 233	CPSF160: Cleavage and polyadenylation specificity factor subunit 1 (Arabidopsis thaliana)	Annual Mean Temperature
SNP 333	Ascc1: Activating signal cointegrator 1 complex subunit 1 (Mus musculus)	Mean Temperature of Driest Quarter
SNP 455	GLCAT14A: Beta-glucuronosyltransferase GlcAT14A (Arabidopsis thaliana)	Precipitation of Driest Month
SNP 619	GBP6: Guanylate-binding protein 6 (Pongo abelii)	Precipitation of Driest Month
SNP 834	NAC098: Protein CUP-SHAPED COTYLEDON 2 (Arabidopsis thaliana)	Longitude
SNP 880	TPP1: Thylakoidal processing peptidase 1%2C chloroplastic (Arabidopsis thaliana)	Mean Temperature of Warmest Quarter
SNP 1134	EMB2654: Pentatricopeptide repeat-containing protein At2g41720 (Arabidopsis thaliana)	Mean Temperature of Driest Quarter
SNP 1589	At1g19525: Pentatricopeptide repeat-containing protein At1g19525 (Arabidopsis thaliana)	Temperature Seasonality

9 Figures

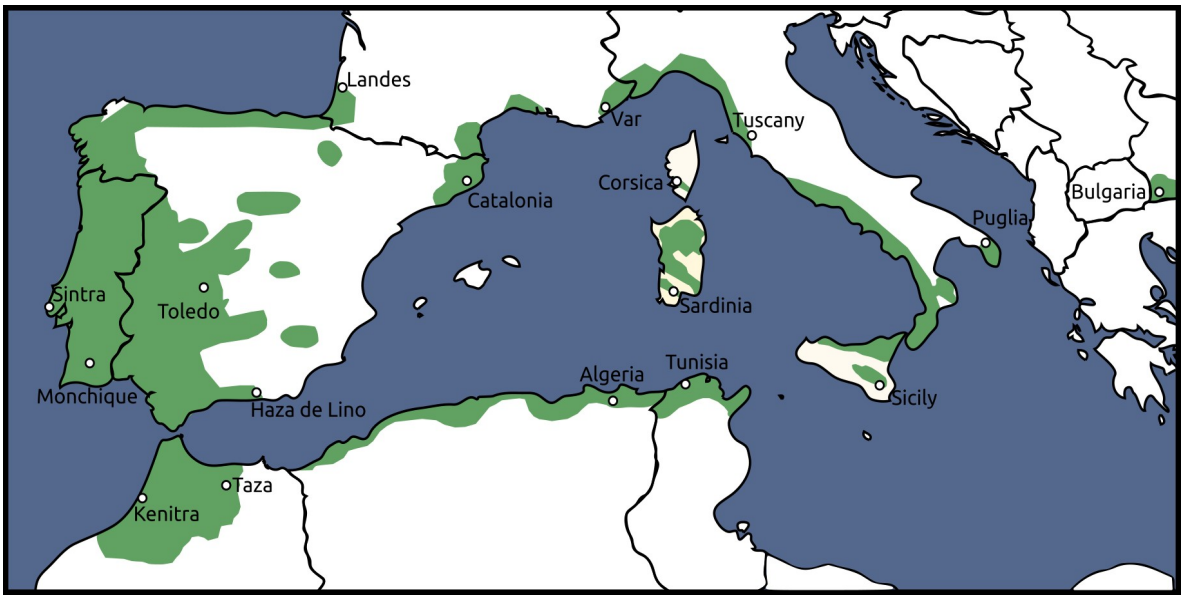

Figure 1: A map of cork oak (*Quercus suber*) distribution. Shaded land areas represent the species' range. White dots represent the sampling locations. Adapted from EUFORGEN 2009 (www.euforgen.org).

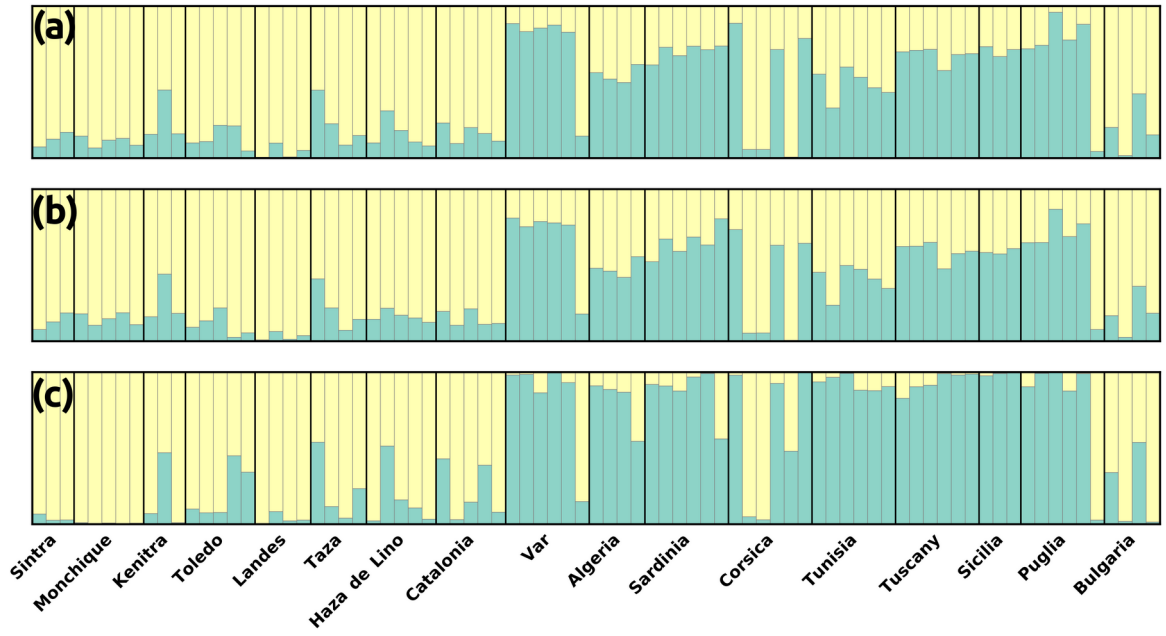

Figure 2: *MavericK* clustering plots for K=2. Sampling sites are presented from West to East. “a” is the Q-value plot for the dataset with all loci, “b” is for the dataset with only “neutral” loci, and “c” if for the dataset with only “non-neutral” loci.

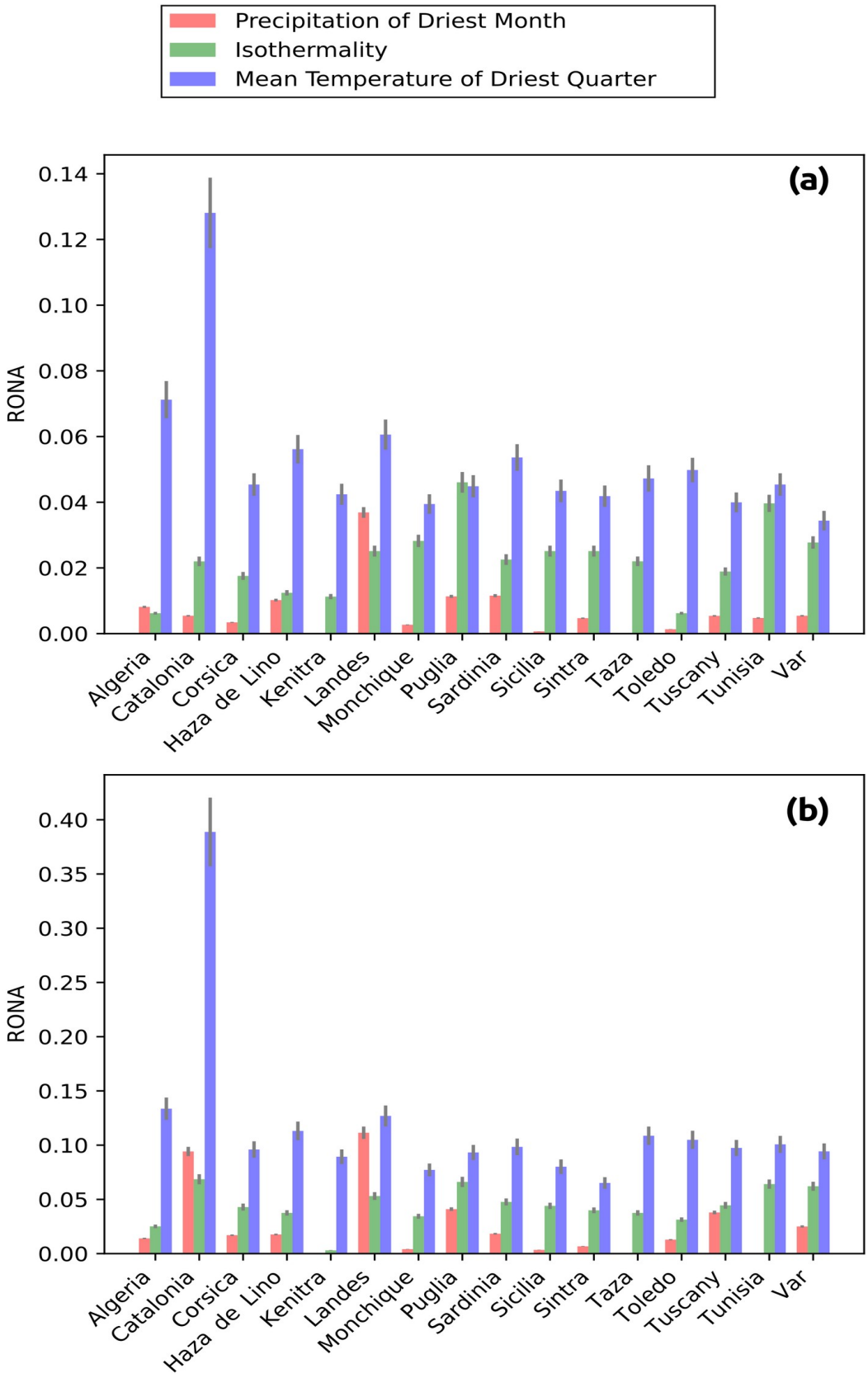

Figure 3: Risk of Non-Adaptedness plot for the three SNPs with most associations. Bars represent weighted means (by R^2 value) and lines represent standard error. (a) is the plot for RCP26 and (b) is for RCP85 prediction models.