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

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

Abstract

Species respond to global climatic changes in a local context. Understanding this process is paramount due to the pace of these changes. 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. The potential response of *Q. suber* to global climatic changes has also been studied, under ecological modelling. In this work, "Genotyping by Sequencing" (GBS) is used to derive 2,547 SNP markers to assess the species' evolutionary history from a nuclear DNA perspective, gain insights on how local adaptation may be shaping the species' genetic background, and to forecast how *Q. suber* may respond to global climatic changes from a genetic perspective. Results reveal an essentially unstructured species, where a balance between gene flow and local adaptation keeps the species' gene pool somewhat homogeneous across its distribution, but at the same time allows 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, the cork oak does not require large shifts in allele frequencies to survive the predicted climatic changes. However, more research is required to integrate these results with those of ecological modelling.

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

1 Introduction

1.1 Adaptation

Global climatic changes have been shown to cause alterations in species' traits (Benito Garzón, Alía, Robson, & Zavala, 2011; Walther et al., 2002). Understanding how species respond to such alterations in their environmental context is becoming an increasingly important question due to the pace at which they are taking place (Kremer et al., 2012; Primack et al., 2009). To avoid obliteration, species may respond to climatic changes by either altering their distribution range, effectively going extinct in the original location but persisting somewhere else, 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, 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 et al., 2012; Kremer, Potts, & Delzon, 2014). Despite this tree's

ecological and economic importance, little is known regarding the consequences of global climatic change on its future (Benito Garzón, Sánchez de Dios, & Sainz Ollero, 2008).

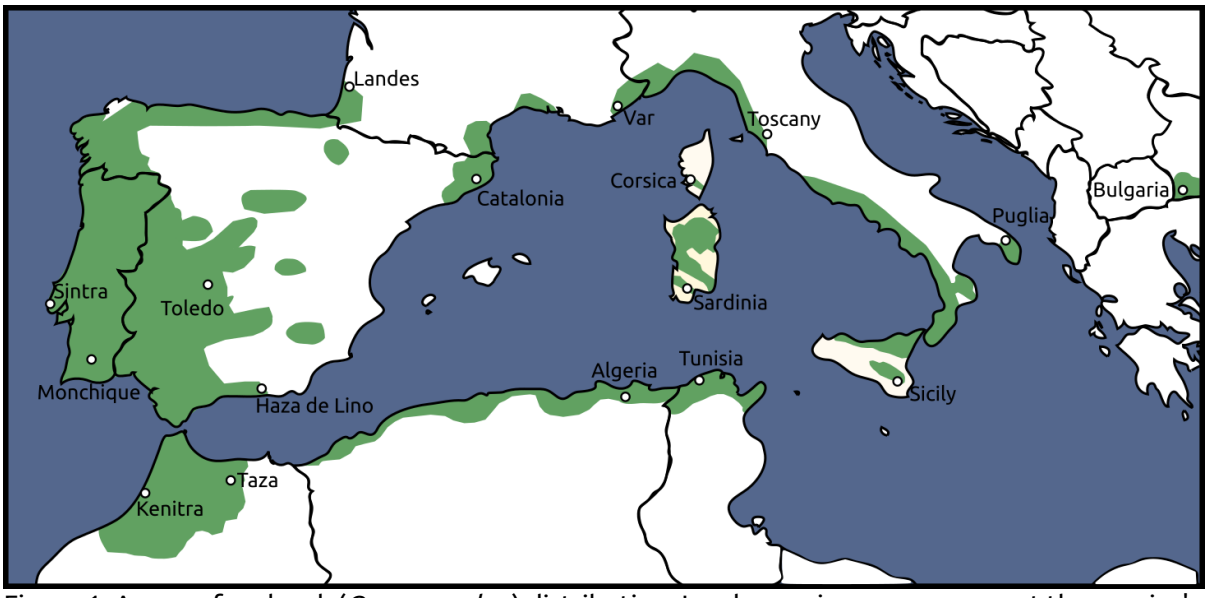

Figure 1: A map of cork oak (*Quercus suber*) distribution. Land areas in green represent the species' range. White dots represent the sampling locations. Adapted from EUFORGEN 2009 (www.euforgen.org).

Some recent works have been performed to attempt 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*. However, studies integrating genetic information and response to climatic alterations of *Q. suber* are rare and of small scale (Modesto et al., 2014) when compared with other oak species (Rellstab et al., 2016). Studies such as (Jose Alberto Ramírez-Valiente, Valladares, Huertas, Granados, & Aranda, 2011) have revealed that some traits can be associated to genetic variants, however, these were performed on a local scope and using a relatively low number of markers, which limits their utility in a larger scope. Knowing gene flow and local

adaptation dynamics of *Q. suber* is paramount to understanding the species' potential to endure rapid climatic changes through adaptation (Savolainen, Lascoux, & Merilä, 2013).

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). Applying this kind of methodology on *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. The Risk of Non-Adaptedness (RONA) method was developed in (Rellstab et al., 2016) with this very goal, however, no public implementation is provided in the mentioned work.

1.2 Population structure

In order 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 overcome the obstacles caused by the entanglement of 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).

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. 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.

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 attain 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).

2 Material & Methods

2.1 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).

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**

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 "Current conditions ~1960-1990" and "Future" predictions for 2070 (using two different *Representative Concentration Pathways* (RCPs), *rcp26* and *rcp85* conditions 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 as these are available under permissive licenses and calculated for both *rcp26* and *rcp85*). An average of the mentioned datasets was obtained for each coordinate and variable used in the analyses (Supplementary 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 (Supplementary Table 3).

2.2 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.

2.3 Genomic data analyses

The raw GBS data was analysed using the program *ipyrad* v0.5.15, which is based on *pyrad* (Eaton, 2014), using the provided "*conda*" environment - *MUSCLE* v3.8.31 (Edgar, 2004) and *VSEARCH* v2.0.3 (Rognes, Flouri, Nichols, Quince, & Mahé, 2016). Sequence assembly was performed for the GBS *datatype*, with the parameters *clustering threshold* of 0.95, *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 Supplementary 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 is presented as Supplementary Datafile 2. 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/). 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 55% 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 0.05 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 *PGDSpider* did not handle either of these formats at the time of writing.

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.

2.4 Outlier detection and environmental associations

Outlier detection was performed using two programs: *SelEstim* v1.1.4 (Vitalis, Gautier, Dawson, & Beaumont, 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 available in Supplementary 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 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://github.com/StuntsPT/pyRona/blob/master/Baypass_workflow.R) as of commit "5b406fb" 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. Similar to what was done for the Mantel tests, association analyses were performed excluding individuals from *Bulgaria* sampling site.

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.

2.5 Population Structure

Three 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), STRUCTURE (Pritchard, Stephens, & Donnelly, 2000) and *MavericK* (Verity & Nichols, 2016).

Principal Components Analysis analysis 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 STRUCTURE method was performed with STRUCTURE v2.3.4, (Pritchard et al., 2000) using the admixture model with an inferred *alpha.* To achieve best results using STRUCTURE, 20 replicates of each "K" were run at 200,000 iterations (10% burnin), and the three best values of delta K were then run for a single replicate at 2,000,000 iterations (10% *burnin*). 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). In this case, two runs were performed: an initial single "pilot" run of 5,000 iterations, with a *burnin* of 500 using an admixture model, a free *alpha* parameter of "1" and "thermodynamic integration" (TI) turned off. Tuned *alpha* and *alphaPropSD* values were extracted from the pilot run and used in the "tuned" run as parameters for the admixture model. This run 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. Both programs were wrapped under *Structure_threader* v 1.2.2 (Pina-Martins, Silva, Fino, & Paulo, 2016) for values of "K" between 1 and 8. The most suitable value of "K" was calculated using

the *evanno* (Earl & vonHoldt, 2012) and TI methods for and STRUCTURE and *MavericK* respectively. Full parameter files are available as Supplementary Datafile 2.

In order to obtain an unbiased population structure, the same methodology was used on two more datasets derived from the original data. On one, only SNPs considered outliers or that were associated with environmental variables were used ("non-neutral" dataset), and on the other one, these markers were removed ("neutral" dataset).

2.6 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 github, under a GPLv3 license, and can be downloaded free of charge at https://github.com/StuntsPT/pyRona.

In short, for every significant association between a SNP and an environmental variable, the RONA method plots each location's individuals' allele frequencies (corrected by *Baypass* to eliminate any possible effects of population structure) 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

current values and the prediction, the more *Q. suber* should have to shift its allele frequencies to survive in the location under the new conditions.

Two alternative climate prediction models were used to calculate a RONA value for each location, a low emission scenario (RCP26) and a high emission scenario (RCP85) (IPCC, 2014) in order to account for uncertainties in the models' assumptions.

The software version 0.1.3 was used and any associations flagged by *Baypass* with a BF above 15 were considered relevant and included in the RONA analysis. Results for the three most frequent non-geospatial environmental variables associated with most SNPs, were selected as the most interesting for determining generic RONA values.

3 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 7,456 distinct contiguous sequence fragments (genomic loci), from which an initial set of 12,330 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, 2,547 SNPs remained, which were used for all further analyses. This filtering process also further removed two samples due more having more than 55% missing data, and therefore, of the 83 remaining samples, only 81 were used in the analyses (Table 2).

Table 2: Number of individuals used in analysis, F_{IS} values, and Hardy-Weinberg Equilibrium (HWE) *p*-values for each sampling site.

Sample site	Number of individuals used in analysis	F_{IS}	HWE (Het. Def. P-value)	HWE (Het. Exc. P-value)
Algeria	4	0.09	**0**	1
Bulgaria	4	0.01	0.26	1
Catalonia	5	0.03	**0**	1
Corsica	6	0.1	**0**	1
Haza de Lino	5	0.04	**0**	1
Kenitra	3	0.06	**0**	1
Landes	4	-0.02	0.94	0.57
Monchique	5	0.03	**0**	1
Puglia	6	0.1	**0**	1
Sardinia	6	0.07	**0**	1
Sicilia	3	0.09	**0**	1
Sintra	3	0.09	**0**	1
Taza	4	0.09	**0**	1
Toledo	5	0.02	**0.02**	1
Tunisia	6	0.05	**0**	1
Tuscany	6	0.06	**0**	1
Var	6	0.01	**0.02**	1
Total:	**81 -**		**15**	**0**

The calculated F_{IS} values for each sampling site are available in Table 2. These range from -0.0234 (*Landes*) to 0.0987 (*Puglia*) with an average value of 0.0531. Pairwise F_{ST} values are available in Figure 2 and Supplementary Table 4. These range from 0.0038 between *Sintra* and *Monchique* to 0.1225 between *Kenitra* and *Var* (average F_{ST} of 0.0553).

Hardy-Weinberg Equilibrium tests revealed that a heterozygote deficit exists in most sampling sites (Table 2), in fact, only *Bulgaria* and *Landes* sampling sites seem not to have an excess of homozygote individuals. When looking at HWE results per marker, of the 2,547 SNPs, only 109 reveal a heterozygote deficit, whereas 23 reveal a deficit of homozygotes. Performing the same test on all individuals as a single large population also revealed a deficit of heterozygotes. The performed Mantel test revealed no evidence of IBD among *Q. suber* individuals.

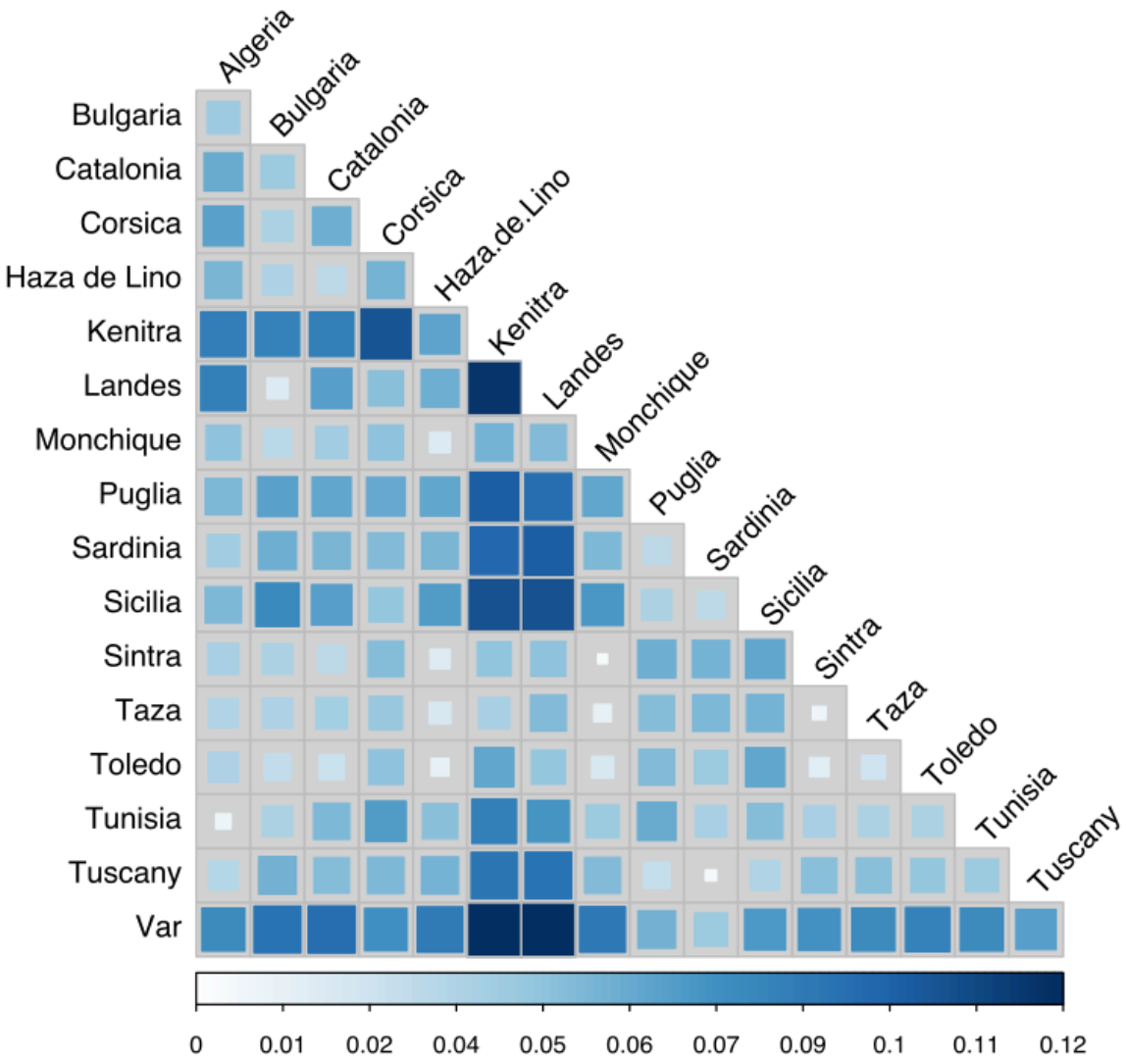

Figure 2: Pairwise F_{ST} plot between sampling sites. Darker blue represents a higher pairwise F_{ST} value, and lighter blue represents a lower value.

3.1 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 32 and 48 outlier SNPs respectively (Supplementary Table 5). Most of the loci considered under outliers by *BayeScan* were also present in the set of loci flagged as

outlier by *SelEstim*. This set of 31 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. Ten of the 31 sequences revealed significant matches to *Q. lobata* genome scaffolds. Of these, seven were not annotated, and four could be matched to an annotated region (Table 3).

Table 3: Summary of best BLAST hit results for loci with SNPs considered outliers against the genome of *Q. lobata*.

SNP name	Scaffold name	Seq. length	evalue	Identity %	Match length	Scaffold start	Scaffold end	Conserved protein domain	Description (Similar to)	GO term
SNP 37	scaffold3209	51	8.2E-13	49	53	56022	56074	InterPro:IPR015590, Pfam:PF00171	Aldehyde dehydrogenase family 7 member A1 (*Malus domestica*)	GO:0008152, GO:0016491, GO:0055114
SNP 490	scaffold1024	65	2.1E-20	60	63	161458	161396	InterPro:IPR018108, Pfam:PF00153	At3g20240: Probable mitochondrial adenine nucleotide transporter BTL1 (*Arabidopsis thaliana*)	
SNP 497	scaffold8324	52	1.1E-16	49	51	23136	23086	InterPro:IPR000916, Pfam:PF00407	NCS2: S-norcoclaurine synthase 2 (*Papaver somniferum*)	GO:0006952, GO:0009607
SNP 1896	scaffold1118	103	5E-44	100	103	250562	250664	InterPro:IPR002100, InterPro:IPR002487, Pfam:PF00319, Pfam:PF01486	AGL104: Agamous-like MADS-box protein AGL104 (*Arabidopsis thaliana*)	GO:0003677, GO:0003700, GO:0005634, GO:0006355, GO:0046983

The ecological association approach was carried out using the software *BayPass* and yielded 374 associations between 329 SNPs and 14 of the 16 tested environmental variables (no associations were found with neither “Temperature Annual Range” nor “Precipitation Seasonality”). These associations can be found in Supplementary Table 6. Despite this relatively high number of associations, it is important to note that 72 of these associations were between a SNP and a geospatial variable: 9 associations with “Latitude”, 55 with “Longitude” and 8 with “Altitude”. Of all environmental variables, the one with most markers associated is “Precipitation of Driest Month” with 79

associations, followed by “Mean Temperature of Driest Quarter” with 51 associations, and “Temperature Seasonality” with 33 associations.

Sequences containing 144 of the 329 markers associated with environmental variables were matched to entries in the *Q. lobata* genome, however, of these only 47 were annotated (Table 4).

The union of the outlier loci set and the set of loci associated with at least one environmental variable resulted in a dataset of 341 SNPs which were deemed “non-neutral” (19 SNPs were common to both loci sets). The remaining 2206 SNPs were grouped in another sub-dataset, deemed “neutral”.

3.2 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. The classic method for the STRUCTURE software, the *evanno* method revealed that K=2 had the best ΔK, followed by K=3 and K=4 on all datasets. It is, however, important to note that the *evanno* method is not able to evaluate the ΔK value for K=1. Despite this assessment, the presented plots are always with K=2, but with strong evidence that that the most likely scenario is that there is no structuring of any kind.

Table 4: 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 37	Aldehyde dehydrogenase family 7 member A1	Longitude	
SNP 70	PAT23: Probable protein S-acyltransferase 23	Longitude	
SNP 76	UGT74E2: UDP-glycosyltransferase 74E2	MTWQ	
SNP 346	KINESIN-13A: Kinesin-13A	Longitude	
SNP 442	tea1: Tip elongation aberrant protein 1	MTDQ	
SNP 490	At3g20240: Probable mtDNA adenine nucleotide transporter BTL1	Precip. of Driest Month	
SNP 497	NCS2: S-norcoclaurine synthase 2	MTDQ	
SNP 513	LTA3: Dihydrolipoyllysine-residue acetyltransferase component 1 of pyruvate dehydrogenase complex mtDNA	Longitude	
SNP 545	AVT1: Vacuolar amino acid transporter 1	Precip. of Driest Month	
SNP 618	TCTP: Translationally-controlled tumor protein homolog	MTDQ	
SNP 626	At4g13010: Putative quinone-oxidoreductase homolog cpDNA	Isothermality	
SNP 638	FAAH: Fatty acid amide hydrolase	Annual Mean Temp.	MTDQ
SNP 690	At1g22950: Uncharacterized PKHD-type hydroxylase At1g22950	Precip. of Driest Month	
SNP 892	ARGF: Ornithine carbamoyltransferase cpDNA	MTDQ	
SNP 896	PIR: Protein PIR	Isothermality	
SNP 910	FOLD1: Bifunctional protein FolD mtDNA	MTDQ	
SNP 975	LPP2: Lipid phosphate phosphatase 2	Annual Mean Temp.	MTWQ
SNP 985	NUDT8: Nudix hydrolase 8	Longitude	
SNP 1267	RABH1B: Ras-related protein RABH1b	MTDQ	
SNP 1279	NPF4.6: Protein NRT1/ PTR FAMILY 4.6	Precip. of Wettest Month	
SNP 1317	FH20: Formin-like protein 20	Precip. of Driest Month	
SNP 1381	ATG18F: Autophagy-related protein 18f	Min Temp. of Coldest Month	
SNP 1391	BETAC-AD: Beta-adaptin-like protein C	MTDQ	
SNP 1515	C7-dimethyl-8-ribityllumazine synthase cpDNA	Mean Temp. of Warmest Quarter	
SNP 1568	yipf6: Protein YIPF6 homolog	Latitude	
SNP 1621	At5g10080: Aspartic proteinase-like protein 1	Min Temp. of Coldest Month	
SNP 1645	ERDJ3A: DnaJ protein ERDJ3A	Latitude	
SNP 1663	PIGS: GPI transamidase component PIG-S	Altitude	Annual Mean Temp.
SNP 1680	66 kDa stress protein	Isothermality	
SNP 1733	SBT5.4: Subtilisin-like protease SBT5.4	Precip. of Driest Month	
SNP 1742	MCM8: Probable DNA helicase MCM8	MTDQ	
SNP 1748	ATOBGM: Probable GTP-binding protein OBGM mtDNA	Precip. of Driest Month	
SNP 1774	LDL2: Lysine-specific histone demethylase 1 homolog 2	Isothermality	
SNP 1779	VIT_19s0014g04930:	MTWQ	
SNP 1922	Stearoyl-[acyl-carrier-protein] 9-desaturase cpDNA	Isothermality	
SNP 1959	Tbc1d15: TBC1 domain family member 15	Annual Precip.	
SNP 1982	ALDH3F1: Aldehyde dehydrogenase family 3 member F1	Longitude	
SNP 2068	CAJ1: Protein CAJ1	Mean Diurnal Range	
SNP 2213	PAT04: Probable protein S-acyltransferase 4	Mean Diurnal Range	
SNP 2253	APK1B: Protein kinase APK1B cpDNA	Temp. Seasonality	
SNP 2272	UPL4: E3 ubiquitin-protein ligase UPL4	MTDQ	
SNP 2282	Os04g0338000: Probable aldo-keto reductase 2	Precip. of Driest Month	
SNP 2361	CRS1: cpDNA group IIA intron splicing facilitator CRS cpDNA	Precip. of Driest Month	
SNP 2413	At1g11300: G-type lectin S-receptor-like serine/threonine-protein kinase At1g11300	Longitude	
SNP 2525	XYL1: Alpha-xylosidase 1	Isothermality	
SNP 2539	TIG: Trigger factor-like protein TIG cpDNA	Temp. Seasonality	Annual Precip.
SNP 2540	At1g54610: Probable serine/threonine-protein kinase At1g54610	MTDQ	Precip. of Driest Month

The Q-matrix plot showing the relatedness of each genotype to each considered deme of *MavericK*'s results produced using all loci (Figure 3A) 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 assigned to cluster "1" and Eastern ones (from locations *Var*, *Algeria*, *Sardinia*, *Corsica*, *Tunisia*, *Tuscany*, *Sicilia*, *Puglia* and *Bulgaria*), 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 (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 mostly attributed to cluster "1" (*Var 21*, *Corsica 3, Corsica 11, Corsica 14* and *Puglia 5*), and individuals from *Tunisia* are almost equally split between both clusters.

The Q-plot obtained using the "neutral" loci subset (Figure 3B) is nearly identical to the one with all the loci, and can be interpreted in the same way.

The Q-plot produced using only the 305 (13.4%) "non-neutral" loci (Figure 3C), however does bear a different clustering pattern from the previous ones. In this case, the East – West split is more evident, as individual genomes' attribution to each cluster is not as evenly split, but rather displays a more pronounced individual genome attribution to either cluster.

The Q-plot obtained from STRUCTURE (Supplementary Figure 1) reveals a similar pattern to that of *MavericK* on all datasets.

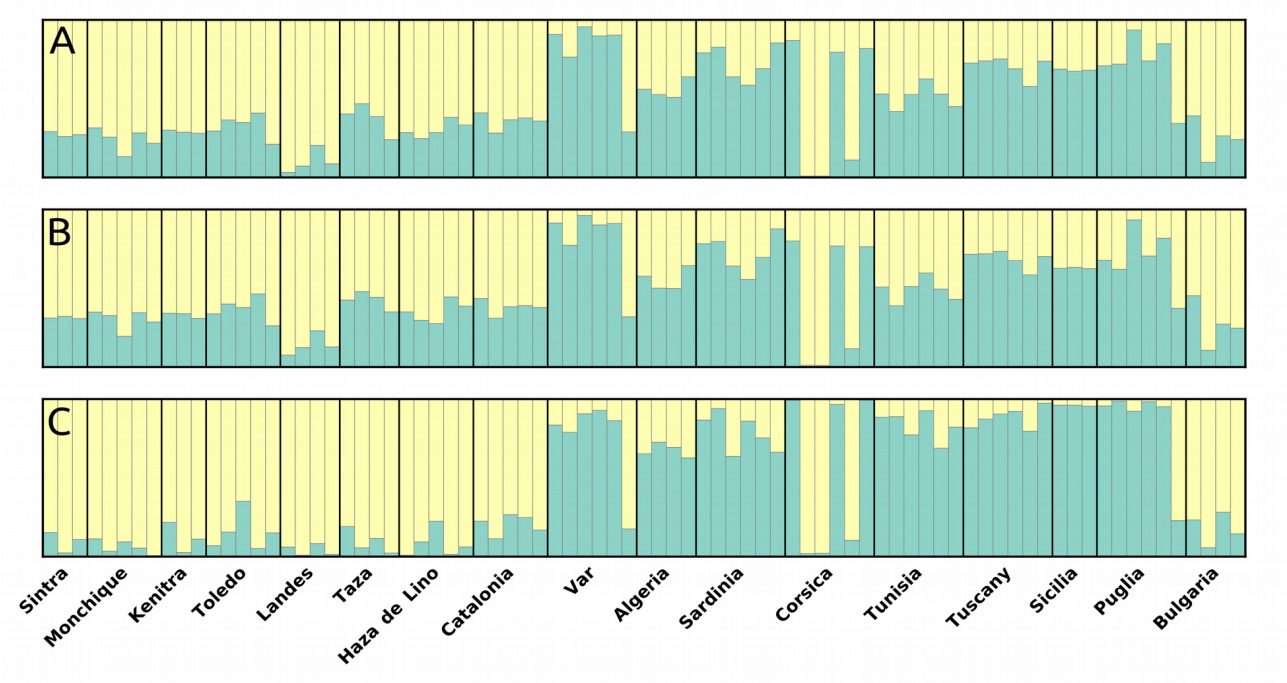

Figure 3: *MavericK* clustering plots for K=2. Sampling sites are presented from West to East. "A" is the Q-value plot for the dataset with all loci, "B" is for the dataset with only "neutral" loci, and "C" if for the dataset with only "non-neutral" loci.

The PCA clustering method (largest eigenvector values of 0.0431 and 0.0241) is essentially concordant with the previous methods, revealing two loosely defined groupings (Supplementary Figure 2). The first group containing individuals from *Algeria*, *Corsica*, *Puglia*, *Sardinia*, *Sicilia*, *Tuscany*, *Tunisia* and *Var* and the second group containing individuals from *Bulgaria*, *Catalonia*, *Corsica*, *Haza de Lino*, *Kenitra*, *Landes*, *Monchique*, *Puglia*, *Sintra*, *Taza*, *Toledo* and *Var*. The groups are loosely defined, because they somewhat resemble an East – West split, but individuals from *Corsica*, *Puglia* and *Var* are present in both groups. Just as in the Q-plots, Bulgarian individuals group with Western ones, despite existing on the edge of the species' Eastern range. Finally, a less pronounced sub-grouping is discernible: one comprising three individuals from *Corsica*; a second comprising all *Landes* individuals, plus three individuals from *Bulgaria*; and a third sub-group consisting of two individuals from *Puglia* and three from *Var*.

3.3 Risk of non-adaptedness (RONA)

A summary of the RONA analyses for both a low emission scenario (RCP26) and a high emission scenario (RCP85) predictions can be found in Figure 4 and Supplementary Table 7. The most represented environmental variables are “Precipitation of Driest Month” (79 SNPs, mean R^2=0.1597), “Mean Temperature of Driest Quarter” (51 SNPs, mean R^2=0.1466) and “Temperature Seasonality” (33 SNPs, mean R^2=0.1545). 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 (0 mm^2), regardless of the model).

Under the RCP26 predictions, the highest RONA values for “Mean Temperature of Driest Quarter” is *Landes* (0.1482), for “Temperature Seasonality” is *Toledo* (0.0690) and for “Precipitation of Driest Month” is *Landes* (0.0356). Under the RCP85 predictions, *Catalonia* presents the highest values of RONA for “Mean Temperature of Driest Quarter” (0.3921), *Landes* presents the highest RONA for “Precipitation of Driest Month” (0.1157), whereas *Toledo* has the highest value (0.1478) for “Temperature Seasonality”. It is important to note that the high RONA values of *Catalonia* are twice as high as the second highest RONA value on the RCP26 prediction and more than three times as high for RCP85.

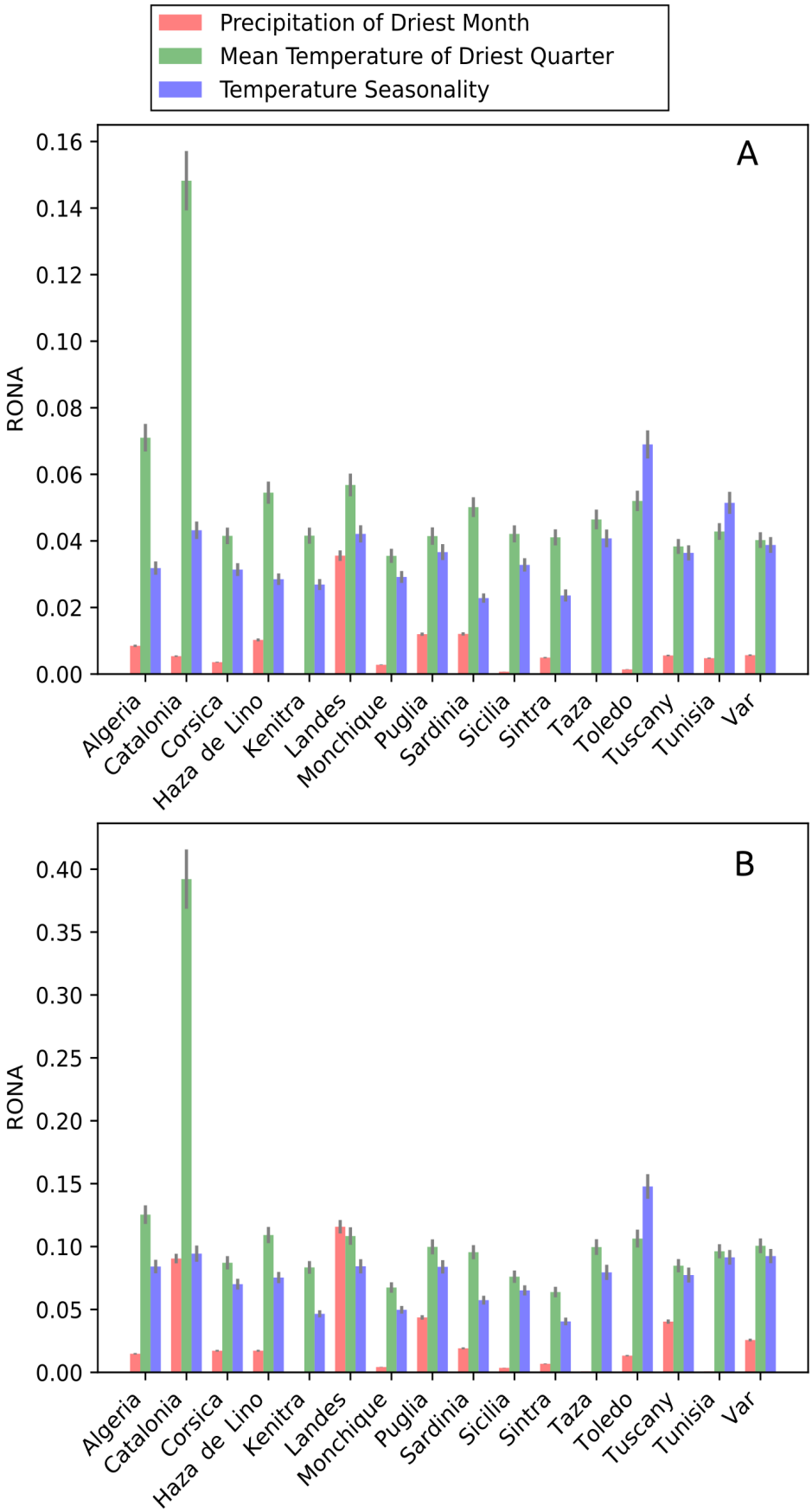

Figure 4: Risk of Non-Adaptedness plot for the three SNPs with most associations. Bars represent weighted means (by R^2 value) and lines represent standard error. (A) is the plot for RCP26 and (B) is for RCP85 prediction models.

4 Discussion

In this study, *Quercus suber* individuals were sampled across the species' distribution range to assess the 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 2,547 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.

4.1 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.

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.

Only two sampling sites did not reveal significant deviations from HWE (*Bulgaria* and *Landes*) regarding heterozygote deficiency. No sampling site exhibited heterozygote excess. Although this pattern is not usual, few individual markers deviate from HWE (4.28% reveal excess heterozygotes and 0.90% deficit heterozygotes). This may be due to the fact that each sampling site does not represent a real biological population (albeit the same pattern arises when all individuals are merged into a single group, or in Eastern and Western groups), or to non random mating across the species distribution range.

Similar to what is observed with F_{IS}, F_{ST} values are on average (0.0553) higher than on unstructured trees species (0.0125) (Guichoux et al., 2013), but lower than other well structured trees such as eucalypts (0.095) (Cappa et al., 2013). This data supports what the clustering analyses reveal: an incomplete segregation in two clusters, as seen on Figure 3. 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, which could hint at some form of local adaptation.

A comparative Q-plot analysis between "neutral" and "non-neutral", however, reveals the most contrasting differences regarding *Q. suber*'s population structure.

In Figure 3C, where the Q-plot was produced using only loci putatively under selection, the division between Western and Eastern individuals is clearer than in Figure 3A and Figure 3B. Conversely, in Figure 3B, which was drawn based on loci deemed "neutral", a pattern very similar to the Q-plots of all loci emerges, which supports a scenario of an incomplete segregation between individuals from Eastern and Western locations. 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). It could be argued that these plastidial lineages exist due to population contractions and expansions from glacial refugia, but a high gene flow would have erased any evidence of their existence in the nuclear genome, as is thought to have occurred in other tree species (Eidesen et al., 2007). However it seems just as likely to assume a scenario without refugia, where *Q. suber* maintained a relatively large population effective, even during glacial periods, and the cpDNA segregation is a consequence of different dispersal capacities of pollen and acorns (Sork, 1984).

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, species wide gene flow counters this with an homogenising effect. Population contractions in refugia locations during glacial periods explain both the plastidial lineages and, to some extent, the small difference between Eastern and Western locations. (2) In this scenario, balance between gene flow and adaptation is responsible for maintaining the current genetic pattern, but not for their origin, which is explained by 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). Combination of the two above phenomena is the cause for the small East-West differentiation. 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), since *Q. suber* always acts as a pollen donor in these hybridization events (Boavida, Silva, & Feijó, 2001). This would cause a particularly large difference in effective population size between nuDNA and cpDNA, which explains why cpDNA can be divided into lineages – random mutations crop up in different geographic locations and are differentially maintained by drift. The SNP data is compatible with both hypothesis, but not sufficient to confirm any of them, and as such, the issue will remain open for investigation.

4.2 Outlier detection and environmental association analyses

The method used to detect outlier loci flagged ~1.2% 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 31 outlier markers found, only four had a match to an annotated location in *Q. lobata*'s genome. This low proportion 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. Of particular note is SNP 493, whose sequence is a match to a region of the *Q. lobata* genome, annotated as "Similar to NCS2: S-norcoclaurine synthase 2 (*Papaver somniferum*)", a protein family member usually expressed upon infections and stressful conditions (van Loon, Rep, & Pieterse, 2006). This can be a particularly interesting marker for downstream studies regarding adaptation to infection response.

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 ~14.3% of the queried sequences could be matched to such regions.

Of the 47 SNPs associated with an environmental variable that returned hits to annotated regions of *Q. lobata*'s genome, four are likely located in a mitochondrial region, seven in chloroplastidial regions, and 36 in nuclear regions. While all these associations are potentially interesting to explore, doing so falls outside the grander

scope of this work. Nevertheless, 6 SNPs are particularly interesting to take a closer look at, mostly due to how much information is available regarding the identified genomic region function.

In addition to being identified as an outlier, SNP 497 is also associated with the variable "Mean Temperature of Driest Quarter". It is interesting to assess that a marker located in a genetic region known to be expressed during stressful conditions is associated with an environmental variable that cork oak is known to be sensitive to. This makes SNP 497 a very interesting candidate for downstream studies.

SNP 638 is located in a sequence annotated as "Similar to FAAH: Fatty acid amide hydrolase". This is a family of proteins that are known to play a role in the transport of fixed nitrogen from bacteroids to plant cells in symbiotic nitrogen metabolism (Shin et al., 2002). *Q. suber* is known to have symbiotic associations with mycorrhizae (Sebastiana et al., 2014) and the association of this marker with both "Annual Mean Temperature" and "Mean Temperature of Driest Quarter" can lead to important findings on downstream studies.

SNP 1621 and SNP 1733 are located in sequences that matched regions whose annotation indicates they may be involved in pathogen defence signalling (Figueiredo, Monteiro, & Sebastiana, 2014; Xia et al., 2004). The matched annotations are "Similar to At5g10080: Aspartic proteinase-like protein 1" and "Similar to SBT5.4: Subtilisin-like protease SBT5.4" respectively. SNP 1621 is associated with the variable "Min Temperature of Coldest Month", and SNP 1733 is associated with "Precipitation of driest month". Like the above, these markers can be potentially very interesting for downstream analyses regarding pathogen response.

SNP 1645 is located in a sequence that matched a region annotated as "Similar to ERDJ3A: DnaJ protein ERDJ3A". This protein is known to play a role in pollen tube formation during heat stress (Yang et al., 2009). In this case, the maker is associated with "Latitude", which might be working as a proxy for some temperature related variable that was not used in this study.

The sequence where SNP 2272 is found can be matched to a region annotated as "Similar to UPL4: E3 ubiquitin-protein ligase UPL4". This family of proteins is known to be involved in leaf senescence processes (Miao & Zentgraf, 2010). Its association with "Mean Temperature of Driest Quarter" makes SNP 2272 a good candidate for downstream research regarding *Q. suber*'s leaf development.

4.3 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). The implementation developed for this work, named *pyRONA* suffers from most of the same limitations as the original application, even though it is based on an arguably superior association detection method (Gautier, 2015), but introduces a correction to the average values based on the R^2 of each marker association (by using weighted means). The automation brought by this new implementation, easily allows two different emission scenarios (RCP26 and RCP85) to be tested and compared.

With the exception of *Catalonia*, which seems to have an exceptionally high highest RONA value under both prediction models, the other locations present relatively low RONA values for the tested variables. The variable "Mean Temperature of Driest Quarter" appears to be the tested variable that requires the greatest changes in allele

frequencies to ensure adaptation of the species to the local projected changes, although "Temperature Seasonality" is not far behind. These RONA values, are nevertheless smaller than those presented in (Rellstab et al., 2016). This might be due to various factors, such as the different variables tested, the geographic scope of the study, the species' respective tolerance to environmental ranges, the differences between species' standing genetic variation, the association detection method, or likely a combination of several of these factors.

Notwithstanding, the obtained results seem to indicate that *Q. suber* is generally well genetically equipped to handle climatic change in most of its current distribution (with the notable exception of *Catalonia*). Despite cork oak's long generation time, it seems reasonable that during the considered time frame current populations are able to shift their allele frequencies (2% to 10% on average, depending on the predictive model) due to the species relatively high standing genetic variation, which according to (Kremer et al., 2012) should really work in the species' favour in the presence of strong selective pressures.

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

5 Conclusions

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

Despite the relatively large geographic distances involved, the nuclear markers used in this work indicate lesser genetic structuring than previously thought from cpDNA markers, that clearly segregated the species in several well defined demes (Magri et al., 2007). The SNP data from this work can thus be used to propose two new hypotheses to replace the current view of a genetic structure carved by population recessions and expansions from glacial refugia. The observed genetic structure origin and maintenance can be explained either by balance between gene flow and local adaptation, or alternatively, differential hybridization of *Q. suber* with *Q. ilex s.l.* in the West and *Q. cerris* in the East is responsible for the geographic differences, which are then maintained by the mentioned balance between gene flow and local adaptation (albeit more research is 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 the *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 being 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. One aspect that could thoroughly improve its reliability would be the availability of more genomic resources, especially a thoroughly annotated genome of the species. Such resource would allow the identification of more markers, and assess the reliability of more associations, which would also allow a more refined method for assessing which loci are more likely to be under the effects of selection. Fortunately, such efforts are underway, and further work in this area should benefit from it in the near future.

6 Acknowledgements

We would like to thank R. Nunes, A. S. Rodrigues, C. Ribeiro and I. Modesto, for their help during sample collection. Funding was provided by projects SOBREIRO/0036/2009 (under the framework of the Cork Oak ESTs Consortium) and UID/BIA/00329/2013 from Fundação para a Ciência e Tecnologia (FCT) – Portugal. F. Pina-Martins was funded by FCT grant SFRH/BD/51411/2011, under the PhD program "Biology and Ecology of Global Changes", Univ. Aveiro & Univ. Lisbon, Portugal.

7 References

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. doi:10.1073/pnas.1607612113

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. doi: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. doi:10.1111/gcb.12181

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. doi: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. doi:10.1534/genetics.109.112391

Benito Garzón, M., Alía, R., Robson, T. M., & Zavala, M. A. (2011). Intra-specific variability and plasticity influence potential tree species distributions under climate change: Intra-specific variability and plasticity. *Global Ecology and Biogeography*, *20*(5), 766–778. doi:10.1111/j.1466-8238.2010.00646.x

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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi:10.1534/genetics.112.140749

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. doi: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. doi: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. doi: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. doi:10.1111/mec.12813

Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*(2), 359–361. doi:10.1007/s12686-011-9548-7

Eaton, D. A. R. (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, *30*(13), 1844–1849. doi: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. doi: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. doi: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. doi: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. doi: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. doi:10.1016/j.tig.2012.03.009

Figueiredo, A., Monteiro, F., & Sebastiana, M. (2014). Subtilisin-like proteases in plant–pathogen recognition and immune priming: a perspective. *Frontiers in Plant Science*, *5*. doi:10.3389/fpls.2014.00739

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. doi: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). doi: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. doi:10.1111/mec.13513

Gautier, M. (2015). Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics*, genetics.115.181453. doi: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. doi: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. doi:10.1111/mec.12125

Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, *195*(1), 205–220. doi: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. doi:10.1002/joc.1276

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. doi: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. doi: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. doi: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. doi: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. doi:10.1016/S0169-5347(02)02497-7

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. doi: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. doi: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. doi: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. doi:10.1016/j.gde.2006.09.004

Miao, Y., & Zentgraf, U. (2010). A HECT E3 ubiquitin ligase negatively regulates Arabidopsis leaf senescence through degradation of the transcription factor WRKY53. *The Plant Journal*, *63*(2), 179–188. doi:10.1111/j.1365-313X.2010.04233.x

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. doi: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. doi: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. doi: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. doi: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*. doi: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. doi: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. doi:10.1007/s11295-014-0752-y

Ramírez-Valiente, J. A., 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. doi:10.1007/s11295-010-0331-9

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. doi: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*. doi: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. doi: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. doi: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. doi:10.1038/nrg3522

Sebastiana, M., Vieira, B., Lino-Neto, T., Monteiro, F., Figueiredo, A., Sousa, L., … Paulo, O. S. (2014). Oak Root Response to Ectomycorrhizal Symbiosis Establishment: RNA-Seq Derived Transcript Identification and Expression Profiling. *PLOS ONE*, *9*(5), e98376. doi:10.1371/journal.pone.0098376

Shin, S., Lee, T.-H., Ha, N.-C., Koo, H. M., Kim, S., Lee, H.-S., … Oh, B.-H. (2002). Structure of malonamidase E2 reveals a novelSer-cisSer-Lys catalytic triad in a new serine hydrolase fold that is prevalent in nature. *The EMBO Journal*, *21*(11), 2509–2516. doi:10.1093/emboj/21.11.2509

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. doi: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. doi: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. doi:10.1016/j.ppees.2007.09.004

van Loon, L. c., Rep, M., & Pieterse, C. m. j. (2006). Significance of Inducible Defense-related Proteins in Infected Plants. *Annual Review of Phytopathology*, *44*(1), 135–162. doi:10.1146/annurev.phyto.44.070505.143425

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. doi: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. doi: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. doi:10.1534/genetics.113.152991

Walther, G.-R., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T. J. C., … Bairlein, F. (2002). Ecological responses to recent climate change. *Nature*, *416*(6879), 389–395. doi:10.1038/416389a

Xia, Y., Suzuki, H., Borevitz, J., Blount, J., Guo, Z., Patel, K., … Lamb, C. (2004). An extracellular aspartic protease functions in Arabidopsis disease resistance signaling. *The EMBO Journal*, *23*(4), 980–988. doi:10.1038/sj.emboj.7600086

Yang, K.-Z., Xia, C., Liu, X.-L., Dou, X.-Y., Wang, W., Chen, L.-Q., … Ye, D. (2009). A mutation in THERMOSENSITIVE MALE STERILE 1, encoding a heat shock protein with DnaJ and PDI domains, leads to thermosensitive gametophytic male sterility in Arabidopsis. *The Plant Journal*, *57*(5), 870–882. doi:10.1111/j.1365-313X.2008.03732.x

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. doi:10.1139/g98-006