Tree Genetics & Genomes
DOI 10.1007/s11295-014-0786-1

ORIGINAL PAPER

Identifying signatures of natural selection in cork oak (*Quercus suber* L.) genes through SNP analysis

Inês S. Modesto · Célia Miguel · Francisco Pina-Martins · Maria Glushkova · Manuela Veloso · Octávio S. Paulo · Dora Batista

Received: 10 January 2014 / Revised: 4 August 2014 / Accepted: 7 August 2014

Abstract Cork oak (*Quercus suber* L.) is an evergreen tree species endemic to the western Mediterranean Basin with a major economical, social and ecological relevance, associated with cork extraction and exploitation. In the last years, cork oak stands have been facing a significant decline, which may be aggravated by the climate changes that are predicted to occur within cork oak distribution range during this century. Under this scenario, the assessment of adaptive genetic variation is essential to understand how cork oak may cope with these threats and to delineate strategies for the management of its genetic resources. In this study, six candidate genes possibly significant for environmental adaptation were analysed in cork oak populations from its entire distribution range. Signatures of natural selection were investigated using population genetic statistics and environmental association tests under alternative scenarios of population genetic structure. Signals of balancing selection were detected in the putative *non-expressor of pathogenesis-related gene 1* (*NPR1*), involved in plant defence response against pathogens, in *auxin response factor 16* (*ARF16*), a gene implicated in root development, in *RAN3*, also involved in developmental processes, and in *glutamine synthetase nodule isozyme* (*GS*), involved in nitrogen fixation. Furthermore, for *ARF16*, a *class I heat shock protein* (*sHSP*) and *GS*, associations were found between SNP allele and haplotype frequencies and several spatial and climatic variables, suggesting that these genes may have a role on cork oak local adaptation. In this study, the first steps were taken into gathering information on cork oak adaptation to environmental conditions.

Keywords Adaptive genetic variation · Balancing selection · Candidate gene · Environmental association · Western Mediterranean

Communicated by A. Kremer

Electronic supplementary material The online version of this article (doi:10.1007/s11295-014-0786-1) contains supplementary material, which is available to authorized users.

I. S. Modesto (✉) · F. Pina-Martins · O. S. Paulo · D. Batista
Centro de Biologia Ambiental (CBA), Computational Biology and Population Genomics Group (CoBiG2), Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1759-016 Lisboa, Portugal
e-mail: ines.sbm@gmail.com

C. Miguel
Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa (ITQB-UNL), Av. da República, 2780-157 Oeiras, Portugal

C. Miguel
Instituto de Biologia Experimental e Tecnológica (IBET), Apartado 12, 2781-901 Oeiras, Portugal

F. Pina-Martins
Centro de Estudos do Ambiente e do Mar (CESAM) e Departamento de Biologia, Universidade de Aveiro, 3810-193 Aveiro, Portugal

M. Glushkova
Department of Forest Genetics, Physiology and Plantations, Forest Research Institute of B.A.S., 132 "St. Kliment Ohridski" blvd., 1756 Sofia, Bulgaria

M. Veloso
Unidade de Biotecnologia e Recursos Genéticos, Instituto Nacional de Investigação Agrária e Veterinária, I.P. (INIAV), Quinta do Marquês, 2784-505 Oeiras, Portugal

D. Batista
Centro de Investigação das Ferrugens do Cafeeiro/Instituto de Investigação Científica Tropical (CIFC-Biotrop/IICT), Quinta do Marquês, 2784-505 Oeiras, Portugal

Published online: 22 August 2014

Introduction

Identifying genes and allelic variations involved in local adaptation can help us to understand how species have adapted to their environment and characterize the underlying genetic basis of the adaptive process. This knowledge can be of major relevance at the present time, as it may highlight how species will respond to future climate change. In regions such as the Mediterranean Basin, studying the adaptation of species is of particular relevance, as severe climate change is expected to occur in this region during this century, with a predicted increase of at least 2–4 °C and a great decrease in precipitation (IPCC 2007; Giorgi and Lionello 2008). If climate in the Mediterranean Basin changes as fast as predicted, forest climate zone boundaries could move quicker than forest tree species are able to migrate (Higgins and Harte 2006) and consequently their survival will depend primarily on their plasticity and their ability to adapt to new environmental conditions (Davis and Shaw 2001; Valladares et al. 2007).

Cork oak (*Quercus suber* L.) is one of the keystone forest tree species in Mediterranean ecosystems. It is endemic to the western part of this region and occurs across a vast range of climatic conditions. Cork oak distribution is rather discontinuous, ranging from the Atlantic coast of North Africa and Iberian Peninsula to Southeastern Italy (Fig. 1) (Pausas et al. 2009). Moreover, it can also be found as an introduced species in Croatia (Trinajstic 2005) and in Bulgaria (Fig. 1, dark grey), where it withstands extreme cold temperatures, up to −27.5 °C (Alexandrov et al. 2001).

Cork oak has long been explored for the extraction of its outer bark, the cork, mainly in a unique agroforestry-pastoral system managed by man known as *montado* in Portugal and *dehesa* in Spain. It is due to the commercialization of cork that this species has a great economical and social relevance in the countries where it is naturally distributed. Cork oak stands also have a great ecological significance, contributing to the survival of many native plant and animal species and to prevent desertification of the areas where they are cultivated (Gil and Varela 2008). Despite its relevance, cork oak stands have been facing a significant decline by the lack of regeneration, mainly due to severe drought periods, the dependence on aged adult trees and inadequate management practices (Pulido and Diaz 2005; Otieno et al. 2006; Soto et al. 2007; Sousa et al. 2007) as well as susceptibility to several diseases (Brasier 1996; Cabral and Ferreira 1999; Moreira and Martins 2005). This decline may be aggravated by the pronounced climate change predicted to occur in the Mediterranean Basin. Therefore, understanding the processes of local adaptation and thus the species' ability to cope with environmental changes and with pests and diseases is of major relevance.

In previous studies, resorting to common garden experiments or provenance trials based on phenotypic and ecophysiological traits, evidences for cork oak local adaptation have been detected (e.g. Aranda et al. 2005; 2007; Gandour et al. 2007). For instance, in a Portuguese provenance trial under the framework of the Concerted Action EU/FAIR 1 CT 95–0202 (Varela 2000), contrasting differences were observed in survival, height, time of bud burst and water use (Nunes et al. 2008). In other studies, differential responses to low temperatures were reported between individuals from different populations (Aranda et al. 2005) and differences in survival were observed between northern, continental and southern populations from the Iberian Peninsula in drought conditions (Ramirez-Valiente et al. 2009b). Some of the adaptive traits studied in Ramirez-Valiente et al. (2011) were demonstrated to be heritable. Furthermore, Ramirez-Valiente et al. (2009a, 2010) reported one microsatellite (QpZAG46) correlated with leaf size and its population allele frequency correlated with temperature. Despite these studies, little is still known about cork oak adaptive genetic variation and no reports have been

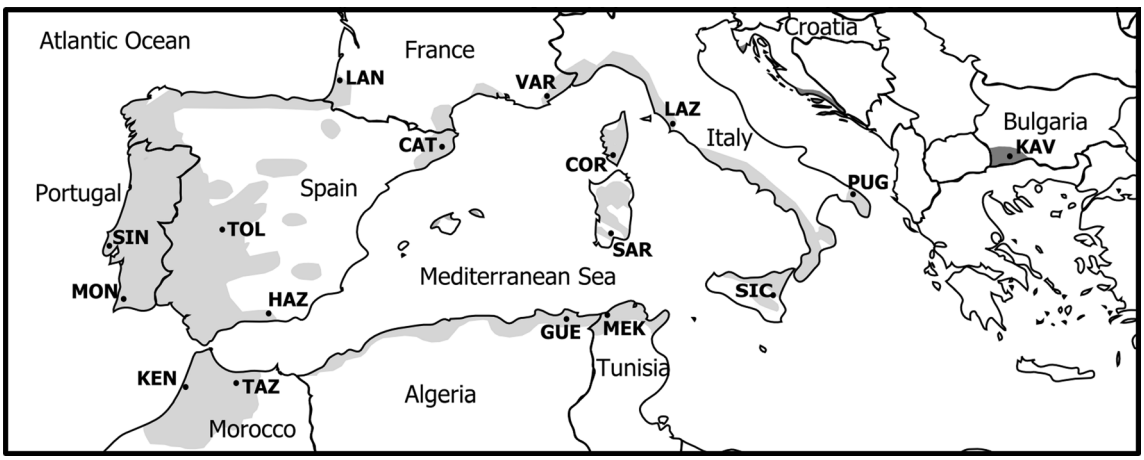

Fig. 1 Map of cork oak (*Q. suber*) geographical distribution. *Light grey* represents natural distribution; *dark grey* represents introduced, somewhat naturalized populations. Localities of the sampled populations used for this study are identified by the following codes: *SIN*, Sintra; *MON*, Monchique; *HAZ*, Haza de Lino; *TOL*, Montes de Toledo; *CAT*, Cataluña; *KEN*, Kenitra; *TAZ*, Taza; *MEK*, Mekna; *LAN*, Landes; *VAR*, Var; *COR*, Corse; *LAZ*, Lazio; *SAR*, Sardegna; *SIC*, Sicilia; *PUG*, Puglia; *KAV*, Kavrakirovo

made about genes underlying local adaptation mediated by abiotic or biotic stress responses. In contrast, in other oak species, adaptation imprints were detected through the study of nucleotide diversity, analyses of deviations from neutral models and association studies of genetic variation with phenotypes and environmental conditions, resorting to candidate genes and single nucleotide polymorphism (SNP) analyses (e.g. Homolka et al. 2013; Sork et al. 2010; Derory et al. 2006; Alberto et al. 2013).

The ability to detect signatures of natural selection in population sequence data depends on the nature and the strength of the selection events (Nielsen 2005), on the evolutionary scale at which they occur (Zhai et al. 2009) and on the sensitivity of the methods to discard other evolutionary forces that can mimic selection, such as demography and population structure (Biswas and Akey 2006). Therefore, it is important to adopt an approach combining several complementary methods, such as different neutrality tests and environmental associations that look at different evolutionary scales and types of selection, and try to account for hidden spatial genetic structure when applying these methods.

A strong geographical structure has been reported for cork oak chloroplastidial DNA (cpDNA) (Magri et al. 2007; Simeone et al. 2009; Costa et al. 2011). However, for the nuclear genome, a lack of genetic structuring seems to be evident from several independent nuclear neutral data, as the nuclear marker ITS (Simeone et al. 2009) and nuclear neutral SNP data (Pina-Martins et al. Mined ESTs SNPs bring new insights into Cork Oak population structure, in prep.). As an alternative, less probable scenario, an East and West structure is also plausible based on the independent nuclear SNP data (Pina-Martins et al., in prep.). This contrast between cpDNA and nuclear data has been reported for other long-lived outcrossing species, with long-distance pollen dispersal (Austerlitz et al. 2000), and is to be expected.

Recently, in the scope of a Portuguese consortium for the generation of a comprehensive expressed sequenced tags (ESTs) database (the Cork Oak EST Consortium—COEC) (Pereira-Leal et al. 2014), cork oak transcriptome was pyrosequenced from several different tissues, developmental stages and biotic and abiotic stress conditions. In the course of one of the projects included in this consortium ("*Polymorphism detection and validation*," FCT project SOBREIRO/0036/2009), under a population framework including 8 populations naturally growing on climatic divergent regions, more than 400 high-quality SNPs were identified in annotated contigs. Subsequently, a set of these SNPs was validated through Sanger sequencing. Being located in transcribed regions, these SNPs can be of major interest when trying to understand selective pressures acting on genes related with local adaptation to the environment (Vera et al. 2008; Renaut et al. 2010; Horton et al. 2012).

The main goals of this study were to detect genetic signatures of natural selection, in a population framework, and to test for associations of the obtained population genetic variation with environmental variables potentially relevant for cork oak local adaptation. For this purpose, six genes with putative functions in developmental processes and stress responses, retrieved from the population 454 transcriptome dataset and comprising validated SNPs, were selected to investigate imprints of natural selection, resorting to neutrality tests and environmental association tests. These analyses were performed considering three plausible underlying scenarios of population genetic structure in order to account for possible cofounding effects of historical population structure and adaptation. This is the first report providing knowledge concerning adaptive genetic variation within natural populations of cork oak, which can be integrated in future management and conservation strategies for this species.

Material and methods

Sampled populations and environmental data

Sixteen cork oak populations were sampled spanning the full distribution range of the species from an international provenance trial (FAIR I CT 95 0202) established at Monte Fava, Alentejo, Portugal (8°7′ W, 38°00′ N) (Varela 2000), except for the native Portuguese and Bulgarian populations, which were collected directly from their original locations (Table 1, Fig. 1). Populations were selected considering both geographical distribution and environmental heterogeneity between the original sampling locations, prioritizing populations that represent contrasting environments (Table 1). Three samples from holm oak (*Quercus ilex* subsp. *rotundifolia* Lam.) were also collected from the original populations [Fátima, Portugal (coordinates 39° 37′ N, 8° 40′ W); Serra da Estrela, Portugal (coordinates 40° 34′ N, 7° 18′ W); Alentejo, Portugal (coordinates 38° 5′ N, 7° 9′ W)]. Within each cork oak population, six trees were selected at random for sequencing fragments of the candidate genes (CGs) selected. The collected leaves were stored at −80 °C until DNA extraction.

Three spatial variables were recorded for each population from the original sampling sites: altitude, latitude and longitude (Varela 2000). Climatic data was gathered from WorldClim database at 30 arc-seconds resolution (about 1 km) (Hijmans et al. 2005) using DIVA-GIS version 7.5.0 (Hijmans et al. 2001). Eleven bioclimatic variables were collected from this database (Online Resource 1), as well as the maximum and minimum monthly temperatures necessary to estimate maximum temperature (T_{max}) of wettest quarter, minimum temperature (T_{min}) of wettest quarter, T_{max} of driest quarter, T_{min} of driest quarter, T_{max} of warmest quarter and T_{min} of coldest quarter. Four composed variables accounting for temperatures and precipitation were estimated through the multiplication of the following variables: T_{max} of driest quarter

Table 1 Geographic location and climatic conditions of the 16 cork oak populations sampled for this study

Code	Population	Country	Spatial variables			Climatic variables				
			Long (deg)	Lat (deg)	Alt (m)	AMT (°C)	Isothermality	P (mm)	PDQ (mm)	P season.
SIN	Sintra	Portugal	9°25′ W	38°45′ N	528	14.9	4.2	819	37	64
MON	Monchique	Portugal	8°34′ W	37°19′ N	902	13.3	4.3	731	34	63
LAZ	Lazio, Toscany	Italy	11°57′ E	42°25′ N	160	15.1	3.3	709	102	34
PUG	Puglia, Brindisi	Italy	17°40′ E	40°34′ N	45	15.9	3.5	574	66	40
SIC	Sicilia, Catania	Italy	14°30′ E	37°07′ N	250	16.1	3.4	432	20	65
SAR	Sardegna, Cagliari	Italy	8°51′ E	39°05′ N	200	13.1	3.3	757	33	54
VAR	Var, Bomes les Mimoses	France	6°15′ E	43°08′ N	155	15.0	3.6	730	75	43
LAN	Landes, Soustons	France	1°20′ W	43°45′ N	20	13.6	4.0	1,289	239	23
COR	Corse, Sartene	France	8°58′ E	41°37′ N	50	14.7	3.0	616	48	49
TOL	Montes de Toledo, Cañamero	Spain	5°21′ W	39°22′ N	800	15.1	3.6	469	35	45
CAT	Cataluña, Sta Coloma Farnes	Spain	2°32′ E	41°51′ N	500	12.7	3.1	887	178	21
HAZ	Haza de Lino	Spain	3°18′ W	36°50′ N	1,300	12.8	3.8	541	35	49
KEN	Kenitra, Ain Johra	Marocco	6°35′ W	34°05′ N	160	18.1	4.5	547	8	77
TAZ	Taza, Bab Azhar	Marocco	4°15′ W	34°12′ N	1,130	18.7	3.9	521	14	66
MEK	Mekna, Tabarka	Tunisia	8°51′ E	36°57′ N	12	18.2	4.0	825	24	70
KAV	Kavrakirovo	Bulgaria	23°10′ E	41°26′ N	200	24.1	3.3	467	86	22

Long Longitude, *Lat* Latitude, *Alt* Altitude, *AMT* Annual Mean Temperature, *P* Annual Precipitation, *PDQ* Precipitation of the Driest Quarter, *P season.* Precipitation seasonality

and precipitation of driest quarter; T_{max} of warmest quarter and precipitation of warmest quarter; T_{min} of coldest quarter and precipitation of coldest quarter; and T_{min} wettest quarter and precipitation in wettest quarter. In total, 21 environmental variables were selected. Correlations between variables were investigated using Spearman's correlation coefficient, and four environmental variables were excluded due to high correlation ($r>0.95$). Association analyses were then preformed with 17 environmental variables and 3 spatial variables (Online Resource 1).

Candidate genes loci

In this study, nucleotide polymorphisms were accessed in partial fragments of six CGs for adaptation to the environment (Table 2). CGs were selected from a database of ESTs containing SNPs generated in the course of one of the projects included in the COEC ("*Polymorphism detection and validation*," FCT project SOBREIRO/0036/2009) (NCBI accession number ERP001762), according to their functional roles described for model plants. The gene fragments sequenced in this study correspond to the six following orthologous genes of *Arabidopsis thaliana*: *RAS-related nuclear protein 3* (*RAN3*), *non-expressor of pathogenesis-related gene 1* (*NPR1*), *pathogenesis-related gene 1* (*PR1*), *auxin response factor 16* (*ARF16*), a *class I small heat shock protein* (*sHSP*) and *glutamine synthetase nodule isozyme* (*GS*). RAN3 is putatively involved in nucleocytoplasmic transport and cell cycle progress (Haizel et al. 1997; Meier and Brkljacic 2010); NPR1 is a key signalling protein of systemic acquired resistance (SAR) pathogen defence pathway (Pieterse and Van Loon 2004); PR1 is a defence protein from SAR involved in plant-pathogen interactions (Niderman et al. 1995; Rauscher et al. 1999); ARF16 is involved in root development and root cap cell differentiation (Wang et al. 2005; Ding and Friml 2010); sHSP is possibly involved in response to stress (Wang et al. 2004; Bondino et al. 2012); and GS is a cytosolic protein involved in nitrogen fixation in the root (Bernard and Habash 2009). The deduced amino acid sequences of the studied fragments were searched for protein conserved domains using BLASTp (http://blast.ncbi.nlm.nih.gov).

DNA extraction and sequencing

Total genomic DNA was extracted from liquid nitrogen-grounded leaves using the DNeasy Plant Mini Kit (Qiagen), according to the manufacturer's protocol. Primers were designed to amplify fragments of the genes *RAN3*, *NPR1*, *PR1*, *ARF16*, *sHSP* and *GS* using PerlPrimer v1.1.10 (Marshall 2004) (Table 2). PCRs were carried out in a total volume of 15 μL, containing 0.4–0.75 ng of genomic DNA, 0.4 U GoTaq DNA Polymerase (Promega), 1× reaction buffer (Promega), 0.4 μM of each primer, 0.1 mM dNTPs mix and 3.2 mM $MgCl_2$. Negative controls were included in all sets of PCR reactions. Amplification cycles started with 5 min

Table 2 Summary data of candidate genes (CGs)

CG	Annotation	Length[a]	Primer[b]	Ta (°C)[c]
RAN3	*RAS-related nuclear protein 3*	627	Fwd: TATCTTGCCAGGAAGCTTGC Re: GGTCTATGGTCAATAGCCGAC	53
NPR1	*Non-expressor of pathogenesis-related gene 1*	270	Fwd: ACAGAGCTCCTTGATCTTGC Re: GAGATCATCACCTGCCATAGC	53
PR1	*Pathogenesis-related gene 1*	257	Fwd: CAACCGATGAATGTGCCTCC Re: TGGACCTATAACATGGGACGC	64
ARF16	*Auxin response factor 16*	234	Fwd: GAATATCTTCAGAAGATCTCCACC Re: CATTTAGAAATCTGCTCCTCAGTG	65
sHSP	*Class I small heat shock protein*	374	Fwd: GTGTTCAAAGCTGATCTTCC Re: ACCTTCTGACAAGTAAACCC	56
GS	*Glutamine synthetase nodule isozyme*	514	Fwd: GCCCTTCTGTTGGTATATCTGC Re: GTTTCATGTCGGCCAGTGAG	62

[a] Length of the gene fragments sequenced

[b] Primer sequence (5′–3′)

[c] Annealing temperatures of each pair of primers

denaturation at 94 °C, followed by 35–40 cycles of 30 s at 94 °C, 30 s at variable annealing temperatures (Table 2) and 1 min at 72 °C, with a final extension step at 72 °C for 15 min. PCR products were purified using SureClean (Bioline) and sequenced on ABI PRISM 310 or ABI 3730XL (Applied Biosystems) genetic analysers. The obtained sequences were edited with Sequencher v4.0.5 (Gene Codes Corporation) and aligned using ClustalW (Thompson et al. 1994). The heterozygous phase was determined using the program PHASE v2.1.1 (Stephens et al. 2001; Stephens and Scheet 2005) for *Q. suber* and *Q. rotundifolia* separately, with default parameters, or Champuru v1.0 when indels were present (Flot et al. 2006; Flot 2007).

Statistical analyses

Gene diversity

Number of polymorphic sites (S), nucleotide diversity (π), diversity at non-synonymous sites (π_A), diversity at silent sites (π_S), number of haplotypes (*Hap*), haplotype diversity (H) and number of synonymous and non-synonymous substitutions were computed for *Q. suber* using the program DnaSP v5 (Librado and Rozas 2009). Amino acid replacements were assessed through the translation of the putative ORFs in BioEdit v7.1.3.0 (Hall 1999).

Population genetic structure

In order to investigate if the patterns of genetic variation found in the CGs reflect a historical population structure, analyses of molecular variance (AMOVA) were performed employing ARLEQUIN v3.5 (Excoffier and Lischer 2010). Three potential scenarios of genetic structure were tested: scenario 1—lack of genetic structure (one single group), as suggested by nuclear neutral SNPs (Pina-Martins et al., in prep.), from onwards designated as LS scenario; scenario 2—East vs West regions (two groups), as marginally suggested by the same dataset, from onwards designated as East/West scenario; scenario 3—four lineages, as described by Magri et al. (2007) based on cpDNA microsatellite data (four groups), from onwards designated as cpLineage scenario. AMOVAs were conducted without the Bulgarian population Kravakirovo as this is an introduced population and may lead to biased results.

Neutrality tests

Two neutrality tests based on within-species population genetic data, Tajima's D (Tajima 1989) and Fu's Fs (Fu 1997), were conducted using ARLEQUIN v3.5. These tests were performed for the three alternative scenarios described above, since historical population structure can affect the detection of signals of natural selection (Excoffier et al. 2009).

For site-specific sequence analysis of selective pressures acting on each CG, a maximum likelihood approach was implemented using CODEML from PAML v4.6 software package (Yang 2007). This analysis was conducted using *Q. rotundifolia* as outgroup. To test for positive selection acting on different sites across the protein sequence, three site models were tested: M0, that assumes one site rate for all codon sites, M1, which corresponds to neutrality and assumes two values for ω (ratio between the non-synonymous mutations per non-synonymous sites, d_N, and the synonymous mutations per synonymous sites, d_S) ($\omega=1$ and $\omega<1$), and M2, that estimates three values of ω ($\omega=1$, $\omega<1$ and $\omega>1$) and accounts for positive selection. Likelihood ratio tests (LRT) were performed to compare the three models, and a χ^2 distribution was used to check for significant differences

between the log likelihoods of the models as implemented in the software package. Posterior probabilities of the inferred positively selected sites were estimated by the Bayes empirical Bayes (BEB) approach that takes sampling errors into account (Yang et al. 2005).

Environmental association analysis

Correlations between genetic data (SNP allele or haplotype frequencies) and spatial and climatic variables were tested using MatSAM v2 (Joost et al. 2007). This program computes series of univariate logistic regression models. Significance of the correlations was assessed through two LRTs (*G* and Wald tests), and the null hypothesis of no association between the genetic and the environmental data was rejected at a 5 % significance level, after Bonferroni correction. To account for the effect of putative population structure, the association tests were conducted separately for each of the groups included in the three alternative scenarios previously considered (see above in population genetic structure sub-section). As before, Kravakirovo population was excluded from these analyses to avoid biased results.

Results

Diversity and population structure

For the six studied CGs, sequences were obtained for 59 to 95 individuals from 16 populations, depending on the loci (Online Resource 2). The regions analysed covered a total of approximately 2.3 kb, ranging from 234 bp (*ARF16*) to 627 bp (*RAN3*). From this total, 1,372 bp corresponded to coding sequence and 904 bp to non-coding sequence (introns and 3′-UTR). The number of polymorphic sites observed per gene varied between 4 (*ARF16*) and 13 (*RAN3*), giving a total of 44 SNPs detected (Table 3). Of these, 15 were non-synonymous SNPs (34.1 %), 8 synonymous (18.2 %) and 21 in non-coding regions (47.7 %). In addition, four indels were detected in *GS*, located in introns. In total, 47 mutations were detected (SNPs and indels), giving an average of one mutation per 48 bp. One non-synonymous mutation was detected in *GS*, leading to a non-conservative amino acid replacement (in which the original amino acid is replaced by another with different physicochemical properties) (Table 4), while in *NPR1* and *ARF16*, three non-synonymous mutations were identified, two of which corresponding to non-conservative substitutions. In *PR1* and *sHSP*, four non-synonymous and non-conservative mutations were detected. *RAN3* was the only CG that presented only synonymous mutations or mutations in non-coding regions.

The average total nucleotide diversity (π) was 0.0060, varying between 0.0027 at *GS* and 0.0069 at *NPR1*. The levels of haplotypic diversity (*H*) are also heterogeneous among loci, with higher values detected at *sHSP* (0.836) and lower values found at *GS* (0.553), with an average of 0.635. Nucleotide diversity at non-synonymous sites (π_A) was higher than diversity at synonymous sites (π_S) for three of the six CGs (*PR1*, *ARF16* and *GS*). By contrast, for the other three CGs, π_S was higher than π_A, although for *NPR1* and *sHSP* the difference between the two values was small.

Examining the haplotype distribution throughout the sampled populations (Online Resource 3), no prominent genetic structure was detected. All the analysed CGs presented two to three more frequent haplotypes common to all or almost all populations, along with a few less frequent and less spread ones.

In the results obtained from the AMOVAs (Table 5), the overall source of variation is observed within populations for all tested groupings. Significant values were found for the variance among groups when considering the East/West scenario for the CGs *sHSP* and *GS*. However, the percentage of variance explained by the differences among groups is very low (1.42 and 6.99 % for each CG, respectively). No significant values of differentiation among groups were found for any of the CGs when considering the cpLineage scenario described by Magri et al. (2007).

Neutrality tests

For the LS scenario, the Tajima's *D* and Fu's *Fs* tests rejected the null neutral model for *NPR1* and *ARF16* (Table 6). For both CGs, the values of *D* and *Fs* obtained were positive, indicating an excess of intermediate frequency alleles in the first, consistent with balancing selection or population decline, and a deficit of allele number in the latter, which also suggests balancing selection or a population decline. Considering the East/West scenario, positive significant values were detected for *NPR1* in East and West groups (*D* and *Fs*), for *RAN3* in East (*Fs*) and West groups (*D* and *Fs*), for *ARF16* in the East group (*D*) and for *GS* also in the East group (*Fs*) (Table 6). Similarly, the results obtained for the cpLineages reflect equivalent trends (Table 6).

Analyses with PAML were performed for all CGs except for *RAN3*, as no non-synonymous mutations were found in this gene fragment. The selection model (M2) was not significantly more adjusted to the data than the neutral model (M1) for any of the five CGs investigated. However, for *sHSP*, M2 likelihood was higher than M1, even though not significantly, and the selection model detected three positions potentially under positive selection, one of them, SNP position 51 (amino acid position 17), with a significant *p* value ($p<0.05$) (Online Resource 4).

Table 3 Summary data on the amplified gene fragments and respective diversity indexes

CG	*N*	Amp. region	*S*	Non Cod	Syn	*Non Syn*			Indels	π	π_S	π_A	Hap	*H*
						Total	*Con*	*Non Con*						
RAN3	65	I/E/I/E/3′ UTR	13	11	2	–	–	–	–	0.0068	0.0268	0.0000	11	0.597
NPR1	105	E	5	–	2	3	1	2	–	0.0069	0.0074	0.0068	4	0.558
PR1	94	E	5	–	1	4	–	4	–	0.0052	0.0010	0.0035	5	0.609
ARF16	96	E	4	–	1	3	1	2	–	0.0067	0.0048	0.0073	5	0.648
sHSP	94	E/3′ UTR	10	4	2	4	–	4	–	0.0062	0.0052	0.0041	14	0.836
GS	93	E/I/E/I/E/I/E	7	6	–	1	–	1	3	0.0027	0.0000	0.0031	6	0.553

N number of individuals sequenced, *Amp. region* amplified region, *I* intron, *E* exon, *UTR* untranslated region, *S* number of polymorphic sites, *Non Cod* number of SNPs in non-coding regions, *Syn* number synonymous SNPs, *Non Syn* number of non-synonymous SNPs, *Con* number of conservative amino acid changes, *Non Con* number of non-conservative amino acid changes, *Indels* number of indels, π genetic diversity, π_S diversity at silent sites, π_A diversity at non-synonymous sites, *Hap* number of haplotypes, *H* haplotype diversity

Environmental association analysis

Environmental associations were tested both at SNP and haplotype levels, and significant results were obtained in four of the CGs studied, *NPR1*, *ARF16*, *sHSP* and *GS* (Table 7).

When considering the LS scenario, SNP position 72 of the CG *ARF16* was detected as being correlated with precipitation of the driest quarter (Table 7) (significance detected only with *G* test). Allele G is directly proportional to this variable, while allele A is inversely proportional, meaning that allele A is more frequent in populations from locations with lower precipitation in the driest quarter of the year (Table 7, Online Resource 5). The frequency of haplotype 5, the only haplotype detected with A in SNP position 72, was also found as being negatively correlated with precipitation of the driest quarter (Table 7). When considering the East/West scenario, three additional associations were found in the West group, with latitude, precipitation of the driest quarter, precipitation of the warmest quarter and precipitation seasonality (Table 7). In this case, Allele A is inversely proportional to latitude, precipitation of the driest quarter and precipitation in the warmest quarter, and directly proportional to precipitation seasonality. Correlations were also found in the West group for haplotype 5 with these same four variables and further on with T_{min} of the driest quarter (positive correlation) (Table 7). Finally, when considering the cpLineage scenario, associations were only detected in Lineage 2 between SNP 72AG and longitude and precipitation of the warmest quarter (Table 7). In this lineage, allele A is negatively correlated with longitude and positively correlated with precipitation of the warmest quarter. In the same lineage, correlations were found between haplotype 5 and also longitude (negative correlation) and precipitation of the warmest quarter (positive correlation). Furthermore, haplotype 5 was detected as being negatively correlated to the composed variable T_{min} of the coldest quarter × precipitation of the coldest quarter, in Lineage 1 (Table 7).

Table 4 Non-synonymous mutations of candidate genes and corresponding amino acid changes

Gene	SNP position	SNP allele	Amino acid	Properties	SNP allele	Amino acid	Properties
NPR1	22	G	Ala	Non-polar, aliphatic	T	Ser	Polar, uncharged
	35	A	Lys	Positively charged	T	Met	Non-polar, aliphatic
	87	C	Asp	Nagatively charged	G	Glu	Nagatively charged
PR1	3	A	Gln	Polar, uncharged	G	Arg	Positively charged
	163	G	Asp	Nagatively charged	A	Asn	Polar, uncharged
	164	A	Asp	Nagatively charged	C	Ala	Non-polar, aliphatic
	174	G	Gly	Non-polar, aliphatic	A	Glu	Nagatively charged
ARF16	13	A	His	Positively charged	G	Arg	Positively charged
	72	G	Ala	Non-polar, aliphatic	A	Thr	Polar, uncharged
	180	G	Glu	Nagatively charged	A	Lys	Positively charged
sHSP	51	A	Asp	Nagatively charged	G	Gly	Non-polar, aliphatic
	142	C	Leu	Non-polar, aliphatic	A	Phe	Aromatic
	189/190	AG	Lys	Positively charged	TT	Ile	Non-polar, aliphatic
GS	501	T	Leu	Non-polar, aliphatic	G	Arg	Positively charged

Table 5 Analyses of molecular variance (AMOVAs) performed with *RAN3*, *NPR1*, *PR1*, *ARF16*, *sHSP* and *GS*

Fragment	Groups	Variance components (%)		
		Within populations	Among populations	Among groups
RAN3	LS scenario	97.13	2.87	-
	East/West scenario	95.51	1.08	3.41
	cpLineage scenario	97.19	3.09	−0.28
NPR1	LS scenario	91.26	8.74*	-
	East/West scenario	91.80**	9.48**	−1.28
	cpLineage scenario	91.84**	11.28**	−3.13
PR1	LS scenario	94.91	5.09*	-
	East/West scenario	93.48*	3.29	3.23
	cpLineage scenario	94.88*	4.98*	0.14
ARF16	LS scenario	95.12	4.88	-
	East/West scenario	95.77	5.70*	−1.47
	cpLineage scenario	95.32	5.66*	−0.98
sHSP	LS scenario	101.24	−1.24	-
	East/West scenario	100.57	−1.99	1.42*
	cpLineage scenario	101.22	−1.31	0.09
GS	LS scenario	96.74	3.26	-
	East/West scenario	93.58	−0.57	6.99*
	cpLineage scenario	95.89	−0.12	4.23

Populations were grouped according to three scenarios of population structure: (1) LS scenario—lack of genetic structure (one group); (2) East/West scenario—East and West regions (two groups); (3) cpLineage scenario—four lineages based on chloroplastidial data (four groups). LS scenario: all populations included in one group; East/West scenario: [East group] Var, Mekna, Corse, Sardegna, Puglia, Lazio, Sicilia; [West group] Taza, Cataluña, Haza de Lino, Montes de Toledo, Sintra, Monchique, Kenitra, Landes; cpLineage scenario: [Lineage 1] Puglia, Lazio, Sicilia; [Lineage 2] Var, Mekna, Corse, Sardegna; [Lineage 3] Montes de Toledo, Sintra, Monchique, Kenitra, Landes; [Lineage 4] Taza, Cataluña, Haza de Lino

*$p<0.05$; **$p<0.01$

Table 6 Tajima's *D* and Fu's *Fs* neutrality tests, considering the three potential scenarios of population structure tested

Scenarios			*RAN3*	*NPR1*	*PR1*	*ARF16*	*sHSP*	*GS*
LS		*D*	1.95363	2.25774*	1.06458	2.31433*	0.81103	0.26697
		Fs	2.3659	4.98473**	2.03900	2.78840*	−1.65000	4.39741
East/West	East group	*D*	1.65122	2.01752*	0,90490	2.20442*	0.99803	0.75852
		Fs	4.21815*	3.99675*	0.17942	2.15718	−1.9162	6.27781*
	West group	*D*	2,19791*	2.71787**	0,66411	1.89793	0.91394	−0.81306
		Fs	3.71188*	5.50045**	1.00123	2.9683	−1.26549	0.45613
cpLineage	L1	*D*	1.16642	2,08909*	1.00586	1.38372	0.06481	2,49878*
		Fs	3.47437*	3,76558**	0.97073	0.70715	−3.12445	5,86523**
	L2	*D*	0.43039	1.7043	1.04527	2,22551*	0.92407	0.54439
		Fs	5,42507**	2,96797*	1.61303	2,71033*	−0.81879	4,87983**
	L3	*D*	2.00432	2,38577*	0.58101	2.00537	1.2347	0.68796
		Fs	3,08560*	4,89809***	1.94438	2,70912*	0.16854	1.5657
	L4	*D*	1.74874	2,64532**	−0.02618	0.76300	0.18860	−1.22136
		Fs	3,16621*	4,27243***	1.05141	2.49314	−2.55188	1.35372

LS, lack of genetic structure scenario; *East/West*, East/West structure scenario; *cpLineage*, four cpDNA lineages structure scenario; *L1*, lineage 1; *L2*, lineage 2; *L3*, lineage 3; *L4*, lineage 4

*$p<0.05$; **$p<0.01$

Table 7 Significant associations detected between SNP allele and haplotype frequencies and environmental variables, with MatSAM analyses, considering the potential scenarios of population structure tested

CG	SNP position/Haplotypes	SNP allele	Correlation (+//−)	Scenarios						
				LS	East/West		cpLineage			
					East group	West group	L1	L2	L3	L4
RAN3	–	–	−	–	–	–	–	–	–	–
NPR1	Hap 1	–	−	–	–	–	–	Longitude *	–	–
PR1	–	–	−	–	–	–	–	–	–	–
ARF16	SNP 72	A/G	−//+	P driest Q #	–	Latitude * P driest Q ## P warmest Q ##	–	Longitude #	–	–
		A/G	+//−	–	–	P seasonality #	–	P warmest Q #	–	–
	Hap 5	–	−	P driest Q #	–	Latitude * T driest Q # P driest Q ## P warmest Q ##	T_{min} coldest Q x P coldest Q #	Longitude #	–	–
			+			P seasonality #		P warmest Q ##		
HSP	SNP 51	A	+	P driest Q # P warmest Q ##	–	P warmest Q #	–	–	–	–
			−	P seasonality #	–	P seasonality #	–	–	–	–
	Hap 13	–	+	P seasonality #	–	–	–	–	–	–
GS	SNP 28	A/T	+//−	Longitude *	–	–	–	–	–	–
		A/T	−//+	Isothermality *	Isothermality #	–	–	T_{mean} diurnal range ## P annual # Isothermality #	–	–
	SNP 242	T/C	+//−	Longitude *	–	–	–	–	–	–
		T/C	−//+	Isothermality *	Isothermality #	–	–	T_{mean} diurnal range ## P annual # Isothermality #	–	–
	Indel 409	del/ins	+//−	Longitude *	–	–	–	–	–	–
		del/ins	−//+	Isothermality *	Isothermality #	–	–	T_{mean} diurnal range ## P annual # Isothermality #	–	–
	SNP 501	G/T	+//−	Longitude *	–	–	–	–	–	–
		G/T	−//+	Isothermality *	Isothermality #	–	–	T_{mean} diurnal range ## P annual #	–	–

Table 7 (continued)

CG	SNP position/Haplotypes	SNP allele	Correlation (+/−)	Scenarios						
				LS	East/West		cpLineage			
					East group	West group	L1	L2	L3	L4
	Hap 1	−	+	Longitude *	−	−	−	−	−	−
								Isothermality #		

+, positive correlation; −, negative correlation; *LS*, lack of structure scenario; *East/West*, East/West structure scenario; *cpLineage*, four cpDNA lineages structure scenario; *L1*, lineage 1; *L2*, lineage 2; *L3*, lineage 3; *L4*, lineage 4; *del*, deletion; *ins*, insertion; *P*, precipitation; *Q*, quarter of the year; T_{min}, minimum temperature; T_{mean}, mean temperature

*$p<0.05$ (Wald and *G* significance tests); #$p<0.05$ (*G* significance test); ##$p<0.01$ (*G* significance test)

For *NPR1*, only one association was detected, between haplotype 1 and longitude in the cp Lineage 2 (Table 7), with a negative correlation.

For s*HSP*, when considering the LS scenario, a positive correlation was detected between the frequency of allele A in SNP position 51 (AG) and precipitation of the driest quarter and of the warmest quarter (Table 7). Moreover, a negative correlation was found between the same allele frequency and precipitation seasonality (Table 7). At the haplotype level, a positive correlation was also found between haplotype 13 and precipitation seasonality (Table 7). For the East/West scenario, allele A of SNP 51 was also correlated in the West group with precipitation of the warmest quarter (positive correlation) and precipitation seasonality (negative correlation) (Table 7).

For *GS*, when considering the LS scenario, SNPs 28AT, 242TC and 501TG and the indel in position 409 were associated with longitude and isothermality (Table 7). Alleles A from SNP 28, T from SNP 242 and G from SNP 501 and the deletion in position 409 were positively correlated with longitude and negatively correlated with isothermality. Haplotype 1 was also positively correlated with longitude (Table 7). When considering the East/West scenario, equivalent correlations were found between the same SNPs and indel and isothermality in the East group (Table 7). However, when considering the cp Lineages, these SNPs and the indel were associated with mean temperature diurnal range, annual precipitation and isothermality in Lineage 2 (Table 7). Alleles A from SNP 28, T from SNP 242 and G from SNP 501 and the deletion in position 409 were negatively correlated with all the three variables.

Discussion

The levels of overall nucleotide diversity detected in the present study (π=0.00600) are consistent with those reported for other studies with CGs of oak species. Quang et al. (2008) made a population survey in 11 genes of *Quercus crispula*, detecting an overall diversity of 0.00693, while in a study with 9 candidate genes of *Quercus petraea*, an average of total nucleotide diversity of 0.00615 was detected (Derory et al. 2010). In a more recent study, slightly lower values were found in eight candidate genes of *Q. petraea* (π=0.00374) and *Quercus robur* (π=0.00365) (Homolka et al. 2013).

For *RAN3*, the estimated nucleotide diversity at silent sites (π_S) was considerably higher than diversity at non-synonymous sites (π_A), which is consistent with what is expected for coding regions, as they are likely to be under strong purifying selection to preserve the protein structure and function. For *NPR1* and *sHSP*, the differences between π_A and π_S were small, suggesting that purifying selection may be relaxed in these genes or that they may be under positive

selection. For *PR1*, *ARF16* and *GS*, π_A was higher than π_S, indicating also that these genes may be under relaxed purifying selection or positive selection. Other than this, analysis of *PR1* did not suggest any signs of selection. The statistical methods used in this study may not have enough resolution to detect selection in this fragment, which has only 257 bp length. Moreover, the association tests did not reveal any significant result with the abiotic variables used, as it is more likely that biotic factors are the major selective pressure acting on this gene.

On the contrary, it was possible to identify different selection patterns in *RAN3*, *NPR1*, *ARF16*, *sHSP* and *GS*, allowing for the first steps to be taken into gathering important information and insights on cork oak adaptation to environmental conditions.

The patterns of genetic variation detected from the AMOVA results in the six CGs are not in agreement with the historical population structure observed on chloroplastidial markers (Magri et al. 2007; Simeone et al. 2009; Costa et al. 2011). This may reflect the lack of genetic structure in the nuclear genome suggested by other nuclear markers (ITS, Simeone et al. 2009; SNPs, Pina-Martins et al. in prep.) or a confounding effect of natural selection. Only two CGs showed significant AMOVA results (*GS* and *sHSP*), namely for the East/West groups' division, although expressed by low differentiation values.

For *GS*, three SNPs and one indel were detected as being significantly associated with environmental variables and when these are excluded from the AMOVA analysis, the East/West structure signal dissolves (from 6.99 to 0.94 %). Thus, the AMOVA results found in *GS* seem to result from natural selection rather than reflecting historical population structure.

Among these significantly associated mutations, one is a non-synonymous and non-conservative mutation (SNP 501TG) that leads to the replacement of a leucine (Leu) residue, non-polar and aliphatic, for an arginine (Arg) residue, positively charged (Table 4). This amino acid change may have an impact in the structure of the protein and, as it is located in the catalytic domain, may also alter its function. This SNP seems to be in linkage disequilibrium with the other three mutations (Online Resource 6), which suggests a hitchhiking effect resulting from positive selection. The association analyses revealed a correlation of these mutations with isothermality both for the LS scenario and the East/West scenario, although in this case only in the East group, and with longitude only for the LS scenario. Thus, allele G from SNP 501(Arg) (and alleles 28A, 242 T and deletion in position 409) tends to be more frequent in populations with lower isothermality values, i.e. with higher temperature oscillations, characteristic of higher longitudes (Murphy 1985). Haplotype 1 is the only haplotype with alleles 28A, 242 T, 501G and deletion in position 409 and is also correlated with longitude in the LS scenario, being also more frequent in populations located at higher longitudes (Eastern populations). Since longitude is not directly perceived by living organisms, this association may reflect the co-varying environmental variables such as winter severity, seasonality, annual temperature range (Murphy 1985) and in particular isothermality, since a significant correlation was found between isothermality and longitude in our data ($r=-0.618$, $p<0.05$; data not shown).

The *GS* investigated in our study codes for a nodule isozyme, which is a cytosolic protein involved in nitrogen fixation in the root (Bernard and Habash 2009; Andrews et al. 2013). In previous works, it has been demonstrated in different plant species that the activity of this GS protein and the expression of this *GS* gene respond to several factors, including biotic and abiotic stresses, such as drought and salt stress (e.g. Yan et al. 2005; Teixeira and Pereira 2007) and high temperature stress (Hungria and Kaschuk 2013). Accordingly, in our work, the *GS* candidate gene was detected as being associated with a temperature variable.

For *GS*, although no signals of balancing selection were detected in the LS scenario, Fu's *Fs* neutrality test was significantly positive in the East group, from the East/West scenario. In this group, associations with environmental variables were also detected, suggesting that *GS* is under directional selection. Therefore, the balancing selection signal may outcome from considering together populations with different allele frequencies, resulting from directional selection along an environmental gradient. On the other hand, the signal may result from dividing the populations in artificial groups, which may lead to false positives (Excoffier et al. 2009).

Considering the results obtained for *GS*, it seems that this gene is under positive selection, being involved in adaptation to temperature.

A significant population differentiation was also identified for *sHSP* CG with the AMOVA analysis, for the East/West groups' division. When performing this analysis without the SNP for which associations with environmental variables were detected, the East/West structure signal remains significant (1.80 %, $p<0.05$), nonetheless weak. *sHSP* PAML analysis indicated that the amino acid position 17 of the inferred peptide chain may be under positive selection, although the selection model (M2) was not significantly more adjusted to the data than the neutral model (M1). This mutation corresponds to SNP 51 of the nucleotide sequence that leads to a non-conservative amino acid change, in which an aspartate (Asp) residue, negatively charged, is replaced by glycine (Gly), a non-polar residue. The same SNP was detected as being associated with precipitation variables both for the LS scenario and the West group of the East/West scenario. Therefore, allele A (Asp) tends to be more frequent in locations with higher precipitation and less precipitation seasonality (Online Resource 5). Haplotype 13 seems to be also associated with precipitation seasonality in the LS scenario, being more

frequent in populations with high seasonality values. Therefore, *sHSP* may be involved in cork oak local adaptation to drought, being the individuals with allele G in SNP 51 likely to be more tolerant to drought conditions and marked precipitation seasonality and individuals with haplotype 13 more tolerant to marked precipitation seasonality. Our results are in accordance with previous studies that report several small heat shock proteins class I as being involved in response to drought stress in different plant species (e.g. Coca et al. 1994; Sato and Yokoya 2008) including *HSP17* in cork oak, a gene that is induced by water stress in somatic embryos (Puigderrajols et al. 2002).

As a significant population differentiation was detected, we cannot exclude the hypothesis that the significant associations result from this genetic structure and not from the action of directional selection. However, this differentiation is very low and PAML results (not affected by population genetic structure) also suggest that *sHSP* may be under positive selection.

From the four genes showing no signs of differentiation from the AMOVA analysis, three (*NPR1*, *ARF16* and *RAN3*) displayed significantly positive values of Tajima's *D* and Fu's *Fs*, which suggest that these genes may be under balancing selection or that cork oak population is declining. However, if that was the case, demographic effects should be affecting all the genome in a similar way, and thus it seems more probable that these three CGs should be under balancing selection. Therefore, this selection may have erased a possible signal from historical population structure, resulting in the absence of significant signals in the AMOVA analysis.

For *NPR1*, the signal of balancing selection is consistent for both the LS scenarios and the East/West scenario. Interestingly, Caldwell and Michelmore (2009) have also shown evidences of balancing selection for *A. thaliana NPR1*, similar to what is reported here for the putative orthologous gene in cork oak, suggesting that it may be under balancing selection in different plant species.

Cork oak NPR1 protein is probably involved in plant defence response to several pathogens and in the cross-communication between the three known plant defence pathways, as in other plant systems (Pieterse and Van Loon 2004). Consequently, changes in the peptide chain may affect the plant defence capacity. In the studied fragment, three non-synonymous mutations were found in the ankyrin repeat binding domain, the same domain in which signals of selection were found in the *A. thaliana NPR1*. Two of these correspond to non-conservative amino acid changes. One leads to the replacement of an alanine (Ala) residue by a serine (Ser) residue, involving a change in polarity, while the other corresponds to the replacement of lysine (Lys) residue, positively charged, by a methionine (Met) residue, with a non-polar aliphatic R group. As the original amino acid is replaced by one with different physicochemical properties, these mutations are likely to alter at some degree the structure and can possibly change the function of the protein, as it is located in a binding domain. The conservative mutation, located at the same domain, is less likely to be under selection, as it is less probable to alter the protein, although together with the other amino acid changes it may have some impact on the protein's structure. The interaction of ankyrin repeat domain with TGA transcription factors enhances their DNA binding activity to the promoter elements of *Pathogenesis-related* (*PR*) genes and is, therefore, thought to be critical for defence gene activation. In previous studies, SNP mutations in the ankyrin repeat domain of NPR1 were demonstrated to abolish interaction with TGA factors and the activation of *PR* genes (e.g. Zhou et al. 2000; Shearer et al. 2009). Consequently, natural variation within this domain is expected to affect the expression profile of *PR* genes in response to pathogens by altering the affinity of NPR1 for TGA transcription factors. For *NPR1*, no environmental associations were found when considering the LS scenario or the East/West scenario. Selective pressure acting on this CG is then probably exerted by pathogens rather than by climatic conditions, through pathogen effector proteins that may target defence pathway signalling proteins to suppress resistance (Caldwell and Michelmore 2009). Different mutations in *NPR1* may hence be associated with resistance to different pathogens or strains, or different levels of resistance.

The *A. thaliana* ARF16 protein is a transcription factor involved in root cap cell differentiation (Wang et al. 2005; Ding and Friml 2010) and in the regulation of the abaxial identity of leaves (Liu et al. 2011). In the orthologous *ARF16* of cork oak, SNP 72 was identified as being associated with precipitation variables, both for the LS scenario and the West group from East/West scenario, suggesting that this CG may be under directional selection. This finding seems to be inconsistent with the balancing selection signal previously detected in the LS scenario. In the LS scenario, the apparent signal of balancing selection may be a product of considering, as a single group, populations that have different allele frequencies in response to an environmental gradient Accordingly, when dividing the populations in East and West groups, the signal fades and is detected only in the East group, and a higher number of associations are found in the West group than in the LS scenario. This supports the idea that the West populations may be under directional selection and that the balancing selection signal in the LS scenario is probably a result of mixing East and West groups. Moreover, the East group may be indeed under balancing selection, although only one of the neutrality tests was significant, which would indicate that lineage-specific selection may be occurring.

The mutation in SNP 72 is non-synonymous and corresponds to a non-conservative amino acid change from an alanine (Ala) residue to a threonine (Thr) residue, with a change in polarity. Similar association signals with precipitation variables were obtained for haplotype 5, the only

haplotype with allele A in SNP 72. Allele A (Thr) in this SNP and haplotype 5 are therefore more frequent in populations exposed to lower precipitation and higher precipitation seasonality (Online Resources 5). This could indicate that *ARF16* is involved in adaptation to drought, being the individuals with allele A (Thr) and haplotype 5 more tolerant to this type of stress. Confirming the involvement of ARF16 in root development, previous studies have shown that abolishing the expression of this transcription factor leads to uncontrolled root growth (Wang et al. 2005). Moreover, in a study of *Q. robur* transcriptome in drought conditions, *ARF16* was downregulated in stress conditions (Spiess et al. 2012), which may, accordingly, stimulate root growth. Therefore, if the putative ARF16 has a conserved function in cork oak, the mutation in SNP 72 may have an impact on root growth in response to drought stress.

For *RAN3*, a gene that may be involved in development pathways (Merkle 2011), signals of balancing selection were detected when considering the East/West scenario, but not the LS scenario. Such signal pattern differences between different scenarios were also detected for other CGs (*GS* and *ARF16*). This complexity of selection signals, both of balancing selection and directional selection, when considering different geographical scales and different groups, suggests that cork oak may be subjected to a complex pattern of selection forces and types of selection acting on different scales of the species range for these CGs.

Ignoring spatial genetic structure when studying signatures of natural selection can lead to the identification of false positives (Excoffier et al. 2009). Therefore, although the evidences suggest a lack of genetic structure in the nuclear genome of cork oak, we performed association considering an East and West population structure (Pina-Martins et al., in prep.) and a four cpDNA lineage structure (Magri et al. 2007), in addition to a global dataset analysis. For CGs *sHSP* and *GS*, the associations detected for East/West scenario were the same as the ones detected for the LS scenario or a subset of these. However, association signals obtained for the East/West scenario were weaker and could only be detected in one of the two groups. East and West groups are unequal in the range of environmental conditions represented, and SAM analysis depends on a number of individuals representing many different landscapes in order to enhance the power of this method to detect associations with environmental variables (Joost et al. 2007). Therefore, dividing the global dataset can diminish the statistical power of this method, leading to weaker association signals or their disappearance. On the other hand, there may be a lineage-specific adaptation in one of the two groups. However, if this was the case, the association signal should be diluted when joining the East and West populations in the global analysis and not the contrary. In contrast to the *sHSP* and *GS* results, for *ARF16* additional associations were detected with different variables, with stronger correlations signals, when considering the East/West scenario, in the West group only. As the new detected variables are correlated with the variable also detected in the global analysis, they may have emerged as significant due to the reduction of the number of models that the Bonferroni correction accounts for. On the other hand, stronger association signals in the West group may indicate a lineage-specific adaptation that would have its signal diluted when adding the East populations in the global analysis. When dividing the population in the cp Lineages, most associations are only found in Lineage 2 for all the CGs and these association signals tend to be considerably different from the ones detected in the two other scenarios. Therefore, separating our populations in these cpDNA lineages could have generated false positives, as populations are divided in an excessive number of (artificial) groups (Excoffier et al. 2009). On the other hand, we cannot completely discard the hypothesis of the association signals being detected in the larger groups (LS scenario and East/West scenario) due to hidden population genetic structure. Further studies are needed to deeply assess and confirm these SNP association results.

In conclusion, important evidence was obtained for the understanding of cork oak adaptation to conditioning abiotic factors. Two genes were found to be probably involved in cork oak local adaptation to drought stress, *ARF16* and a *sHSP*, each bearing one mutation associated with precipitation variables. Furthermore, *GS* seems to be involved in adaptation to a temperature variable. This can be particularly relevant in the light of climate change, as an increase in temperatures and a great decrease in precipitation is expected in the Mediterranean Basin during this century (IPCC 2007; Giorgi and Lionello 2008). Furthermore, signals of balancing selection were detected in *RAN3*, a gene possibly involved in development pathways, and in *NPR1*, probably due to selective pressure wielded by pathogens. Several pests and diseases, such as *Phytophthora cinnamomi*, have been pointed out as relevant factors involved in the decline of cork oak (Brasier 1996; Cabral and Ferreira 1999; Moreira and Martins 2005). However, the patterns of natural selection found seem to be extremely complex, with different selective forces acting on different geographical scales and regions, which may hamper our ability to detect clear signatures of natural selection. Nevertheless, the findings here reported may provide seminal information for the identification of functionally important adaptive genetic variation within natural populations of cork oak, which is of major importance for the definition of management and conservation strategies for this relevant species.

Acknowledgments We thank José Conde (CISE - Centro de Interpretação da Serra da Estrela, Seia, Portugal) for the indispensable support in the exhaustive survey and sample collection at Serra da Estrela, and Maria Helena Almeida (Instituto Superior de Agronomia, UTL) for sample access and advice regarding the international provenance trial (FAIR I CT 95 0202) established at Monte Fava, Alentejo, Portugal. This work was funded by Portuguese funds through FCT - Fundação para a

Ciência e Tecnologia (projects PTDC/AGR-GPL/104966/2008 and SOBREIRO/0036/2009). We thank the anonymous reviewers for their constructive comments and suggestions which helped us to improve the manuscript.

Data archiving statement Sequence data has been submitted to GenBank (National Center for Biotechnology Information) and can be accessed through the accession numbers KF988869-KF989346. A complete list of accession numbers is provided in Online Resource 2.

References

Alberto FJ, Derory J, Boury C, Frigerio JM, Zimmermann NE, Kremer A (2013) Imprints of natural selection along environmental gradients in phenology-related genes of *Quercus petraea*. Genetics 195(2): 495–512

Alexandrov AH, Genov K, Popov E (2001) Country reports: Bulgaria. In: Borelli S, Vare MC (eds) Mediterranean oaks network, Report of the first meeting, 12–14 October 2000, Antalya, Turkey. IPGRI, Rome, Italy

Andrews M, Raven JA, Lea PJ (2013) Do plants need nitrate? The mechanisms by which nitrogen form affects plants. Ann Appl Biol 163:174–199

Aranda I, Castro L, Alia R, Pardos JA, Gil L (2005) Low temperature during winter elicits differential responses among populations of the Mediterranean evergreen cork oak (*Quercus suber*). Tree Physiol 25(8):1085–1090

Aranda I, Pardos M, Puertolas J, Jimenez MD, Pardos JA (2007) Water-use efficiency in cork oak (*Quercus suber*) is modified by the interaction of water and light availabilities. Tree Physiol 27(5): 671–677

Austerlitz F, Mariette S, Machon N, Gouyon PH, Godelle B (2000) Effects of colonization processes on genetic diversity: Differences between annual plants and tree species. Genetics 154(3):1309–1321

Bernard SM, Habash DZ (2009) The importance of cytosolic glutamine synthetase in nitrogen assimilation and recycling. New Phytol 182(3):608–620

Biswas S, Akey JM (2006) Genomic insights into positive selection. Trends Genet 22(8):437–446

Bondino HG, Valle EM, ten Have A (2012) Evolution and functional diversification of the small heat shock protein/alpha-crystallin family in higher plants. Planta 235(6):1299–1313

Brasier CM (1996) *Phytophthora cinnamomi* and oak decline in southern Europe. Environmental constraints including climate change. Ann Sci For 53(2–3):347–358

Cabral MT, Ferreira MC (1999) Pragas dos Montados. Estação Florestal Nacional, Lisboa

Caldwell KS, Michelmore RW (2009) *Arabidopsis thaliana* genes encoding defense signaling and recognition proteins exhibit contrasting evolutionary dynamics. Genetics 181(2):671–684

Coca MA, Almoguera C, Jordano J (1994) Expression of sunflower low-molecular-weight heat-shock proteins during embryogenesis and persistence after germination-localization and possible functional implications. Plant Mol Biol 25(3):479–492

Costa J, Miguel C, Almeida H, Oliveira MM, Matos JA, Simões F, Veloso M, Pinto Ricardo C, Paulo OS, Batista D (2011) Genetic divergence in Cork Oak based on cpDNA sequence data. IUFRO Tree Biotechnology Conference 2011: from genomes to integration and delivery. BMC Proc 5:(Suppl 7), 13

Davis MB, Shaw RG (2001) Range shifts and adaptive responses to Quaternary climate change. Science 292(5517):673–679

Derory J, Leger P, Garcia V, Schaeffer J, Hauser MT, Salin F, Luschnig C, Plomion C, Glossl J, Kremer A (2006) Transcriptome analysis of bud burst in sessile oak (*Quercus petraea*). New Phytol 170(4):723–738

Derory J, Scotti-Saintagne C, Bertocchi E, Le Dantec L, Graignic N, Jauffres A, Casasoli M, Chancerel E, Bodenes C, Alberto F, Kremer A (2010) Contrasting relations between diversity of candidate genes and variation of bud burst in natural and segregating populations of European oaks. Heredity 105(4):401–411

Ding Z, Friml J (2010) Auxin regulates distal stem cell differentiation in *Arabidopsis* roots. Proc Natl Acad Sci U S A 107(26):12046–12051

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10(3):564–567

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity 103:285–298

Flot JF (2007) CHAMPURU 1.0: a computer software for unraveling mixtures of two DNA sequences of unequal lengths. Mol Ecol Notes 7(6):974–977

Flot JF, Tillier A, Samadi S, Tillier S (2006) Phase determination from direct sequencing of length-variable DNA regions. Mol Ecol Notes 6(3):627–630

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147(2): 915–925

Gandour M, Khouja ML, Toumi L, Triki S (2007) Morphological evaluation of cork oak (*Quercus suber*): Mediterranean provenance variability in Tunisia. Ann For Sci 64(5):549–555

Gil L, Varela MC (2008) Cork oak (*Quercus suber*). In: EUFORGEN (ed) Technical guidelines for genetic conservation and use. IPGRI, Rome

Giorgi F, Lionello P (2008) Climate change projections for the Mediterranean region. Global Planet Change 63(2–3):90–104

Haizel T, Merkle T, Pay A, Fejes E, Nagy F (1997) Characterization of proteins that interact with the GTP-bound form of the regulatory GTPase ran in *Arabidopsis*. Plant J 11(1):93–103

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids 41:95–98

Higgins PAT, Harte J (2006) Biophysical and biogeochemical responses to climate change depend on dispersal and migration. Bioscience 56: 407–417

Hijmans RJ, Guarino L, Cruz M, Rojas E (2001) Computer tools for spatial analysis of plant genetic resources data: 1 DIVA-GIS. Plant Genet Resour Newslett 127:15–19

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25(15):1965–1978

Homolka A, Schueler S, Burg K, Fluch S, Kremer A (2013) Insights into drought adaptation of two European oak species revealed by nucleotide diversity of candidate genes. Tree Genet Genomes 9(5):1179–1192

Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, Muliyati NW, Platt A, Sperone FG, Vilhjalmsson BJ, Nordborg M, Borevitz JO, Bergelson J (2012) Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. Nat Genet 44(2):212–216

Hungria M, Kaschuk G (2013) Regulation of N_2 fixation and NO_3–/NH_4+ assimilation in nodulated and N-fertilized Phaseolus vulgaris L. exposed to high temperature stress. Environ Exp Bot 98:32–39

IPCC (2007) Climate Change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change. IPCC Secretariat, Geneva

Joost S, Bonin A, Bruford MW, Despres L, Conord C, Erhardt G, Taberlet P (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. Mol Ecol 16(18):3955–3969

Librado P, Rozas J (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452

Liu ZY, Jia LG, Wang H, He YK (2011) HYL1 regulates the balance between adaxial and abaxial identity for leaf flattening via miRNA-mediated pathways. J Exp Bot 62(12):4367–4381

Magri D, Fineschi S, Bellarosa R, Buonamici A, Sebastiani F, Schirone B, Simeone MC, Vendramin GG (2007) The distribution of *Quercus suber* chloroplast haplotypes matches the palaeogeographical history of the western Mediterranean. Mol Ecol 16(24):5259–5266

Marshall OJ (2004) PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. Bioinformatics 20(15): 2471–2472

Meier I, Brkljacic J (2010) The *Arabidopsis* nuclear pore and nuclear envelope. Arabidopsis Book//Am Soc Plant Biol 8:139

Merkle T (2011) Nucleo-cytoplasmic transport of proteins and RNA in plants. Plant Cell Rep 30:153–176

Moreira AC, Martins JMS (2005) Influence of site factors on the impact of *Phytophthora cinnamomi* in cork oak stands in Portugal. For Pathol 35(3):145–162

Murphy EL (1985) Bergmann's rule, seasonality and geographic variation in body size of house sparrows. Evolution 39:1327–1334

Niderman T, Genetet I, Bruyere T, Gees R, Stintzi A, Legrand M, Fritig B, Mosinger E (1995) Pathogenesis-related PR-1 proteins are antifungal - Isolation and characterization of three 14-kilodalton proteins of tomato and of a basic PR-1 of tobaco with inhibitory activity against *Phytophthora infestans*. Plant Physiol 108(1):17–27

Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197–218

Nunes A, Almeida M, Monteiro M, Patricio M (2008) Reslutados preliminares em ensaios genéticos de sobreiro. FPFP - Federação dos Produtores Florestais de Portugal, Lisboa

Otieno DO, Kurz-Besson C, Liu J, Schmidt MWT, Do RVL, David TS, Siegwolf R, Pereira JS, Tenhunen JD (2006) Seasonal variations in soil and plant water status in a *Quercus suber* L. Stand: roots as determinants of tree productivity and survival in the Mediterranean-type ecosystem. Plant Soil 283(1–2):119–135

Pausas JG, Pereira JS, Aronson J (2009) The tree. In: Aronson J, Pereira JS, Pausas JG (eds) Cork oak woodlands on the edge. Island Press, Washington DC, pp 11–21

Pereira-Leal JB, Abreu IA, Alabaça CS, Almeida MH, Almeida P, Almeida T, Amorim MI, Araújo S, Azevedo H, Badia A, Batista D, Bohn A, Capote T, Carrasquinho I, Chaves I et al (2014) A comprehensive assessment of the transcriptome of cork oak (*Quercus suber*) through EST sequencing. BMC Genomics (Accepted)

Pieterse CM, Van Loon L (2004) NPR1: the spider in the web of induced resistance signaling pathways. Curr Opin Plant Biol 7(4):456–464

Puigderrajols P, Jofre A, Mir G, Pla M, Verdaguer D, Huguet G, Molinas M (2002) Developmentally and stress-induced small heat shock proteins in cork oak somatic embryos. J Exp Bot 53(373):1445–1452

Pulido FJ, Diaz M (2005) Regeneration of a Mediterranean oak: a whole-cycle approach. Ecoscience 12(1):92–102

Quang ND, Ikeda S, Harada K (2008) Nucleotide variation in *Quercus crispula* Blume. Heredity 101(2):166–174

Ramirez-Valiente JA, Lorenzo Z, Soto A, Valladares F, Gil L, Aranda I (2009a) Elucidating the role of genetic drift and natural selection in cork oak differentiation regarding drought tolerance. Mol Ecol 18(18):3803–3815

Ramirez-Valiente JA, Valladares F, Gil L, Aranda I (2009b) Population differences in juvenile survival under increasing drought are mediated by seed size in cork oak (*Quercus suber* L.). For Ecol Manag 257(8):1676–1683

Ramirez-Valiente JA, Lorenzo Z, Soto A, Valladares F, Gil L, Aranda I (2010) Natural selection on cork oak: allele frequency reveals divergent selection in cork oak populations along a temperature cline. Evol Ecol 24(5):1031–1044

Ramirez-Valiente AJ, Valladares F, Delgado Huertas A, Granados S, Aranda I (2011) Factors affecting cork oak growth under dry conditions: local adaptation and contrasting additive genetic variance within populations. Tree Genet Genomes 7(2):285–295

Rauscher M, Adam AL, Wirtz S, Guggenheim R, Mendgen K, Deising HB (1999) PR-1 protein inhibits the differentiation of rust infection hyphae in leaves of acquired resistant broad bean. Plant J 19(6):625–633

Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). Mol Ecol 19:115–131

Sato Y, Yokoya S (2008) Enhanced tolerance to drought stress in transgenic rice plants overexpressing a small heat-shock protein, sHSP17.7. Plant Cell Rep 27(2):329–334

Shearer HL, Wang LP, DeLong C, Despres C, Fobert PR (2009) NPR1 enhances the DNA binding activity of the *Arabidopsis* bZIP transcription factor TGA7. Botany 87(6):561–570

Simeone MC, 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 VL, Davis FW, Westfall R, Flint A, Ikegami M, Wang HF, Grivet D (2010) Gene movement and genetic association with regional climate gradients in California valley oak (*Quercus lobata* Nee) in the face of climate change. Mol Ecol 19(17):3806–3823

Soto A, Lorenzo Z, Gil L (2007) Differences in fine-scale genetic structure and dispersal in *Quercus ilex* L. and *Q. suber* L.: consequences for regeneration of Mediterranean open woods. Heredity 99(6):601–607

Sousa E, Santos M, Varela M, Henriques J (2007) Perda de vigor dos montados de sobro e azinho: Análise da situação e perspectivas. Documento Sintese. Eds MADRP, DGRF, INRB, Lisbon

Spiess N, Oufir M, Matusikova I, Stierschneider M, Kopecky D, Homolka A, Burg K, Fluch S, Hausman J-F, Wilhelm E (2012) Ecophysiological and transcriptomic responses of oak (*Quercus robur*) to long-term drought exposure and rewatering. Environ Exp Bot 77:117–126

Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76(3):449–462

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68(4):978–989

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595

Teixeira J, Pereira S (2007) High salinity and drought act on an organ-dependent manner on potato glutamine synthetase expression and accumulation. Environ Exp Bot 60:121–126

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal W—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673–4680

Trinajstic I (2005) Hrast plutnjak (Quercus suber L.) u dendroflori Hrvatske. Rad Šumar Inst 40:199–206

Valladares F, Gianoli E, Gomez JM (2007) Ecological limits to plant phenotypic plasticity. New Phytol 176(4):749–763

Varela MC (2000) Handbook of the EU concerted action on cork oak, FAIR I CT 95 0202. INIA- Estação Florestal Nacional, Oeiras

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol 17(7):1636–1647

Wang WX, Vinocur B, Shoseyov O, Altman A (2004) Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. Trends Plant Sci 9(5):244–252

Wang JW, Wang LJ, Mao YB, Cai WJ, Xue HW, Chen XY (2005) Control of root cap formation by microRNA-targeted auxin response factors in *Arabidopsis*. Plant Cell 17(8):2204–2216

Yan SP, Tang ZC, Su W, Sun WN (2005) Proteomic analysis of salt stress-responsive proteins in rice root. Proteomics 5:235–244

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24(8):1586–1591

Yang ZH, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22(4):1107–1118

Zhai W, Nielsen R, Slatkin M (2009) An investigation of the statistical power of neutrality tests based on comparative and population genetic data. Mol Biol Evol 26(2):273–283

Zhou JM, Trifa Y, Silva H, Pontier D, Lam E, Shah J, Klessig DF (2000) NPR1 differentially interacts with members of the TGA/OBF family of transcription factors that bind an element of the PR-1 gene required for induction by salicylic acid. Mol Plant Microbe Interact 13(2):191–202