UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA ANIMAL


# Assessing adaptive genetic variation in cork oak expressed genes

Inês Sofia Barrios Modesto

DISSERTAÇÃO

MESTRADO EM BIOLOGIA EVOLUTIVA E DO DESENVOLVIMENTO

2012

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA ANIMAL


# Assessing adaptive genetic variation in cork oak expressed genes

Inês Sofia Barrios Modesto

DISSERTAÇÃO

MESTRADO EM BIOLOGIA EVOLUTIVA E DO DESENVOLVIMENTO

Orientadores:

Professor Doutor Octávio Paulo

Doutora Dora Batista

2012

## Nota prévia

Este trabalho foi realizado no âmbito dos projectos SOBREIRO/0036/2009 "Polymorphism detection and validation", incluído no Cork Oak ESTs (expressed sequence tags) Consortium (COEC), e PTDC/AGR-GPL/104966/20082008 "Assessment of genetic and genomic resources of Cork Oak: the basis towards a prospective management".

O trabalho foi escrito em inglês e em forma de artigo para facilitar a posterior publicação e divulgação científica. A formatação da bibliografia encontra-se de acordo com a revista científica internacional *Molecular Ecology*.

## Agradecimentos

No final desta tese, não poderia deixar de agradecer a todos os que contribuíram para a sua realização e para que chegasse a bom porto:

Em primeiro lugar, gostaria de agradecer aos meus orientadores, Professor Octávio Paulo e Doutora Dora Batista, por tornaram possível a realização deste projecto, pelas discussões e interpretações do trabalho, transmissão de conhecimentos e pelas várias sugestões que se revelaram essenciais para realizar esta tese e para o meu crescimento como cientista e investigadora.

Agradeço à Doutora Célia Miguel pela sua colaboração e pelas suas sugestões e perspectivas diferentes que foram fundamentais para enriquecer este trabalho.

À Doutora Manuela Veloso pela sua colaboração neste projecto e pela sua disponibilidade.

Aos meus colegas do CoBiG$^2$ pela amizade, companheirismo e momentos de boa disposição e por toda a ajuda que me deram a nível do laboratório, da análise de dados, das sugestões que deram em relação a este trabalho e pelas discussões entusiasmantes sobre vários temas da biologia evolutiva.

Queria agradecer em especial à Vera Nunes por me ajudar nos primeiros passos no laboratório e pelas sugestões, discussões e perspectivas sempre perspicazes e relevantes, que contribuíram para o meu crescimento como cientista e, como tal, para o enriquecimento deste trabalho. À Joana Costa, Bruno Vieira, Eduardo Marabuto, Carla Silva, Francisco Pina-Martins e Rui Nunes pela ajuda com a amostragem em campo. À Carla Silva pela ajuda na maceração e extração das amostras, pelos conhecimentos de laboratório que me transmitiu e pelos momentos de amizade, conversa e boa disposição. À Joana Costa pelos conhecimentos que me transmitiu sobre o trabalho laboratorial. Ao Francisco Pina-Martins pela ajuda que me deu, principalmente no início do trabalho de validação de SNPs e na compreensão dos métodos de NGS e 454. À Renata Martins, que me acompanhou de perto todos os dias durante o primeiro ano da tese, principalmente no laboratório, agradeço especialmente pela amizade e boa disposição que me transmitiu. Ao Diogo Silva pela paciência que teve em me ensinar e ajudar com quase todas as análises de dados. À Carla Silva que também me ajudou com parte da

análise de dados. À Patrícia Brás pela ajuda e discussão sobre vários aspectos da análise estatística de dados. À Catarina Dourado por estar sempre disposta a ajudar e resolver problemas, seja a nível do laboratório, da análise de dados, ou mesmo da parte gráfica da tese. Ao Eduardo Marabuto pelas fotos fantásticas de sobreiro que me cedeu. Á Sara Silva, Ana Sofia, Sofia Seabra, Vera Nunes, Catarina Dourado, Carla Silva, Eduardo Marabuto e Franscisco Pina-Martins por todas as correcções e sugestões que deram em relação à tese, por estarem sempre dispostos a ajudar e principalmente pelo apoio que me deram e por me aturarem nesta fase final do trabalho.

Agradeço à Ana Vieira pela amizade e por todas as vezes que me deu boleia para Oeiras, tornando as manhãs sempre mais animadas e bem-dispostas.

Agradeço também aos meus colegas do ITQB: Andreia Matos, Andreia Miguel, Andreia Rodrigues, Liliana Marum, Ana Milhinhos, Inês Chaves e José Bartol. Obrigada por me aturarem neste segundo ano da tese, pela ajuda que me deram no laboratório e pelos momentos de boa disposição e descontração, especialmente durante os almoços.

Não podia deixar de agradecer também aos meus amigos, que me apoiaram durante todo este processo, em especial à Patrícia Nunes, Inês Gonçalves, Ana Filipa Dias e à minha madrinha Joana Serra.

Ao Gonçalo, cujo apoio foi essencial para que eu conseguisse fazer e terminar esta tese, quero deixar um agradecimento especial. Sem as suas palavras de “força nisso” e “ tu consegues”, sem o seu amparo nos momentos mais difíceis e sem a calma que me transmitiu não conseguiria ter chegado até aqui.

Por fim, agradeço a minha família por me ter possibilitado frequentar este mestrado, por me ter acompanhado de perto durante todo este processo e pelo apoio incondicional que me deu.

## Resumo

O sobreiro (*Quercus suber* L.) é uma árvore de folha perene endémica do Oeste da Bacia do Mediterrâneo. A sua área de distribuição natural estende-se desde a costa atlântica da Península Ibérica e de Marrocos até ao Sudeste da Península Itálica, sendo em Portugal que se encontram os maiores povoamentos de sobreiro. É possível também encontrar esta espécie na Bulgária, como resultado de introdução humana. O sobreiro caracteriza-se principalmente pela cortiça que produz, continuamente e de forma renovável, a qual protege a árvore de factores externos agressivos, tais como o fogo. A cortiça tem propriedades físico-químicas únicas que lhe conferem um elevado valor comercial, contribuindo de modo significativo para a economia dos países onde o sobreiro se encontra naturalmente distribuído e onde é comercialmente cultivado e explorado. Portugal é o maior produtor mundial de cortiça, tendo por isso o sobreiro um grande valor económico e social a nível nacional. Do ponto de vista ecológico, detém também uma grande importância, uma vez que o montado de sobro (ou *dehesa* em Espanha) constitui um sistema de características únicas, essencial para a sobrevivência de um grande número de espécies nativas de plantas e animais e para a prevenção da desertificação. Dada a sua grande relevância nestes vários aspectos, o sobreiro foi considerado recentemente "Árvore Nacional de Portugal", o que expressa a sua enorme importância cultural e patrimonial para o país.

Os povoamentos de sobreiro têm vindo a sofrer um declínio devido à falta de regeneração natural, atribuída essencialmente a períodos de seca intensa, à dependência de árvores adultas envelhecidas e à má gestão do montado. Para além disso, estão previstas para este século grandes alterações climáticas, com especial impacto na Bacia Mediterrânica, que podem vir a aumentar a pressão sobre as populações de sobreiro. Nesta região, espera-se um grande aumento das temperaturas e períodos de seca mais severos e prolongados. Prevê-se que estas alterações ocorram numa escala temporal tão curta que as espécies florestais, incluindo o sobreiro, poderão não conseguir acompanhá-las. O declínio do sobreiro tem sido também associado a várias pragas e doenças, principalmente à doença da tinta, provocada pelo fungo *Phytophthora cinnamomi*. Deste modo, é essencial e premente estudar a variabilidade genética do sobreiro, de modo a compreender a sua capacidade adaptativa a factores bióticos e abióticos, tendo em vista o delineamento de estratégias de gestão e conservação dos

recursos genéticos desta espécie. Assim, o estudo da variabilidade genética adaptativa do sobreiro constitui o principal objectivo deste trabalho.

Num estudo prévio, foi sequenciado o transcriptoma de folhas de sobreiro através da tecnologia de pirosequenciação 454, o qual foi subsequentemente analisado para detectar *single nucleotide polymorphisms* (SNPs). Mais de 400 SNPs putativos foram encontrados em regiões transcritas do genoma, os quais podem ser de grande interesse, uma vez que permitem obter informação sobre mutações em genes funcionais, possivelmente sob selecção. No entanto, as sequências obtidas por 454 podem conter erros e, para poder explorar os SNPs detectados, é importante proceder primeiro à sua validação para confirmar se esta variação corresponde a polimorfismos reais ou se, por outro lado, resulta de artefactos da tecnologia 454. O primeiro objectivo deste trabalho consistiu assim na validação de 10 a 15 SNPs putativos, de forma a desenvolver marcadores moleculares úteis, possivelmente sob selecção. O segundo objectivo consistiu em analisar cinco dos SNPs validados, do ponto de vista filogeográfico e da genética populacional, compreendendo a realização de vários testes de neutralidade para detectar sinais de selecção. Por fim, teve-se como terceiro objectivo testar associações entre a variabilidade genética encontrada e variáveis ambientais potencialmente relevantes para a adaptação local do sobreiro. A combinação de vários testes de neutralidade e métodos de associação ambiental é importante para estudos de selecção e adaptação, uma vez que diferentes métodos estatísticos têm diferentes capacidades e sensibilidades na detecção de selecção natural.

Tendo em conta estes objectivos, desenharam-se *primers* para amplificar fragmentos de ADN genómico contendo os SNPs putativos e subsequentemente validá-los através de sequenciação Sanger. As amostras utilizadas nesta fase foram as mesmas previamente usadas para a pirosequenciação do transcriptoma. Dos SNPs validados, cinco foram escolhidos para uma análise mais detalhada dos respectivos fragmentos. Em primeiro lugar, foram estimadas as relações entre os haplótipos detectados para cada fragmento através da construção de redes haplotípicas. Em seguida foram efectuadas análises de variância molecular (AMOVA) com o intuito de aferir a estruturação populacional. A variabilidade genética e haplotípica foram estimadas para cada fragmento e vários testes de neutralidade foram efectuados, nomeadamente o *D* de Tajima, o *Fs* de Fu e o teste baseado na comparação entre a taxa de mutações não-sinónimas por posições não-sinónimas ($d_N$) e taxa de mutações sinónimas por posições sinónimas ($d_S$) implementado

pelo software PAML. Por fim, foram feitos testes de associação entre os dados genéticos e diversas variáveis ambientais através de regressões logísticas implementadas pelo software MatSAM.

Em relação ao processo de validação dos SNPs putativos, foi possível amplificar e sequenciar 59% dos fragmentos testados (19 de 32), um valor superior ao descrito noutros estudos similares. Do conjunto de SNPs testados, 11 foram validados, três foram invalidados, e para os restantes cinco não foi possível sequenciar todas as amostras utilizadas na sequenciação do transcriptoma, pelo que, apesar de ainda não se ter encontrado variação, não é possível chegar a qualquer conclusão em relação à sua autenticidade. Assim, obteve-se uma percentagem de 79% de sucesso de validação, um valor não muito inferior ao reportado em trabalhos anteriores.

Dos fragmentos validados, cinco foram escolhidos para subsequente análise. Esses fragmentos encontram-se nos potenciais genes ortólogos de *Arabidopsis thaliana* (L.) Heynh. *RAN3*, *NPR1*, *PR1*, *ARF16* e *HSP*. Para nenhum dos fragmentos analisados foi detectada estruturação geográfica da variabilidade genética, o que pode estar relacionado com o facto de os sobreiros serem organismos longevos, submetidos a condições ambientais variáveis ao longo da sua vida, e terem uma dispersão de pólen pelo vento a longas distâncias, características que normalmente levam a uma baixa estruturação entre populações ao nível de marcadores nucleares.

Em dois dos fragmentos analisados (*NPR1* e *ARF16*) foram encontrados sinais de selecção balanceada, tendo sido estimados para ambos valores significativamente positivos de *D* de Tajima e de *Fs* de Fu. Ambos apresentam ainda padrões filogeográficos e de distribuição geográfica da variabilidade genética semelhantes, sem estruturação. O *NPR1* é um gene envolvido na defesa contra agentes patogénicos, tais como *P. cinnamomi*. As mutações não-sinónimas e não-conservativas (cujo aminoácido mutado tem propriedades físico-químicas diferentes) detectadas neste fragmento encontram-se no domínio *ankirin repeats* (ANK), essencial na activação de factores de transcrição responsáveis pelo controlo de genes de defesa contra agentes patogénicos. Assim, estas mutações podem estar sob pressão selectiva exercida por este stresse biótico, sendo por isso mantidos os polimorfismos. Por outro lado, sinais de selecção balanceada foram encontrados num estudo prévio no mesmo domínio do *NPR1* de *A. thaliana*, corroborando a hipótese deste gene se encontrar sob este tipo de selecção. Em

relação ao *ARF16*, este gene codifica um factor de transcrição envolvido na diferenciação da coifa e foi identificado como gene candidato à resistência à seca em *Quercus robur* L. As mutações não-sinónimas e não-conservativas detectadas neste fragmento encontram-se no domínio da proteína responsável pela sua acção como factor de transcrição, podendo desta forma ter efeitos a nível da transcrição dos genes activados pela proteína ARF16. Assim, estas mutações podem estar sob pressão selectiva exercida por condições ambientais flutuantes ao longo da vida do sobreiro, mantendo o polimorfismo neste gene.

No fragmento do gene *HSP* foi detectada uma posição com sinal de selecção positiva (análise com PAML). Na mesma posição, a frequência do aminoácido aspartato foi positivamente correlacionada com precipitação em Setembro e latitude através dos estudos de associação. As proteínas HSP encontram-se geralmente associadas a condições de stresse, sendo provável que esta esteja envolvida na resposta a stresse hídrico. Setembro é usualmente o primeiro mês após o período de seca mais intensa, pelo que a precipitação neste mês será provavelmente importante para as plantas recuperarem desse período. Deste modo, árvores detentoras de genótipos com aspartato na posição em questão parecem estar mal-adaptadas a períodos de seca prolongada.

Neste estudo foram desenvolvidos marcadores moleculares úteis para compreender a variabilidade genética adaptativa do sobreiro, confirmando-se a utilidade da pirosequenciação do transcriptoma de organismos não-modelo para este fim. Além disso, com esses marcadores foi possível adquirir novos conhecimentos em relação à adaptação do sobreiro ao meio ambiente, o que é essencial para que se possa definir estratégias de gestão e conservação dos recursos genéticos desta espécie.

**Palavras-chave:** adaptação, *Quercus suber*, selecção balanceada, selecção positiva, sobreiro, SNP, validação

## Abstract

Cork oak (*Quercus suber* L.) is an evergreen tree endemic to the Western Mediterranean region with great economical, social and ecological relevance. It has a particular importance in Portugal, where the largest stands can be found. Cork oak stands have been facing a significant decline by the lack of regeneration, mostly due to severe drought periods, the dependence on aged adult trees and bad management. In the context of the climate changes predicted for this century, drought periods are expected to be increasingly longer and more rigorous in the Mediterranean Basin, which can enhance this decline. Moreover, several diseases have also been associated with cork oak populations' decay. In this scenario, evaluating how cork oak populations can cope with these threats is essential to delineate management and conservation strategies for this species. The main goal of this work was therefore to assess cork oak adaptive genetic variation. Putative SNPs detected in cork oak transcriptome were validated in order to develop useful variable markers in functional genes potentially under selection. Five fragments containing validated SNPs were further investigated through a population genetics approach. Several neutrality tests were performed as well as environmental association tests in order to find selection signatures. Different selection signals were detected in the analysed fragments. *NPR1*, a gene involved in plant defence against pathogens, and *ARF16*, a gene implicated in root cap cell differentiation and previously identified as a candidate gene for drought resistance, seemed to be under balancing selection. In *HSP*, a gene possibly involved in response to drought stress, one amino acid position was detected as possibly being under positive selection and associated with latitude and precipitation in September. Therefore, in this study, useful molecular markers for assessing cork oak adaptive genetic variation were developed, allowing for the first steps to be taken into gathering important information and insights on cork oak adaptation to biotic and abiotic environmental conditions.

**Keywords:** adaptation, balancing selection, positive selection, *Quercus suber*, SNP, validation

**Table of contents**

# CHAPTER 1

# General Introduction

## 1.1. Forest trees as models

Forests represent approximately 82% of Earth's biomass (Roy *et al.* 2001), covering approximately 27% of its terrestrial surface. Therefore, it comes as no surprise that they harbour more than 50% of terrestrial biodiversity and play an important role in carbon sequestration, climate regulation and water quality preservation. Forest trees are also very important to humans, providing a variety of essential resources, such as building materials, paper products, firewood, energy and tree-crop foods (Neale and Kremer 2011). Global tree species richness is estimated to range between 60,000 and 100,000 species (Oldfield *et al.* 1998, Grandtner 2005). However, deforestation and other human induced changes have put more than 10% of these species on the path to extinction (Oldfield *et al.* 1998). In this scenario, understanding adaptive genetic variation in forest trees is essential to delineate management and conservation strategies (Krutovsky and Neale 2005) and can be helpful for the improvement of economically important species in combination with traditional phenotypic selection (Neale 2007).

Despite their relevance, forest trees are rarely seen as models in most plant biology lines of research, as their size and life span make them difficult to use in experimental studies (Linhart 1999). However, in recent years they have been gaining attention as non-classical models in population, evolution and ecological genomics studies (Gonzalez-Martinez *et al.* 2006, Neale and Ingvarsson 2008, Neale and Kremer 2011). From an evolutionary point of view, forest trees have several features that make them great models to study adaptive divergence. They usually have outcrossing populations with large effective sizes, high levels of genetic and phenotypic diversity and low population structure (Gonzalez-Martinez *et al.* 2006, Petit and Hampe 2006, Gailing *et al.* 2009). Furthermore, forest trees are long-lived and sessile, growing under temporally varying and spatially heterogeneous environmental conditions, which contributes to diversity maintenance. This high genetic diversity is essential for adaptation to changing environments (Gailing *et al.* 2009) whereas their life-spans and resilience provide them the means to withstand short environmental pulses. In addition, forest trees are unique in the way that they can be found in both domesticated and wild forms, which provides extensive experimental opportunities and the chance to exploit the genetic diversity found within natural populations in selective breeding and genetic improvement (Neale and Ingvarsson 2008, Neale and Kremer 2011).

### 1.1.1. The Fagaceae family

The Fagaceae are a large angiosperm family comprising 8-10 genera and approximately 900 woody species spread throughout the Northern Hemisphere (Singh 2004). They cover large continuous forests and represent important forest resources, with many species being economically important. Furthermore, they play a central role in forest ecosystems, providing the maintenance of terrestrial biodiversity. In many countries, they are also considered important patrimonial and cultural resources (Logan 2005). The only genera in Europe are oaks (*Quercus*), chestnuts (*Castanea*) and beeches (*Fagus*) (Singh 2004).

Oaks occur mainly in temperate, subtropical and semiarid biotopes and have been divided in two clades (Manos *et al.* 1999, Bellarosa *et al.* 2005, Denk and Grimm 2010). One clade comprises the New World species, with sections *Lobatae*, *Protobalanus* and the holarctic *Quercus s.s.*, while the other strictly Eurasian clade, comprises the *Cerris* group, that includes species from temperate and semiarid regions (e.g. *Quercus ilex*, *Quercus suber*) (Denk and Grimm 2010) and *Cyclobalanopsis* group from tropical regions (Deng 2007). The wide ranging distribution and varied life-strategies of these species, growing under a great range of climatic and edaphic conditions offer unprecedented opportunities to investigate the genetic basis of adaptive traits.

Oaks have been considered good model species to study adaptation of forest trees to changing environments (Gailing *et al.* 2009), since they are widely distributed throughout Europe as dominant tree species in many forests. Moreover, genomic resources are increasingly becoming available for this genus (e.g. Derory *et al.* 2006, Soler *et al.* 2007, Ueno *et al.* 2010).

### 1.1.2. Cork oak as a case study

Cork oak (*Quercus suber* L.) (Figure 1.1) is a Mediterranean slow growth and extremely long lived (200-250 years) evergreen tree. It is a monoecious (separate male and female flowers on the same plant) wind-pollinated species with a protandrous system to ensure cross-pollination. As most oaks, cork oak is propagated mainly via

sexual reproduction (natural regeneration) and seed (acorn) (Figure 1.1b) dissemination occurs essentially through gravity and zoochory. The striking characteristic of this tree is its thick, soft and porous bark, the cork (Figure 1.1c), which is continuously and renewably produced, protecting the trees from external damaging factors including forest fires (Gil and Varela 2008).


**Figure 1.1** Cork oak (a), its fruit, the acorn (b) and detail of its bark, the cork (c). Photos by Eduardo Marabuto.

Cork oak current natural distribution is rather discontinuous throughout the Western Mediterranean region (Figure 1.2) (Pausas *et al.* 2009), ranging from the Atlantic Coast of North Africa and Iberian Peninsula to southeastern Italy, including the larger west Mediterranean islands and coastal regions of Maghreb (Algeria and Tunisia), Provence (France) and Catalonia (Spain). In addition, cork oak is present in a somewhat naturalized state in Bulgaria, as a result of recent introductions possibly from Portugal and/or Spain (Alexandrov *et al.* 2001). The largest stands of cork oak are located in its western distribution limit, particularly in Portugal, where it reaches an area of 736,700 hectares (http://www.realcork.org/artigo.php?art=401). This restricted distribution when compared to other oaks is mainly due to cork oak being highly constrained by soil preferences (Serrasolses *et al.* 2009) and by low winter temperatures (Pausas *et al.* 2009). It has a strict acidophilus character, growing mainly on non-calcareous substrates, preferring siliceous and lightly structured soils (Serrasolses *et al.* 2009).

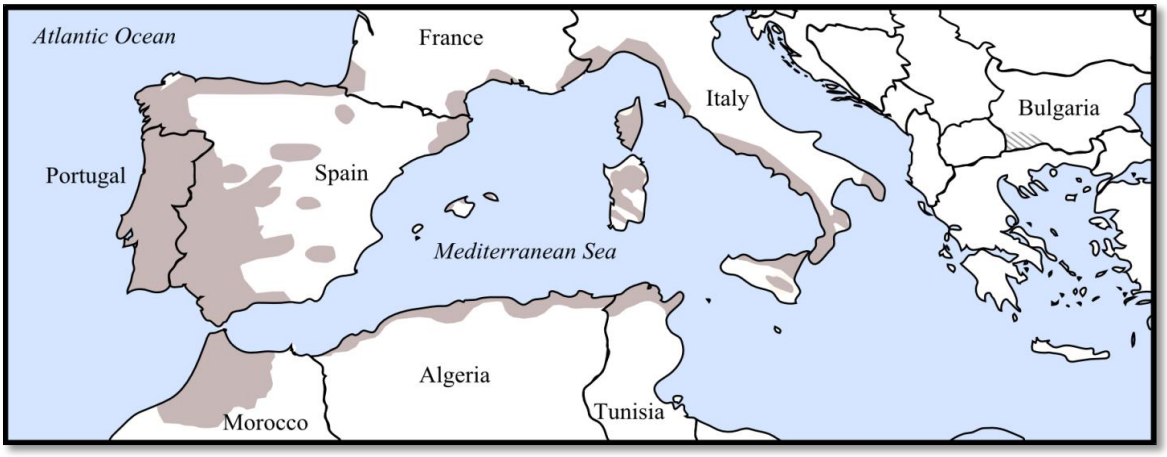



**Figure 1.2.** Map of the present geographical distribution of Cork oak (*Quercus suber*). Full gray represents natural distribution; dashed gray represents introduced, somewhat naturalized populations.

Cork oak has long been explored for the extraction of cork in the countries where it is naturally distributed. This material has unique physic and chemical properties suited for various purposes, such as the production of bottle stoppers, thermal and sound insulation in floors and walls, decoration, products for military and aviation uses, among others. Therefore, cork oak has a great economical and social importance where it is produced, particularly in Portugal, which is responsible for 55% of the world cork production and processing (http://www.amorim.com/cor_glob_cortica.php). This tree also has a great ecological relevance, as its woods are managed by man under a unique ecological system in the world, known as *montado* in Portugal and *dehesa* in Spain. These are open woods with low tree density (50-300 trees/ha) specifically managed for cork production, as well as other goods such as cattle and hunting. These systems contribute to the survival of many native plant and animal species and prevent the desertification of sensitive areas (Gil and Varela 2008). Giving its relevance at these several levels, it is not surprising that cork oak has been recently classified as "National Tree" in Portugal, revealing also its importance from a cultural and patrimonial perspective in this country (Republic of Portugal "Projecto de Resolução N.º 123/XII/1.ª").

Despite its restricted distribution relatively to other oaks, cork oak grows under a wide range of climatic conditions. Therefore, different populations are expected to be under

different selective pressures. Several studies report differences in phenotypic and functional traits among distinct populations in common garden experiments (e.g. Aranda *et al.* 2005, Gandour *et al.* 2007, Ramirez-Valiente *et al.* 2009b, 2010a). In Aranda *et al.* (2005), differential responses to low temperatures were reported, with populations from cold regions displaying more tolerance to this type of stress, whereas Ramírez-Valiente *et al.* (2009) showed that northern populations were not well adapted to drought and that continental populations were intermediately adapted to dry conditions. Some of the ecophysiological traits correlated with environment variables were shown to be heritable, supporting the idea that natural selection has led to local adaptation (Ramirez-Valiente *et al.* 2011). Furthermore, a microssatellite (*Qp*ZAG46) was found to be correlated with leaf size (Ramirez-Valiente *et al.* 2009a) and its population allelic frequency to be correlated with temperature (Ramirez-Valiente *et al.* 2010b), revealing that temperature is an important selective agent in cork oak. From an evolutionary perspective, other than these studies, cork oak has only been explored in phylogenetic and phylogeographic studies (Magri *et al.* 2007, Simeone *et al.* 2009, Denk and Grimm 2010, Costa *et al.* 2011).

In spite of all these studies, little is still known about adaptive genetic variation in this species. However, this information is extremely important to understand the adaptation of cork oak, especially in the current context of climate change, which is expected to have a great impact in the Mediterranean Basin (IPCC 2007). These changes are predicted to occur rapidly, in a way that forest climate zone boundaries may move faster than forest tree species are able to migrate (Higgins and Harte 2006). Thus, the survival of cork oak will depend primarily on its plasticity and ability to adapt under new environmental conditions (Davis and Shaw 2001, Valladares *et al.* 2007). The scenario is more worrying now that cork oak *montados* are facing a great decline by the lack of regeneration, mainly due to severe drought periods and the dependence on aged adult trees (Toumi and Lumaret 1998, Soto *et al.* 2007) because of bad management, which puts this entire ecosystem under threat (Quartau and Mathias 2010). Several diseases have also been highlighted as important reasons for cork oak decline (Brasier 1996, Cabral and Ferreira 1999). Therefore, assessing adaptive genetic variation becomes urgent to understand cork oak adaptation to biotic and abiotic conditions in order to delineate sustainable conservation strategies for this species.

## 1.2. Natural selection and adaptation

Natural selection was initially proposed in *On The Origin of Species* (Darwin 1859) and is currently considered a crucial process in species evolution. Natural selection is the mechanism that explains the evolution of adaptations, which are traits that increase the survival or reproduction success (fitness) of organisms. It acts upon every trait that affects the fitness of a biological entity. Adaptation also refers to the process through which the organisms with an advantageous trait will increase in number by the action of natural selection, replacing the less suitable organisms, which survive and reproduce in a smaller degree (Futuyma 2005).

Two main types of selection are generally defined (Hughes 2007): positive selection, which acts upon advantageous mutations, and negative selection, through which a mutation that reduces fitness is selected against and eliminated. Positive selection has been the focus of many studies (e.g. Lexer *et al.* 2004, Chen *et al.* 2010, Bernhardsson and Ingvarsson 2012), arousing a special interest, as it is associated with adaptation and evolution of new forms and functions. It can be divided into directional selection, which leads to the fixation of an advantageous mutation, and balancing selection, which maintains polymorphism (Hughes 2007). Negative selection and directional selection are generally associated with loss of variation within a population, although directional selection can also increase the variation between populations or species. On the other hand, balancing selection can maintain polymorphism within populations or species in several ways (Nielsen 2005). The most documented type of balancing selection is overdominance or heterozygous advantage, where the heterozygous genotype has a highest fitness relatively to the homozygous, maintaining both alleles across generations (e.g. Banaszek *et al.* 2009, Briggs *et al.* 2011). Balancing selection can also occur for several other reasons, such as fluctuating environment conditions that may favour different genotypes in different generations or different microhabitats/niches, and inverse frequency-dependent selection, in which the rarer a phenotype is, the greater its fitness (Futuyma 2005). According to the neutral theory of molecular evolution (Kimura 1983) and the nearly-neutral variant (Ohta 1973) most of the functional genes are under purifying selection, a type of selection that eliminates new variants, and positive selection would be rare, although its importance to adaptation is not denied by these theories.

Different types of selection leave different molecular signatures, making it possible to detect past and present selection events through population genetics and genomics studies (Jensen *et al.* 2007, Ellegren 2008, Strasburg *et al.* 2012).

### 1.2.1. Detecting signatures of natural selection

A number of statistical methods have been developed to detect molecular signatures of positive selection. In general, these statistical tests can use polymorphism data, divergence data or a combination of both (Nielsen 2005, Jensen *et al.* 2007).

Polymorphism based methods involve sampling multiple copies of a genomic region within populations, often with an orthologous copy from a closely related species to define the ancestral and derived variation states. The site frequency spectrum is assessed across the sampled populations to identify patterns of positive selection (Jensen *et al.* 2007) and many of the classic neutrality tests are based on this. In Tajima's *D*, one of the most commonly used test, the average number of nucleotide differences between pairs of sequences is compared with the total number of segregating sites (Tajima 1989). If differences between these two measures are larger than expected on the standard neutral model, this is rejected. Under the neutrality model, *D* is expected to be zero. Negative values indicate excess of rare haplotypes and therefore positive selection or population growth, while positive values indicate an excess of intermediate frequency haplotypes, i.e. balancing selection or population decline. Other commonly used site frequency spectrum test is Fu's *Fs*. In this test the probability of observing a random neutral sample with a number of alleles similar or smaller than the observed value given the observed number of pairwise differences is estimated (Fu 1997). A negative *Fs* will indicate an excess of rare alleles and thus genetic hitchhiking or population growth, while a positive *Fs* will indicate a deficit of allele number and therefore balancing selection or a bottleneck. Other site frequency based tests are generally used, such as Fu & Li (Fu and Li 1993) and Fay & Wu (Fay and Wu 2000). Despite these site frequency spectrum statistical tests being popular in selection studies, they are affected by demography and fail to distinguish between selection and demographic effects (Nielsen 2005).

Divergence based methods use comparative approaches, involving sequence data from multiple species to detect and locate positive selection (Nielsen 2005, Jensen *et al.* 2007). For coding sequences, comparisons between the rate of nonsynonymous mutations per nonsynonymous sites ($d_N$) and the rate of synonymous mutations per synonymous sites ($d_S$) can provide a great measure of the strength and character of selection. In a neutral scenario, where there are no functional constrains, synonymous and nonsynonymous substitutions would occur at the same rate and thus $d_N/d_S$ ($\omega$) =1. If there are functional constrains and therefore negative selection, we should expect $\omega<1$. A $\omega>1$ is interpreted as evidence for positive selection on amino acid changes (Jensen *et al.* 2007). As most proteins are expected to be under strong purifying selection to preserve their structure and function, the examination of $\omega$ over an entire gene is unlikely to detect positive selection acting on a few sites, since the amount of positive selection needed to elevate $\omega$ above 1 is enormous (Nielsen and Yang 1998, Suzuki and Gojobori 1999). In this way, several methods for estimating the distribution of $\omega$ in the presence of site-to-site variation have been proposed (e.g. Suzuki *et al.* 2001, Yang *et al.* 2005, Wong *et al.* 2006). One of the most commonly used methods is the maximum likelihood approach implemented in Phylogenetic Analysis by Maximum Likelihood (PAML) (Yang 1997, 2007). In maximum likelihood methods, two distinct inferential steps are required to identify sites subjected to positive selection. In the first step, it must be shown that a given alignment contains any sites likely to be under positive selection by means of a model comparison. If a model that includes selection performs better than one that does not allow it (the null model), the protein is considered as being under positive selection. The null model allows different sites to have different values of $\omega$, but not values of $\omega>1$. The alternative model adds a class of sites with $\omega>1$, allowing for positive selection. The two models are compared by a likelihood ratio test and if the alternative model is better suited for the data than the null model, positive selection can be inferred. In the second inferential step, sites under positive selection are detected. For each site, the posterior probability is calculated under each $\omega$ class in the maximum likelihood model. A high posterior probability under the $\omega>1$ class is suggestive of positive selection at that site (Yang and Bielawski 2000, Yang *et al.* 2005, Yang 2007). Several studies have used PAML to detect and locate sites under positive selection (e.g. Padhi and Verghese 2008, Talianova *et al.* 2011).

Natural selection can be reflected in statistical associations between genetic markers and environmental data (Joost *et al.* 2007, Coop *et al.* 2010, Manel *et al.* 2010). This approach has the advantage of allowing the identification of loci under selection and the establishment of hypotheses about the ecological factors that may be driving the species adaptation (Joost *et al.* 2007). A recently developed software, SAM (Joost *et al.* 2007, 2008), has been widely used to perform these associations (e.g. Grivet *et al.* 2010, Nunes *et al.* 2011, Gao *et al.* 2012). The method used in this program is based on multiple univariate logistic regression models to test for association between allelic frequencies at marker loci and environmental variables (Joost *et al.* 2008).

## 1.3. Molecular markers in plant evolution research

Molecular analyses in plants have been traditionally based on chloroplastidial DNA (cpDNA) sequences, used mainly for phylogenetic studies (Borsch and Quandt 2009). However, cpDNA has low evolutionary rate and is exclusively maternally transmitted in Fagaceae, which limits the information that can be withdrawn from this type of markers (Mogensen 1996, Borsch and Quandt 2009). On the other hand, nuclear markers, such as microsatellites (or simple sequence repeats, SSRs), ITS (internal transcribed spacer regions of nuclear ribosomal DNA) or AFLPs (amplified fragment length polymorphisms), are biparentally inherited thus overcoming this limitation of cpDNA. Analysis of ITS sequences have been proven useful in phylogenetic and phylogeography studies (e.g. Manos *et al.* 1999, Bellarosa *et al.* 2005, Simeone *et al.* 2009), although the existence of paralogue genes may distort the phylogenetic signals in plant evolutionary studies (Song *et al.* 2012). On the other hand, SSRs have high mutation rates and therefore high polymorphism, being highly suitable for answering population-level questions. However, high mutation rates lead to homoplasy, which could limit the biological accuracy of the results, and SSRs have null alleles that bias data analysis, in addition to being relatively scarce throughout the genome (Hedrick 1999). AFLPs have the advantage of being distributed throughout the whole genome (Arif *et al.* 2010), giving information at the genomic level. Furthermore, AFLPs can be analysed to find outlier markers possibly under selection (e.g. Manel *et al.* 2009, Poncet *et al.* 2010), but have the disadvantage of being anonymous and presenting reproducibility problems (Arif *et al.* 2010).

Single Nucleotide Polymorphisms (SNPs) are the most recent and promising type of molecular markers whose application in plant genetic studies is growing (e.g. Foster *et al.* 2010, Kelleher *et al.* 2012). They are abundant and widespread in the genomes of many species and can be used as neutral markers or as adaptive markers if they are located in adaptive genes or adaptive regulating regions (Kirk and Freeland 2011).

### 1.3.1. Single nucleotide polymorphisms

SNPs are single base differences between DNA sequences and are the most common molecular markers found in eukaryotic genomes (Coles *et al.* 2005, Westermeier *et al.* 2009). They are widespread through the genomes and can be found both in coding and non-coding regions. There are three different SNP types: transitions (purine to purine or pyrimidine to pyrimidine), transversions (purine to pyrimidine or pyrimidine to purine) and small insertions/deletions (indels). SNPs can be bi-, tri- or tetra-allelic, but they are usually biallelic (Doveri *et al.* 2008), which can be a disadvantage when compared to multiallelic markers. The great abundance of SNPs throughout the genome compensates for the little information given by each SNP as they are able to provide a high density of markers near or in a locus of interest (Doveri *et al.* 2008, Duran *et al.* 2009). SNPs have low mutation rates, which makes them very useful for studying complex genetic traits and for understanding genome evolution (Syvanen 2001). When in coding or regulatory regions, these variations can have a major impact in the development of an organism and its response to the environment, being these particular SNPs of great interest when trying to understand organisms' adaptation.

SNP markers have become popular in ecology and evolution research (Moen *et al.* 2008, Namroud *et al.* 2008, Hao *et al.* 2011, Keller *et al.* 2012), especially in non-model organisms. These markers can have several applications, such as genome mapping, association studies, assessment of genetic diversity, paternity studies, or phylogeography (Duran *et al.* 2009).With genome sequencing becoming more and more affordable, the development of SNP markers have become easier, especially with the emergence of next generation sequencing (NGS) (e.g. Novaes *et al.* 2008, Milano *et al.* 2011, Bundock *et al.* 2012). SNPs in coding regions can be particularly useful, especially in population genetics, functional genomics, conservation and evolutionary

biology studies, as their effects on the protein amino acid composition and function can be more easily assessed (Renaut *et al.* 2010).

### 1.3.2. NGS: an important tool for molecular markers development

Next-generation sequencing (NGS) methods have emerged during the last decade and are evolving rapidly. They have been having a great impact on genomics research, especially in non-model organisms, allowing for the performance of experiments that were previously not feasible or economically challenging. NGS enables whole-genome sequencing, targeted re-sequencing, metagenomics studies, transcriptome sequencing, and mutation detection, among other applications (Voelkerding *et al.* 2009, Glenn 2011). It produces large amounts of sequence data and provides very important tools for molecular marker discovery (e.g. SNPs, SSRs) in non-model organisms for which no genomic tools are available (e.g. Barbazuk *et al.* 2007, Novaes *et al.* 2008, Bundock *et al.* 2012). Several NGS platforms are currently available, such as 454 (Roche), Illumina, SOLiD (Applied Biosystems), Ion Torrent and Helicos (BioSciences), among others, each technology having advantages and drawbacks (Glenn 2011). One of the most commonly used technology is 454.

#### 1.3.2.1. 454 pyrosequencing and SNP discovery

When using 454, sequencing is performed by the pyrosequencing technology. In this method, each incorporation of a nucleotide by the DNA polymerase results in the release of pyrophosphate, which initiates a series of reactions that ultimately produce light by the firefly enzyme luciferase. The amount of light produced is proportional to the number of nucleotides incorporated (Mardis 2008, Shendure and Ji 2008, Voelkerding *et al.* 2009).

454 sequencing has several advantages in relation to other NGS methods, including low cost and the production of longer reads (800bp in average) (Shendure and Ji 2008, Glenn 2011), although the number of reads obtained is relatively small. Longer reads can be assembled more efficiently, which is essential when characterizing a genome or

transcriptome without a reference. A further advantage is the relatively easy processing and analysis of the output data, which does not require great computational resources. Moreover, 454 was the first NGS method commercially available and it is therefore greatly used, so that the analysis software is well developed (Glenn 2011).

Some drawbacks are known for the 454 technology, namely the reported error rates of 1%, which are higher than those produced by the traditional Sanger sequencing (Margulies *et al.* 2005, Huse *et al.* 2007). A major limitation of the 454 technology relates to homopolymers (sequences formed by the repetition of a single type of nucleotide), as these are inferred from signal intensity, which is often difficult to resolve. This can result in ambiguity of homopolymer length, particularly for longer homopolymers. In addition, insufficient flushing between flows can cause single base insertions (carry forward events) usually near, but not adjacent to homopolymers. As a consequence, insertions and deletions are the major errors in 454 sequencing (Huse *et al.* 2007, Shendure and Ji 2008).

The 454 platform has been primarily used for transcriptome characterization, target re-sequencing and *de novo* genome sequencing, among other less frequent applications (Glenn 2011). Transcriptome sequencing, when performed to provide sufficient coverage depth, is a valuable resource for the identification of Single Nucleotide Polymorphisms (SNPs) in transcribed regions of the genome (Barbazuk *et al.* 2007, Novaes *et al.* 2008, Lai *et al.* 2012), which are of major interest when trying to understand how selection acts on functional genes (Vera *et al.* 2008, Renaut *et al.* 2010, Horton *et al.* 2012). Polymorphisms within a gene may have different impacts on its function or expression depending on their genomic location (intron, exon or unstranslated region). Mutations within coding regions are particularly insightful as they can affect the amino acid composition of the protein and therefore its structure and/or function. Therefore, the discovery of SNP markers in transcriptome sequences can facilitate the identification of genes involved in adaptive change (Renaut *et al.* 2010, Bajgain *et al.* 2011).

## 1.4. Thesis aims

In a previous study (unpublished data) developed in the scope of the SOBREIRO/0036/2009 project (financed by Fundação para a Ciência e Tecnologia – FCT), integrated in the Cork Oak ESTs (expressed sequence tags) Consortium (COEC), large-scale 454 trancriptome sequencing was carried out. This consortium had the objective of sequencing the cork oak transcriptome for several tissues, developmental stages and biotic and abiotic stress conditions. In the referred project, the target tissue was leaves and transcriptome sequencing was performed using a pool of individuals from eigth different populations spread throughout cork oak distribution range. The results were then surveyed to identify SNPs with 4pipe4 software (Pina-Martins *et al.*, submited). More than 400 putative SNPs were discovered in transcribed regions, constituting a resource of major interest as they can provide the ground for understanding the selective forces acting on functional genes.

The central purpose of the present study, developed in the scope of the FCT funded Project PTDC/AGR-GPL/104966/2008, was to investigate cork oak adaptive genetic variation through the exploitation of the putative SNPs discovered in this species transcriptome. However, as 454 reads are prone to sequence errors (Margulies *et al.* 2005, Huse *et al.* 2007) and SNP validation rates reported reach only 85% (Barbazuk *et al.* 2007), it was first necessary to experimentally validate them.

Thus, the main goals of this work were the following:

1. To validate by Sanger sequencing 10 to 15 putative SNPs discovered through 454 transcriptome pyrosequencing, in order to determine if they represent true biological variation or result from technical errors.
2. To explore 5 of the putative validated SNPs located in functional genes through a population genetic diversity analysis in order to assess signatures of selection, population structure and patterns of adaptive genetic variation in cork oak.
3. To test for association between the genetic variation found in the selected genes and environmental variables potentially relevant for cork oak local adaptation. Combining this approach with neutrality tests will increase the possibility of finding genetic markers under selection and assessing different aspects of the evolutionary processes acting on cork oak.

## 1.5. References


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

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 Physiology* **25**, 1085-1090.

Arif IA, Bakir MA, Khan HA, Al Farhan AH, Al Homaidan AA, Bahkali AH, Al Sadoon M, Shobrak M (2010) A brief review of molecular techniques to assess plant diversity. *International Journal of Molecular Sciences* **11**, 2079-2096.

Bajgain P, Richardson BA, Price JC, Cronn RC, Udall JA (2011) Transcriptome characterization and polymorphism detection between subspecies of big sagebrush (*Artemisia tridentata*). *BMC Genomics* **12**, 15.

Banaszek A, Taylor JRE, Ochocinska D, Chetnicki W (2009) Robertsonian polymorphism in the common shrew (*Sorex araneus* L.) and selective advantage of heterozygotes indicated by their higher maximum metabolic rates. *Heredity* **102**, 155-162.

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. *Plant Journal* **51**, 910-918.

Bellarosa R, Simeone MC, Papini A, Schirone B (2005) Utility of ITS sequence data for phylogenetic reconstruction of Italian *Quercus* spp. *Molecular Phylogenetics and Evolution* **34**, 355-370.

Bernhardsson C, Ingvarsson PK (2012) Geographical structure and adaptive population differentiation in herbivore defence genes in European aspen (*Populus tremula* L., Salicaceae). *Molecular Ecology* **21**, 2197-2207.

Borsch T, Quandt D (2009) Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. *Plant Systematics and Evolution* **282**, 169-199.

Brasier CM (1996) *Phytophthora cinnamomi* and oak decline in southern Europe. Environmental constraints including climate change. *Annales Des Sciences Forestieres* **53**, 347-358.

Briggs CW, Collopy MW, Woodbridge B (2011) Plumage polymorphism and fitness in Swainson's hawks. *Journal of Evolutionary Biology* **24**, 2258-2268.

Bundock PC, Casu RE, Henry RJ (2012) Enrichment of genomic DNA for polymorphism detection in a non-model highly polyploid crop plant. *Plant Biotechnology Journal* **10**, 657-667.

Cabral MT, Ferreira MC (1999) *Pragas dos Montados*. Estação Florestal Nacional, Lisbon, Portugal.

Chen QH, Han ZX, Jiang HY, Tian DC, Yang SH (2010) Strong positive selection drives rapid diversification of R-genes in *Arabidopsis* relatives. *Journal of Molecular Evolution* **70**, 137-148.

Coles ND, Coleman CE, Christensen SA, Jellen EN, Stevens MR, Bonifacio A, Rojas-Beltran JA, Fairbanks DJ, Maughan PJ (2005) Development and use of an expressed sequenced tag library in quinoa (*Chenopodium quinoa* Willd.) for the discovery of single nucleotide polymorphisms. *Plant Science* **168**, 439-447.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185**, 1411-1423.

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 Proceedings* **5**, (Suppl 7), 13.

Darwin CR (1859) *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*, 1st ed. John Murray, London.

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

Deng M (2007) *Anatomy, taxonomy, distribution, and phylogeny of Quercus subgenus Cyclobalanopsis (Oersted) Schneid. (Fagaceae).* PhD thesis, Chinese Academy of Sciences.

Denk T, Grimm GW (2010) The oaks of western Eurasia: Traditional classifications and evidence from two nuclear markers. *Taxon* **59**, 351-366.

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 Phytologist* **170**, 723-738.

Doveri S, Lee D, Maheswaran M, Powell W (2008) Molecular markers: History, features and applications. In: *Principles and Practices of Plant Genomics* (eds. Kole C, Abbott AG), pp. 23-68. Science Publishers, Enfield, USA.

Duran C, Appleby N, Edwards D, Batley J (2009) Molecular genetic markers: discovery, applications, data storage and visualisation. *Current Bioinformatics* **4**, 16-27.

Ellegren H (2008) Comparative genomics and the study of evolution by natural selection. *Molecular Ecology* **17**, 4586-4596.

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405-1413.

Foster JT, Allan GJ, Chan AP, Rabinowicz PD, Ravel J, Jackson PJ, Keim P (2010) Single nucleotide polymorphisms for assessing genetic diversity in castor bean (*Ricinus communis*). *BMC Plant Biology* **10**, 13.

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

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* **133**, 693-709.

Futuyma DJ (2005) *Evolution*. Sinauer Associates, Sunderland, USA.

Gailing O, Vornam B, Leinemann L, Finkeldey R (2009) Genetic and genomic approaches to assess adaptive genetic variation in plants: forest trees as a model. *Physiologia Plantarum* **137**, 509-519.

Gandour M, Khouja ML, Toumi L, Triki S (2007) Morphological evaluation of cork oak (*Quercus suber*): Mediterranean provenance variability in Tunisia. *Annals of Forest Science* **64**, 549-555.

Gao LX, Tang SQ, Zhuge LQ, Nie M, Zhu Z, Li B, Yang J (2012) Spatial genetic structure in natural populations of *Phragmites australis* in a mosaic of saline habitats in the yellow river delta, China. *PLoS One* **7**, 8.

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

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11**, 759-769.

Gonzalez-Martinez SC, Krutovsky KV, Neale DB (2006) Forest-tree population genomics and adaptive evolution. *New Phytologist* **170**, 227-238.

Grandtner MM (2005) *Elsevier's Dictionary of Trees*, 1st ed. Elsevier, Amsterdam, The Nederlands.

Grivet D, Sebastiani F, Alia R, Bataillon T, Torre S, Zabal-Aguirre M, Vendramin GG, Gonzalez-Martinez SC (2010) Molecular footprints of local adaptation in two mediterranean conifers. *Molecular biology and evolution* **28**, 101-116.

Hao Z, Li X, Xie C, Weng J, Li M, Zhang D, Liang X, Liu L, Liu S, Zhang S (2011) Identification of functional genetic variations underlying drought tolerance in maize using SNP markers. *Journal of Integrative Plant Biology* **53**, 641-652.

Hedrick PW (1999) Perspective: Highly variable loci and their interpretation in evolution and conservation. *Evolution* **53**, 313-318.

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

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. *Nature Genetics* **44**, 212-216.

Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* **99**, 364-373.

Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**, 143.

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, Switzerland.

Jensen JD, Wong A, Aquadro CF (2007) Approaches for identifying targets of positive selection. *Trends in Genetics* **23**, 568-577.

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. *Molecular Ecology* **16**, 3955-3969.

Joost S, Kalbermatten M, Bonin A (2008) Spatial analysis method (SAM): a software tool combining molecular and environmental data to identify candidate loci for selection. *Molecular Ecology Resources* **8**, 957-960.

Kelleher CT, Wilkin J, Zhuang J, Cortes AJ, Quintero ALP, Gallagher TF, Bohlmann J, Douglas CJ, Ellis BE, Ritland K (2012) SNP discovery, gene diversity, and

linkage disequilibrium in wild populations of *Populus tremuloides*. *Tree Genetics & Genomes* **8**, 821-829.

Keller SR, Levsen N, Olson MS, Tiffin P (2012) Local adaptation in the flowering-time gene network of balsam poplar, *Populus balsamifera* L. *Molecular biology and evolution* **29**, 3143-3152.

Kimura M (1983) *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, UK.

Kirk H, Freeland JR (2011) Applications and implications of neutral versus non-neutral markers in molecular ecology. *International Journal of Molecular Sciences* **12**, 3966-3988.

Krutovsky KV, Neale DB (2005) Forest genomics and new molecular genetic approaches to measuring and conserving adaptive genetic diversity in forest trees. In: *Conservation and Management of Forest Genetic Resources in Europe* (eds. Geburek T, Turok J), pp. 369–390. Arbora Publishers, Zvolen.

Lai K, Duran C, Berkman PJ, Lorenc MT, Stiller J, Manoli S, Hayden MJ, Forrest KL, Fleury D, Baumann U, Zander M, Mason AS, Batley J, Edwards D (2012) Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnology Journal* **10**, 743-749.

Lexer C, Heinze B, Alia R, Rieseberg LH (2004) Hybrid zones as a tool for identifying adaptive genetic variation in outbreeding forest trees: lessons from wild annual sunflowers (*Helianthus* spp.). *Forest Ecology and Management* **197**, 49-64.

Linhart YB (1999) Variation in woody plants: molecular markers, evolutionary processes and conservation biology. In: *Molecular Biology of Woody Plants* (eds. Jain SM, Minocha SC), pp. 341-375. Kluwer Academic Publishers, Dordrecht, The Nederlands.

Logan WB (2005) *Oak: The frame of civilization*, 1st ed. W. W. Norton & Company, New York, USA.

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

Manel S, Conord C, Despres L (2009) Genome scan to assess the respective role of host-plant and environmental constraints on the adaptation of a widespread insect. *BMC Evolutionary Biology* **9**, 288.

Manel S, Poncet BN, Legendre P, Gugerli F, Holderegger R (2010) Common factors drive adaptive genetic variation at different spatial scales in *Arabis alpina*. *Molecular Ecology* **19**, 3824-3835.

Manos PS, Doyle JJ, Nixon KC (1999) Phylogeny, biogeography, and processes of molecular differentiation in *Quercus* subgenus *Quercus* (Fagaceae). *Molecular Phylogenetics and Evolution* **12**, 333-349.

Mardis ER (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9**, 387-402.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu PG, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

Milano I, Babbucci M, Panitz F, Ogden R, Nielsen RO, Taylor MI, Helyar SJ, Carvalho GR, Espineira M, Atanassova M, Tinti F, Maes GE, Patarnello T, Bargelloni L (2011) Novel tools for conservation genomics: comparing two high-throughput approaches for SNP discovery in the transcriptome of the european hake. *PLoS One* **6**, e28008.

Moen T, Hayes B, Nilsen F, Delghandi M, Fjalestad KT, Fevolden S-E, Berg PR, Lien S (2008) Identification and characterisation of novel SNP markers in Atlantic cod: Evidence for directional selection. *BMC Genetics* **9**, 18.

Mogensen HL (1996) The hows and whys of cytoplasmic inheritance in seed plants. *American Journal of Botany* **83**, 383-404.

Namroud M-C, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology* **17**, 3599-3613.

Neale DB (2007) Genomics to tree breeding and forest health. *Current Opinion in Genetics & Development* **17**, 539-544.

Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. *Current Opinion in Plant Biology* **11**, 149-155.

Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nature Reviews Genetics* **12**, 111-122.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics* **39**, 197-218.

Nielsen R, Yang ZH (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929-936.

Novaes E, Drost DR, Farmerie WG, Pappas GJ, Jr., Grattapaglia D, Sederoff RR, Kirst M (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**, 312.

Nunes VL, Beaumont MA, Butlin RK, Paulo OS (2011) Multiple approaches to detect outliers in a genome scan for selection in ocellated lizards (*Lacerta lepida*) along an environmental gradient. *Molecular Ecology* **20**, 193-205.

Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96-98.

Oldfield S, Lusty C, MacKinven A (1998) *The world list of threatened trees*. World Conservation Press, Cambrige, UK.

Padhi A, Verghese B (2008) Detecting molecular adaptation at individual codons in the pattern recognition protein, lipopolysaccharicle- and beta-1,3-glucan-binding protein of decapods. *Fish & Shellfish Immunology* **24**, 638-648.

Pausas JG, Pereira JS, Aronson J (2009) The tree. In: *Cork Oak Woodlands on the Edge* (eds. Aronson J, Pereira JS, Pausas JG), pp. 11-21. Island Press, Washington DC, USA.

Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. *Annual Review of Ecology Evolution and Systematics* **37**, 187-214.

Pina-Martins F, Vieira BM, Seabra SG, Batista D, Paulo OS. 4Pipe4 – A 454 data analysis pipeline specialized for SNP detection. *Submited*.

Poncet BN, Herrmann D, Gugerli F, Taberlet P, Holderegger R, Gielly L, Rioux D, Thuiller W, Aubert S, Manel S (2010) Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Molecular Ecology* **19**, 2896-2907.

Quartau JA, Mathias ML (2010) Insects of the understorey in Western Mediterranean forest landscapes: a rich biodiversity under threat. In: *Insect Habitats:*

*Characteristics, Diversity and Management* (eds. Harris EL, Davies NE), pp. 133-142. Nova Science Publishers, Hauppauge NY, USA.

Ramirez-Valiente AJ, Sanchez-Gomez D, Aranda I, Valladares F (2010a) Phenotypic plasticity and local adaptation in leaf ecophysiological traits of 13 contrasting cork oak populations under different water availabilities. *Tree Physiology* **30**, 618-627.

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 Genetics & Genomes* **7**, 285-295.

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. *Molecular Ecology* **18**, 3803-3815.

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

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.). *Forest Ecology and Management* **257**, 1676-1683.

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

Roy J, Saugier B, Mooney HA (2001) *Terrestrial Global Productivity*. Academic Press, London, UK.

Serrasolses I, Pérez-Devesa M, Vilagrosa A, Pausas JG, Sauras T, Cortina J, Vallejo VR (2009) Soil properties constraining cork oak distribution. In: *Cork Oak Woodlands on the Edge* (eds. Aronson J, Pereira JS, Pausas JG), pp. 89-99. Island Press, Washington DC, USA.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**, 1135-1145.

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**, 236-252.

Singh G (2004) *Plant systematics: an integrated approach*. Science Publishers, Enfield, USA.

Soler M, Serra O, Molinas M, Huguet G, Fluch S, Figueras M (2007) A genomic approach to suberin biosynthesis and cork differentiation. *Plant Physiology* **144**, 419-431.

Song H-X, Gao S-P, Jiang M-Y, Liu G-L, Yu X-F, Chen Q-B (2012) The evolution and utility of ribosomal ITS sequences in Bambusinae and related species: divergence, pseudogenes, and implications for phylogeny. *Journal of genetics* **91**, 129-139.

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**, 601-607.

Strasburg JL, Sherman NA, Wright KM, Moyle LC, Willis JH, Rieseberg LH (2012) What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philosophical Transactions of the Royal Society B-Biological Sciences* **367**, 364-373.

Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. *Molecular biology and evolution* **16**, 1315-1328.

Suzuki Y, Gojobori T, Nei M (2001) ADAPTSITE: detecting natural selection at single amino acid sites. *Bioinformatics* **17**, 660-661.

Syvanen AC (2001) Accessing genetic variation: Genotyping single nucleotide polymorphisms. *Nature Reviews Genetics* **2**, 930-942.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595.

Talianova M, Vyskot B, Janousek B (2011) Interkingdom protein domain fusion: the case of an antimicrobial protein in potato (*Solanum tuberosum*). *Plant Systematics and Evolution* **297**, 129-139.

Toumi L, Lumaret R (1998) Allozyme variation in cork oak (*Quercus suber* L.): the role of phylogeography and genetic introgression by other Mediterranean oak species and human activities. *Theoretical and Applied Genetics* **97**, 647-656.

Ueno S, Le Provost G, Leger V, Klopp C, Noirot C, Frigerio J-M, Salin F, Salse J, Abrouk M, Murat F, Brendel O, Derory J, Abadie P, Leger P, Cabane C, Barre A, de Daruvar A, Couloux A, Wincker P, Reviron M-P, Kremer A, Plomion C

(2010) Bioinformatic analysis of ESTs collected by Sanger and pyrosequencing methods for a keystone forest tree species: oak. *BMC Genomics* **11**, 650.

Valladares F, Gianoli E, Gomez JM (2007) Ecological limits to plant phenotypic plasticity. *New Phytologist* **176**, 749-763.

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. *Molecular Ecology* **17**, 1636-1647.

Voelkerding KV, Dames SA, Durtschi JD (2009) Next-Generation Sequencing: from basic research to diagnostics. *Clinical Chemistry* **55**, 641-658.

Westermeier P, Wenzel G, Mohler V (2009) Development and evaluation of single-nucleotide polymorphism markers in allotetraploid rapeseed (*Brassica napus* L.). *Theoretical and Applied Genetics* **119**, 1301-1311.

Wong WSW, Sainudiin R, Nielsen R (2006) Identification of physicochemical selective pressure on protein encoding nucleotide sequences. *BMC Bioinformatics* **7**, 148.

Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**, 1586-1591.

Yang ZH (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**, 555-556.

Yang ZH, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution* **15**, 496-503.

Yang ZH, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular biology and evolution* **22**, 1107-1118.

# CHAPTER 2

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

Inês S. Modesto[1,2], Célia Miguel[1,3], Francisco Pina-Martins [2,4], Maria Glushkova[5], Manuela Veloso[6], Octávio S. Paulo [2]; Dora Batista [2,7]

[1] Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa (ITQB-UNL), Av. da República, 2780-157 Oeiras, Portugal
[2] Computational Biology and Population Genomics Group, Centro de Biologia Ambiental (CBA), Faculdade de Ciências da Universidade de Lisboa, 1759-016 Lisboa, Portugal
[3] Instituto de Biologia Experimental e Tecnológica (IBET), Apartado 12, 2781-901 Oeiras, Portugal
[4] Centro de Estudos do Ambiente e do Mar (CESAM) e Departamento de Biologia, Universidade de Aveiro, 3810-193 Aveiro, Portugal
[5] Department of Forest genetics, physiology and plantations, Forest Research Institute of B.A.S., 132 "St. Kliment Ohridski" blvd., 1756 Sofia, Bulgaria
[6] Instituto Nacional de Investigação Agrária e Veterinária, I.P. (INIAV), L-INIA, Pólo de Oeiras, Unidade Recursos Genéticos, Ecofisiologia e Melhoramento de Plantas, Quinta do Marquês, 2784-505 Oeiras, Portugal
[7] Centro de Investigação das Ferrugens do Cafeeiro-Biotrop/Instituto de Investigação Científica Tropical (CIFC-Biotrop/IICT), Quinta do Marquês, 2784-505 Oeiras, Portugal

## 2.1. Abstract

Cork oak (*Quercus suber* L.) is an evergreen tree species of great ecological, economical and social relevance throughout its distribution range, the Western Mediterranean region. In Portugal, where the largest stands of cork oak are located, this tree is particularly important. Cork oak stands have been facing a significant decline by the lack of regeneration, mostly due to severe drought periods, the dependence on aged adult trees and bad management, as well as susceptibility to several diseases. Drought periods are predicted to become even more severe during this century as a result of the climate changes expected to occur within cork oak distribution range. The ability of this species to deal with such severe climate changes is likely to depend mainly on its plasticity and adaptive potential. In this scenario, the assessment of adaptive genetic variation is essential to understand how cork oak may cope with these threats and to delineate management strategies of its genetic resources. In this work the validation of putative SNPs detected in cork oak transcriptome was performed in order to develop useful variable markers in functional genes, potentially under selection. Five of the validated SNPs were further investigated through a population genetics approach. Several neutrality tests were performed as well as environmental association tests aiming at finding selection signatures. Two gene fragments seemed to be under balancing selection, namely the putative orthologs of *Arabidopsis thaliana* (L.) Heynh. *NPR1*, involved in plant defence response against pathogens, and *ARF16*, a gene implicated in root cap cell differentiation and previously identified as a candidate gene for drought resistance. In another gene fragment, a putative ortholog of an *A. thaliana HSP*, involved in stress response, one amino acid position was found to be possibly under positive selection. Moreover, allele frequency in this same position was associated with latitude and with precipitation in September, revealing its potential relevance in adaptation to local climatic conditions. Therefore, in this study, useful molecular markers for assessing cork oak adaptive genetic variation were developed and important steps were taken in obtaining information about cork oak adaptation to biotic and abiotic environmental conditions.

**Keywords:** adaptation, balancing selection, environmental association, positive selection

## 2.2. Introduction

Cork oak (*Quercus suber* L.) is a long-lived evergreen tree species endemic to the Western Mediterranean region. It is outcrossing and wind-pollinated and its seed dissemination occurs essentially through gravity and zoochory (Gil and Varela 2008). Cork oak occurs in a vast variety of climatic conditions and its distribution is rather discontinuous, ranging from the Atlantic coast of North Africa and Iberian Peninsula to Southeastern Italy (Figure 2.1) (Pausas *et al.* 2009). In addition to this natural distribution, cork oak can also be found as an introduced species, somewhat naturalized, in Bulgaria (Figure 2.1, dashed grey) (Alexandrov *et al.* 2001). Cork oak has a restricted distribution when compared to other oaks, mainly due to cork oak being highly constrained to acidophilus soils (Serrasolses *et al.* 2009) and by low winter temperatures (Pausas *et al.* 2009). Cork oak has been long explored for the extraction of its bark, the cork, which has a great economical and social importance in the countries where this tree is naturally distributed, with special impact in Portugal and Spain (Campos and Aronson 2009). In these two countries, cork oak constitutes a unique ecological system, known as *montado* in Portugal and *dehesa* in Spain, which contribute 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).

Climate changes are expected to be severe in the Mediterranean Basin, with an increase of at least 2-4$^{o}$C and a great decrease in precipitation during this century (IPCC 2007). Therefore, selective pressures exerted by climate within cork oak distribution range are expected to increase. If this change happens as fast as predicted, forest climate zone boundaries could move quicker than forest tree species are able to migrate (Higgins and Harte 2006), so 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 stands are already facing a significant decline by the lack of regeneration, mainly due to severe drought periods, the dependence on aged adult trees (Toumi and Lumaret 1998, Soto *et al.* 2007) and inadequate management practices (Quartau and Mathias 2010). Therefore, it is essential to conduct studies to understand the capacity of cork oak populations to cope with environmental changes, as these may aggravate the decline. Additionally, cork oak is also being threatened by several diseases, such as the deadly ink disease caused by *Phytophthora cinnamomi* (Brasier

1996). Thus, understanding cork oak adaptive capacity in response to pathogens would also be of major interest.

As cork oak grows under a wide range of rainfall and temperatures, different populations are expected to be under distinct selective pressures. Several studies reported differences in phenotypic and functional traits in trees from distinct populations in common garden experiences (e.g. Aranda *et al.* 2005, Gandour *et al.* 2007, Ramirez-Valiente *et al.* 2009a, 2010). In Aranda *et al.* (2005), populations were shown to have differential responses to low temperatures, with populations from cold regions displaying more tolerance to this type of stress. In another study, it was demonstrated that northern populations were maladapted to drought, while continental populations were intermediately adapted to dry conditions (Ramirez-Valiente *et al.* 2009a). Some of the ecophysiological traits correlated with environmental variables were shown to be heritable, supporting the idea that natural selection has led to local adaptations (Ramirez-Valiente *et al.* 2011). Furthermore, a microssatellite (*Qp*ZAG46) was found to be correlated with leaf size (Ramirez-Valiente *et al.* 2009a) and its population allelic frequency to be correlated with temperature (Ramirez-Valiente *et al.* 2010), revealing that temperature is an important selective agent in cork oak. Relatively to cork oak molecular defence responses to pathogens, very little is known. In Coelho *et al.* (2006) a cinnamyl alcohol dehydrogenase was associated with defensive response to infection by *P. cinnamomi*, as this gene was up-regulated in infected root seedlings. However, no other cork oak candidate genes have yet been studied. In this perspective, little is known about adaptive genetic variation of cork oak, despite the high relevance of this information to delineate maintenance and conservation strategies for this species.

The main focus of this study was to assess cork oak adaptive genetic variation. In a previous study (unpublished data), 454 trancriptome sequencing was performed, using RNA extracted from leaves from 8 different populations, and the results were surveyed to identify Single Nucleotide Polymorphisms (SNPs). More than 400 putative SNPs were discovered in transcribed regions, which are of major interest as they can be used to understand the selective forces acting on functional genes. Since 454 reads are described as containing a relatively high rate of sequence errors, detecting variation in sequences obtained by this method can lead to erroneous results (Margulies *et al.* 2005, Huse *et al.* 2007). Thus, validating this variation is a crucial step to further exploit such data (Barbazuk *et al.* 2007, Wiedmann *et al.* 2008). The aims of this study were

therefore: (i) to validate putative SNPs in order to develop useful variable markers potentially under selection and (ii) to perform population genetic studies with a set of selected SNP markers in order to detect selection signatures in putative functional genes. The ability to detect signals 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 combine several methods in selection and adaptation studies. In this study several neutrality tests were used, as well as environmental association analyses to detect genetic markers possibly under selective pressure.

## 2.3. Materials and Methods

### 2.3.1. Sampling and DNA extraction

Nineteen 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 de 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 2.1, Figure 2.1). Populations were selected considering both geographical distribution and environmental heterogeneity between populations, prioritizing populations that represent contrasting environments (Table 2.1). Three samples from holm oak (*Quercus rotundifolia* Lam.) were also sampled from the original populations, as well as a turkey oak (*Quercus cerris* L.) sample. The collected leaves were stored at –80°C until DNA extraction. Genomic DNA was extracted from the liquid nitrogen-grounded leaves using DNeasy Plant Mini Kit (Qiagen), according to the manufacturer’s protocol.

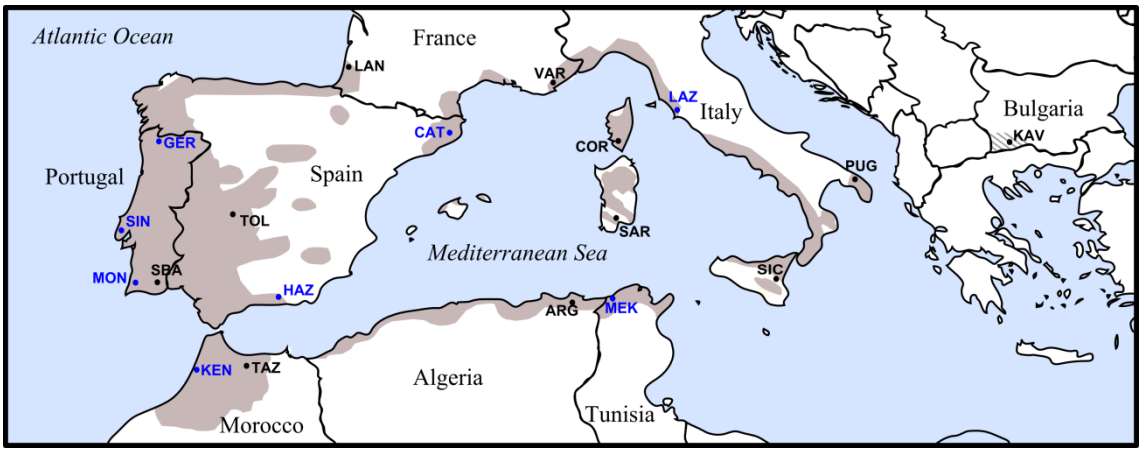



**Figure 2.1.** Cork oak (*Q. suber*) geographical distribution map. Full gray represents natural distribution; dashed gray represents introduced, somewhat naturalized populations. Sampling localities used for the validation process and gene variation assay are represented in blue; sampling localities used only for the gene variation assay are represented in black. GER, Gerês; SIN, Sintra; MON, Monchique; SBA, S. Bras de Alportel; HAZ, Haza de Lino; TOL, Montes de Toledo; CAT, Cataluña; KEN, Kenitra; TAZ, Taza; ARG, Guerbès; MEK, Mekna; LAN, Landes; VAR, Var; COR, Corse; LAZ, Lazio; SAR, Sardegna; SIC, Sicilia; PUG, Puglia; KAV, Kavrakirovo.

**Table 2.1.** Geographic location and climatic conditions of the cork oak populations used in this study.

| Code | Populations | Country | Spatial variables | | | Climatic variables | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Longitude | Latitude | Altitude (m) | AMT (ºC) | SMT (ºC) | WMT (ºC) | P (mm) | Psum (mm) | Ins |
| SIN | Sintra | Portugal | 9°25' W | 38°45' N | 528 | 14.95 | 24.77 | 6.70 | 795 | 45 | 2097 |
| MON | Monchique | Portugal | 8°34' W | 37°19' N | 902 | 16.81 | 29.00 | 4.83 | 950 | 55 | 1894 |
| GER | Gerês | Portugal | 8°10' W | 41°40' N | 381 | - | - | - | - | - | - |
| SBA | S. Bras de Alportel | Portugal | 7°56' W | 37°20' N | 485 | - | - | - | - | - | - |
| LAZ | Lazio, Toscany | Italy | 11°57' E | 42°25' N | 160 | 14.47 | 29.47 | 3.00 | 937 | 125 | - |
| PUG | Puglia, Brindisi | Italy | 17°40' E | 40°34' N | 45 | 16.55 | 29.10 | 7.90 | 588 | 49 | 2341 |
| SIC | Sicilia, Catania | Italy | 14°30' E | 37°07' N | 250 | 17.72 | 28.00 | 9.80 | 448 | 9 | 2586 |
| SAR | Sardegna, Cagliari | Italy | 8°51' E | 39°05' N | 200 | 16.97 | 30.80 | 6.60 | 883 | 29 | 2392 |
| VAR | Var, Bomes les Mimoses | France | 6°15' E | 43°08' N | 155 | 13.73 | 29.30 | 2.13 | 1029 | 111 | 2714 |
| LAN | Landes, Soustons | France | 1°20' W | 43°45' N | 20 | 12.27 | 23.73 | 5.60 | 870 | 140 | 821 |
| COR | Corse, Sartene | France | 8°58' E | 41°37' N | 50 | 13.91 | 27.77 | 5.10 | 691 | 55 | 2631 |
| TOL | Montes de Toledo, Cañamero | Spain | 5°21' W | 39°22' N | 800 | 15.21 | 33.30 | 3.17 | 1067 | 54 | 2290 |
| CAT | Cataluña, Sta Coloma Farnes | Spain | 2°32' E | 41°51' N | 500 | 14.99 | 29.47 | 2.50 | 801 | 140 | 1986 |
| HAZ | Haza de Lino | Spain | 3°18' W | 36°50' N | 1300 | 12.99 | 27.42 | 0.92 | 738 | 26 | 1831 |
| KEN | Kenitra, Ain Johra | Morocco | 6°35' W | 34°05' N | 160 | 12.99 | 26.83 | 8.17 | 536 | 9 | - |
| TAZ | Taza, Bab Azhar | Morocco | 4°15' W | 34°12' N | 1130 | 17.86 | 33.33 | 6.23 | 970 | 29 | - |
| MEK | Mekna, Tabarka | Tunisia | 8°51' E | 36°57' N | 12 | 17.87 | 31.20 | 8.40 | 948 | 27 | 2341 |
| KAV | Kavrakirovo | Bulgaria | 23°10' E | 41°26' N | 200 | 14.43 | 32.30 | 0.60 | 507 | 123 | 2428 |
| ARG | Guerbès | Algeria | - | - | - | - | - | - | - | - | - |

AMT, annual mean temperature; SMT, summer maximum temperature; WMT, winter minimum temperature; P, annual precipitation; Psum, summer precipitation; Ins, insolation.

### 2.3.2. Spatial and environmental data

Three spatial variables, corresponding to the original locations of the populations established at the provenance trial (as described in Varela 2000), were used for each population: altitude, latitude and longitude (Table 2.1). Climatic data was obtained

either by consulting Varela (2000) or the website Weather Online (http://www.weatheronline.co.uk) for the years comprised between 1996 and 2011 (Table 2.1). Fifty-four environmental variables were considered: annual mean temperature (AMT), summer maximum temperature (SMT), winter minimum temperature (WMT), annual precipitation (P), summer precipitation (Psum), insolation (Ins) (Table 2.1) and forty-eight monthly variables (mean temperature per month, maximum mean temperature per month, minimum mean temperature per month and precipitation per month).

### 2.3.3. SNP selection and validation

From the wide range of putative SNPs detected in cork oak 454 transcriptome pyrosequencing (data not published), 32 were selected to be validated (Table 2.2) through Sanger sequencing in order to assess if those SNPs corresponded to real biological variation or were the result of 454 technical artifacts. SNPs that potentially caused nonsynonymous mutations in putative orthologs of known genes were preferred and indels were not considered.

Primers were designed using PerlPrimer v1.1.10 (Marshall 2004) to amplify small fragments, each containing a single selected SNP (supplementary Table S1, Supporting Information). For the validation process, the amplified samples were the same as those used for the transcriptome 454 sequencing, which was carried out with a pool of 5 individuals from each of the following populations: Kenitra (Morocco), Mekna (Tunisia), Puglia (Italy), Haza de Lino, Cataluña (Spain), Gerês, Monchique and Sintra (Portugal).

PCR reactions 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), 1x 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 denaturation at 94 °C, followed by 35-40 cycles of 30 s at 94 °C, 30 s at variable annealing temperatures (supplementary table S2, Supporting Information) and 1 min at 72 °C, with a final extension step at 72 °C for 15 min. PCR products were visualized on 1% agarose gels to confirm amplification and subsequently purified using SureClean

(Bioline) purification protocol. PCR products successfully amplified were sequenced on ABI PRISM 310 or ABI 3730XL (Applied Biosystems) genetic analyzers. The obtained sequences were edited with Sequencher v4.0.5 (Gene Codes Corporation) and aligned using ClustalW (Thompson *et al.* 1994). The samples were gradually sequenced until both expected SNP alleles were detected for each fragment or alternatively all the samples were sequenced.

### 2.3.4. Population genetic analysis based on validated SNPs

From the successfully amplified and validated SNPs, five were selected for further population analyses in order to detect signatures of selection and to test for their potential involvement in adaptation to the environment. This choice was made based on four criteria: (i) the location of the SNP in the genome, with a clear preference for SNPs located in annotated genes; (ii) the putative function of the gene; (iii) the presence of other SNPs in the same fragment; and (iv) the type of mutations comprised in the amplified fragment (synonymous versus nonsynonymous SNPs or conservative versus non-conservative amino acid substitutions). Non-conservative mutations result in changes to amino acids with different physicochemical properties and are thus more prone to cause structural and/or functional modifications in the protein. Therefore, these mutations are of particular interest for selection and adaptation studies. The selected gene fragments were the putative orthologs of *Arabidopsis thaliana* (L.) Heynh. *RAN3*, *NPR1*, *PR1*, *ARF16* and *HSP* (Table 2.2). The open reading frame (ORF) of each gene was confirmed through the translation of the putative ORFs in BioEdit v7.1.3.0 (Hall 1999) and doing a BLASTp (http://blast.ncbi.nlm.nih.gov) subsequently. Five to six individuals from 16 to 19 populations were sequenced for each gene fragment (supplementary Table S2, Supporting Information), with the exception of *RAN3*, for which only two to five individuals were sequenced per population due to technical difficulties. Two to three samples of *Q. rotundifolia* were also sequenced for each gene fragment as outgroup and to account for introgression between the two species, as well as one sample of *Q. cerris*, which was sequenced for each fragment except for *RAN3* (supplementary Table S2, Supporting Information). Amplification, sequencing, sequence editing and alignment were performed as described in 2.3.3.

### 2.3.5. Statistical analyses

After sequence alignment, the heterozygous phase was determined using the program PHASE v2.1.1 (Stephens *et al.* 2001, Stephens and Scheet 2005), with default parameters, for *Q. suber* and *Q. rotundifolia* separately. Alignment files were converted to NEXUS format using CONCATENATOR v1.1.0 (Pina-Martins and Paulo 2008). Median-joining haplotype networks were then constructed for each fragment using NETWORK 4.6.1.0 (Bandelt *et al.* 1999).

Analyses of molecular variance (AMOVA) were performed employing ARLEQUIN v3.5 (Excoffier and Lischer 2010) to assess patterns of genetic differentiation among populations and groups of populations. Populations were grouped by latitude, longitude, east and west populations and according to the genetic structure described in Magri *et al.* (2007).

Number of haplotypes, haplotype diversity (H), number of polymorphic sites and nucleotide diversity ($\pi$) were computed with DnaSP v5.10 (Librado and Rozas 2009).

Hardy-Weinberg equilibrium tests were performed for each fragment, considering both locus by locus and whole haplotype levels using ARLEQUIN v3.5. Tajima's *D* (Tajima 1989) and Fu's *Fs* (Fu 1997) neutrality tests were also estimated using ARLEQUIN v3.5.

For site-specific sequence analysis of selective pressures acting on each fragment a Maximum Likelihood approach was implemented using CODEML from PAML v4.6 software package (Yang 2007). This analysis is based on the $\omega$ parameter, which compares the ratio of nonsynonymous mutations per nonsynonymous sites ($d_N$) to the number of synonymous mutations per synonymous sites ($d_S$). If there is no selection, synonymous and nonsynonymous substitutions should occur at the same rate, thus $\omega$ ($d_N/d_S$) would be expected to be 1. If there is negative selection, $d_N/d_S$ should be smaller than 1 and in case of positive selection $d_N/d_S$ should be higher than 1. 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 $\omega$ ($\omega=1$ and $\omega<1$), and M2, that estimates three values of $\omega$ ($\omega=1$, $\omega<1$ and $\omega>1$) and accounts for positive selection. Likelihood ratio tests (LRT) were performed to compare the three models and a $\chi^2$ 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).

Correlations between genetic data and spatial and climatic variables were tested using the program MatSAM v2 (Joost *et al.* 2007). Four genetic data sets were used: SNPs, nonsynonymous mutations only (amino acid data), haplotypes and haplotypes accounting only nonsynonymous mutations. Associations between genetic markers' frequencies and the spatial and environmental variables were assessed through series of univariate logistic regressions as implemented in the program.

## 2.4. Results

### 2.4.1. SNP validation

Thirty-two putative SNPs were selected to undergo the validation process resorting to Sanger sequencing. Nevertheless, only 19 fragments were successfully amplified and sequenced (59%) (Table 2.2). For the remaining fragments, no PCR products were obtained (SNPs 9-12 and 18), the amplified fragments were too long to be sequenced without designing internal primers (SNPs 5, 14, 21, 23 and 24) or, in a few cases, amplifications were unspecific (SNPs 2, 3 and 16) (Table 2.2). From the 19 SNPs, 11 were validated, confirming the existence of real biological variation, while three SNPs were not validated after sequencing all the samples used for the transcriptome sequencing, suggesting that those corresponded to technical errors. For the remaining five SNPs, variation was not found in the expected positions but it was not possible to sequence all the samples used in the 454 study. Consequently, these last five SNPs were discarded for the rest of the study, as no conclusions can be withdraw regarding to their authenticity. From the 14 SNPs that went through the complete validation process, 79% were validated.

**Table 2.2.** Single Nucleotide Polymorphisms (SNPs) subjected to the validation process, their BLAST protein result, putative protein function, SNP alleles, status of the validation process and observations about the validation process.

| SNP | BLAST protein | Putative protein function | SNP | Validation | Observations |
|---|---|---|---|---|---|
| SNP1 | Sulfoquinovosyldiacylglycerol 1 (SQD1) | sulfolipid biosynthesis | CT | Not validated | Variation not found |
| SNP2 | Gibberellin receptor GID1 | plant growth and development | CT | - | Unspecific PCR amplification |
| SNP3 | Dead box ATP-dependent RNA helicase | RNA methylation | CT | - | Unspecific PCR amplification |
| SNP4 | Nonexpressor of pathogenesis-related genes 1 (NPR1) | plant defense; transcription regulation | CG | Validated | Variation found in expected position |
| SNP5 | V-type proton ATPase catalytic subunit A (V-ATPase subunit A) | vacuolar ATP hydrolysis coupled proton transport | CT | - | Amplified fragment too long for sequencing |
| SNP6 | RAS-related nuclear protein 3 (RAN3) | nucleocytoplasmic transport | CT | Validated | Variation found in expected position |
| SNP7 | ATP binding protein | protein phosphorylation | CT | Not validated | Variation not found |
| SNP8 | Annexin 8 | response to stress | GT | Validated | Variation found in expected position |
| SNP9 | Predicted protein | unkown | AC | - | No PCR products amplified |
| SNP10 | Ribokinase | ribose metabolism | CT | - | No PCR products amplified |
| SNP11 | Predicted protein | unkown | TC | - | No PCR products amplified |
| SNP12 | Predicted protein | unkown | CG | - | No PCR products amplified |
| SNP13 | Autoinhibited calcium ATPase | transmenbrane transport | AG | Not validated | Variation not found |
| SNP14 | S-phase kinase-associated protein 1 (SKP1) | meiosis and mitosis regulation | GT | - | Amplified fragment too long for sequencing |
| SNP15 | Pathogenesis-related protein 1 (PR1) | plant defense | GT | Validated | Variation found in expected position |
| SNP16 | Aquaporin PIP2;3 | response to stress | AC | - | Unspecific PCR amplification |
| SNP17 | MYC2 | plant defense; transcription regulation | AG | Validated | Variation found in expected position |
| SNP18 | Methionine synthase (MS2) | methionine biosynthesis; response to stress | AC | - | No PCR products amplified |
| SNP19 | Xyloglucan 6-xylosyltransferase | xyloglucan synthesis | CG | - | In validation process |
| SNP20 | Class I small heat shock protein (HSP) | response to stress | AC | Validated | Variation found in expected position |
| SNP21 | Glyceraldehyde-3-phosphate dehydrogenase C (GAPC) | response to stress; plant defence | CT | - | Amplified fragment too long for sequencing |
| SNP22 | Glutamine synthetase nodule isozyme (GLN) | nitrogen fixation | GT | Validated | Variation found in expected position |
| SNP23 | 2-methyl-6-phytylbenzoquinone methyltransferase (MPBQ) | plastoquinone and vitamin E biosynthesis | GT | - | Amplified fragment too long for sequencing |
| SNP24 | Malate dehydrogenase (MDH) | carbohydrate metabolic process | AG | - | Amplified fragment too long for sequencing |
| SNP25 | Triacylglycerol lipase | lipid metabolism; regulation of seed germination | CT | Validated | Variation found in expected position |
| SNP26 | Mitogen-activated protein kinase 2 (MPK2) | plant defense | AT | - | In validation process |
| SNP27 | Tubby-like protein 3 (TLP3) | response to stress; transcription regulation | CT | - | In validation process |
| SNP28 | Nuclear transcription factor Y subunit A-7 isoform 1 (NFYA7) | response to stress; transcription regulation | CT | Validated | Variation found in expected position |
| SNP29 | Auxin response factor 16 (ARF16) | plant growth and development; transcription regulation; stress response | AG | Validated | Variation found in expected position |
| SNP30 | Transcription factor bHLH51 | transcription regulation | CT | Validated | Variation found in expected position |
| SNP31 | Protein FRIGIDA-like | flowering | GT | - | In validation process |
| SNP32 | Jasmonate ZIM-domain protein | plant defense; transcription regulation | AG | - | In validation process |

### 2.4.2. Characterization of selected markers and genetic diversity

Five of the validated SNPs were selected for further analysis. The fragments chosen were located in the putative orthologs of *A. thaliana RAN3*, *NPR1*, *PR1*, *ARF16* and *HSP* genes, for which the respective putative function is described in Table 2.2.

The amplified fragment from *RAN3* includes part of an intron, followed by a small exon, a second intron, a second exon with a stop codon and the following non-coding region. These exons correspond to the C-terminal region of the protein. The putative ORF detected in the 454 transcriptome analysis was not the real ORF, which was identified by running BLASTp against the translated nucleotide sequence. For this reason, the putative nonsynonymous SNP was in fact in the non-coding region after the stop codon. Other SNPs were found in the amplified fragment: 10 in the introns and two synonymous mutations in the exons. Therefore, a total of 13 polymorphic sites were detected and a total of 12 haplotypes were found (Table 2.3).

For *NPR1* gene, the amplified fragment comprises a partial exon that includes a segment of the ankirin repeats (ANK) domain and a segment of the C-terminal of the protein encoded. In addition to the validated SNP (nonsynonymous and non-conservative) four polymorphic sites were found, giving a total of five SNPs (Table 2.3). One is a synonymous mutation, another one is a nonsynonymous conservative mutation and the remaining two are nonsynonymous and non-conservative mutations. A total of four haplotypes were detected.

The amplified *PR1* fragment comprises an extracellular SCP-like domain known as "PR1-fold" that is believed to contribute to maintain the protein structure in extracellular environment (Van Loon *et al.* 2006). Besides the validated SNP, four additional polymorphic sites were found (Table 2.3), corresponding to a total of one synonymous mutation and four nonsynonymous and non-conservative mutations. A total of five haplotypes were found.

*ARF16* amplified fragment is likely to correspond to the central non-conserved domain of the protein, which is responsible for its transcriptional activation/repression function. Four polymorphisms were detected in this fragment (Table 2.3), including the validated SNP. One is a synonymous mutation and three are nonsynonymous, two of which are non-conservative. Five haplotypes were detected in total.

**Table 2.3.** Fragments chosen for population genetic studies, length of the amplified fragments and diversity indexes and neutrality tests estimated for these fragments.

| Fragment | Length (bp) | Number of haplotypes | Polymorphic sites | Synonymous mutations | Nonsynonymous mutations | Mutations in non-coding regions | H | $\pi$ | Tajima's *D* | Fu's *Fs* |
|---|---|---|---|---|---|---|---|---|---|---|
| *RAN3* | 627 | 12 | 13 | 2 | - | 11 | 0.597 | 0.00681 | 2.00000 | 1.90400 |
| *NPR1* | 270 | 4 | 5 | 2 | 3 | - | 0.566 | 0.00688 | 2.30457* | 5.19204*** |
| *PR1* | 257 | 5 | 5 | 1 | 4 | - | 0.609 | 0.00519 | 1.06458 | 2.03902 |
| *ARF16* | 234 | 5 | 4 | 1 | 3 | - | 0.648 | 0.00672 | 2.31433* | 2.78844* |
| *HSP* | 374 | 14 | 10 | 2 | 4 | 4 | 0.836 | 0.00619 | 0.81103 | -1.66710 |

H, haplotype diversity; π, genetic diversity. * $P < 0.05$; *** $P < 0.001$.

For *HSP*, the amplified fragment comprises part of the C-terminal alpha crystalline domain (ACD) and the non-codifying sequence after the stop codon. In the coding region, six polymorphic sites were detected (including the validated SNP), corresponding to two synonymous and four nonsynonymous and non-conservative mutations. In the non-coding region, four SNPs were found. In this way, a total of 10 SNPs and 14 haplotypes were detected in this fragment (Table 2.3).

In general, low genetic diversity was found (Table 2.3). The number of haplotypes found is lower for *NPR1*, *PR1* and *ARF16* than for *RAN3* and *HSP*, which included non-coding regions.

At the whole haplotype level analysis, all gene fragments were found to be in Hardy-Weinberg equilibrium (supplementary table S3, Supporting Information), except for *HSP* ($P<0.001$). In this fragment, the locus by locus analysis detected that positions 4 (which corresponds to a nonsynonymous and non-conservative mutation) ($P<0.001$) and 8 (mutation in the non-coding region) ($P<0.001$) had excess of heterozygotes. In *RAN3*, two positions were also detected as in Hardy-Weinberg disequilibrium, positions 9 (synonymous mutation) ($P<0.001$) and 13 (mutation in the non-coding region) ($P<0.05$), both with heterozygote deficit.

### 2.4.3. Phylogeography and population genetic structure

The haplotype networks constructed for the five fragments demonstrated a lack of phylogeographic structure, with haplotypes being frequently shared by geographically distant populations (Figures 2.2 and 2.3, Table 2.4 and supplementary Figure S1, Supporting Information).

**Table 2.4.** Number of haplotypes (N) and haplotype distribution throughout the sampled populations for *RAN3*, *NPR1*, *PR1*, *ARF16* and *HSP* fragments.

| Gene | *RAN3* | | | | | | | | | | | | | | *NPR1* | | | | | *PR1* | | | | | | *ARF16* | | | | | | *HSP* | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | H1 | H2 | H3 | H4 | H5 | H6 | H7 | H8 | H9 | H10 | H11 | H12 | | N | H1 | H2 | H3 | H4 | N | H1 | H2 | H3 | H4 | H5 | N | H1 | H2 | H3 | H4 | H5 | N | H1 | H2 | H3 | H4 | H5 | H6 | H7 | H8 | H9 | H10 | H11 | H12 | H13 | H14 |
| **Populations** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Sintra | 2 | | | | 5 | | 5 | | | | | | | | 3 | 3 | | 1 | 8 | 4 | 6 | 4 | 1 | | 1 | 3 | 9 | 1 | | | 2 | 4 | 2 | 1 | 2 | | | 1 | | | | | | | 4 | |
| Gerês | 2 | | | | 3 | | 5 | | | | | | | | 3 | 3 | | 1 | 6 | - | | | | | | - | | | | | | - | | | | | | | | | | | | | |
| Monchique | 3 | | 1 | | 4 | | 5 | | | | | | | | 3 | 4 | | 3 | 5 | 2 | 10 | 2 | | | | 4 | 3 | 2 | 2 | | 5 | 7 | 2 | 3 | 2 | | 1 | | | | 1 | 1 | | | 2 | |
| S. Bras Alportel | 4 | | | 1 | | | 1 | | | | | 1 | 1 | | 2 | 3 | | | 7 | - | | | | | | - | | | | | | - | | | | | | | | | | | | | |
| Kenitra | 2 | | | | 1 | | 7 | | | | | | | | 2 | | | 4 | 8 | 3 | 4 | 5 | | | 3 | 2 | 6 | | | | 6 | 4 | | | 4 | | | 1 | | | 3 | | | | 2 | |
| Taza | 6 | 1 | | | 1 | | 3 | | | 1 | | 1 | 1 | | 3 | 3 | | 1 | 8 | 4 | 7 | 2 | 2 | | 1 | 2 | 7 | | | | 5 | 6 | 1 | 1 | 4 | 1 | | | | 1 | | | | | 4 | |
| Montes Toledo | 2 | | | | 4 | | 6 | | | | | | | | 2 | 6 | | | 6 | 4 | 4 | 5 | 1 | | 2 | 4 | 5 | 2 | 2 | | 3 | 5 | 2 | | 3 | | | 1 | | | 5 | | | | 1 | |
| Haza de Lino | 2 | | | | 6 | | 4 | | | | | | | | 2 | 6 | | | 6 | 3 | 9 | 2 | | | 1 | 2 | 8 | | | | 4 | 6 | 4 | | 2 | | | 1 | | | 2 | | | | 2 | 1 |
| Guerbès | - | | | | | | | | | | | | | | 3 | 2 | | 2 | 2 | - | | | | | | - | | | | | | - | | | | | | | | | | | | | |
| Landes | 4 | 1 | | | 1 | | 3 | | 1 | | | | | | 3 | 1 | | 4 | 7 | 3 | 9 | 2 | | | 1 | 3 | 3 | 4 | 5 | | | 4 | | 2 | 4 | | | 1 | | | 5 | | | | | |
| Cataluña | 2 | | | | 2 | | 4 | | | | | | | | 3 | 5 | | 3 | 4 | 3 | 6 | 5 | | | 1 | 2 | 10 | 2 | | | | 3 | | 1 | 5 | | | | | | 6 | | | | | |
| Var | 2 | | | | | | 3 | | | | 1 | | | | 2 | 9 | | | 3 | 3 | 5 | 4 | | | 3 | 3 | 1 | 1 | | | 10 | 5 | 1 | 1 | 4 | | | | | | 4 | | | | 2 | |
| Lazio | 4 | | | | 3 | | 4 | | | 1 | 2 | | | | 2 | 3 | | | 9 | 3 | 2 | 9 | 1 | | | 3 | 8 | 1 | | | 3 | 5 | 1 | | 3 | | | 3 | | 1 | 4 | | | | | |
| Corse | 1 | | | | | | 4 | | | | | | | | 2 | 1 | | | 11 | 3 | 9 | 2 | | 1 | | 3 | 8 | | 1 | | 3 | 6 | | | 1 | 1 | | 3 | | | 5 | | 1 | 1 | | |
| Sardegna | - | | | | | | | | | | | | | | 3 | 3 | 1 | | 6 | 3 | 5 | 6 | | | 1 | 4 | 3 | 4 | 1 | | 4 | 7 | 1 | | 2 | 1 | | 3 | 1 | | 3 | | 1 | | | |
| Mekna | 2 | | | | 3 | | 7 | | | | | | | | 3 | 2 | | 1 | 9 | 3 | 5 | 4 | | 1 | | 3 | 7 | 3 | | | 2 | 6 | 2 | | 3 | | | 1 | | | 2 | | 1 | | 3 | |
| Sicilia | 2 | | | | 2 | | 4 | | | | | | | | 2 | 5 | | | 7 | 5 | 7 | 2 | 1 | 1 | 1 | 3 | 4 | 1 | | | 7 | 6 | | | 4 | 1 | 2 | 1 | | | 2 | | | | 2 | |
| Puglia | 2 | | | | | 1 | 7 | | | | | | | | 3 | 3 | | 1 | 8 | 3 | 2 | 4 | | | 4 | 4 | 8 | 2 | 1 | 1 | | 8 | 1 | | 5 | | 1 | 1 | 1 | | 1 | 1 | | 1 | | |
| Kavrakirovo | 4 | | | | 3 | | 4 | 2 | | 1 | | | | | 3 | 5 | | 3 | 5 | 2 | 9 | 3 | | | | 4 | 6 | 2 | 3 | | 1 | 6 | 4 | 1 | 3 | | | 1 | | | 2 | | | | 1 | |
| **Total** | - | 2 | 1 | 1 | 38 | 1 | 76 | 2 | 1 | 3 | 3 | 2 | 2 | | - | 67 | 1 | 24 | 125 | - | 99 | 61 | 6 | 3 | 19 | - | 96 | 25 | 15 | 1 | 55 | - | 21 | 10 | 51 | 4 | 4 | 18 | 2 | 1 | 46 | 2 | 3 | 2 | 23 | 1 |

*NPR1* analysis revealed three common haplotypes spread through almost every population (Table 2.4, Figure 2.2a and b). Haplotype 1 is the second most common haplotype and differs from haplotype 2 only by one synonymous mutation, from haplotype 3 by two nonsynonymous and non-conservative mutations, and from haplotype 4, the most common one, by three nonsynonymous mutations (Figure 2.2a). The most common haplotype is therefore the most derived one (haplotype 4) and is shared with one *Q. rotundifolia* individual, although this haplotype is very distant from the other *Q. rotundifolia* haplotypes, while haplotype 1, the most ancestral one, is shared with *Q. cerris*.

For *ARF16*, the most ancestral haplotype (haplotype 1) is more common than the others and distributed throughout the analyzed populations (Table 2.4, Figure 2.2d). Two other haplotypes are also rather common and dispersed through almost every population (haplotypes 2 and 5). Haplotype 2 diverges from haplotype 1 by one synonymous mutation. On the other hand, haplotype 1 differs from haplotype 3 by two nonsynonymous mutations, one of them non-conservative. Haplotype 5, the most derived and second most common one, differs from haplotype 3 by one nonsynonymous and non-conservative mutation. One *Q. rotundifolia* individual was found to share haplotype 5 with cork oak (Figure 2.2c), although the other *Q. rotundifolia* haplotypes are very distant. Haplotype 1 is shared with cork oak both by *Q. rotundifolia* and *Q. Cerris*.

The haplotype network obtained for *HSP* was much more complex than the other gene networks (Figure 2.2e). However, no evident geographic structure was found (Table 2.4, Figure 2.2f). Two haplotypes are more common and dispersed throughout almost all the analyzed populations (haplotypes 3 and 9). These two haplotypes differ by one synonymous and one nonsynonymous and non-conservative mutation. Haplotypes 1, 6 and 13 are also rather spread throughout cork oak populations, without any apparent geographic structure. Haplotypes 1 and 6 differ from haplotype 3 only by synonymous mutations, while haplotype 13 differs by one nonsynonymous and non-conservative mutation from haplotype 9 and two nonsynonymous and non-conservative mutations from haplotype 3, besides the synonymous mutations. Other rarer haplotypes are also scattered through the analyzed populations without evident structure. Haplotype 3 is shared with *Q. cerris* (Figure 2.2e) and with one *Q. rotundifolia* individual, while haplotype 13 is also shared with *Q. rotundifolia*.

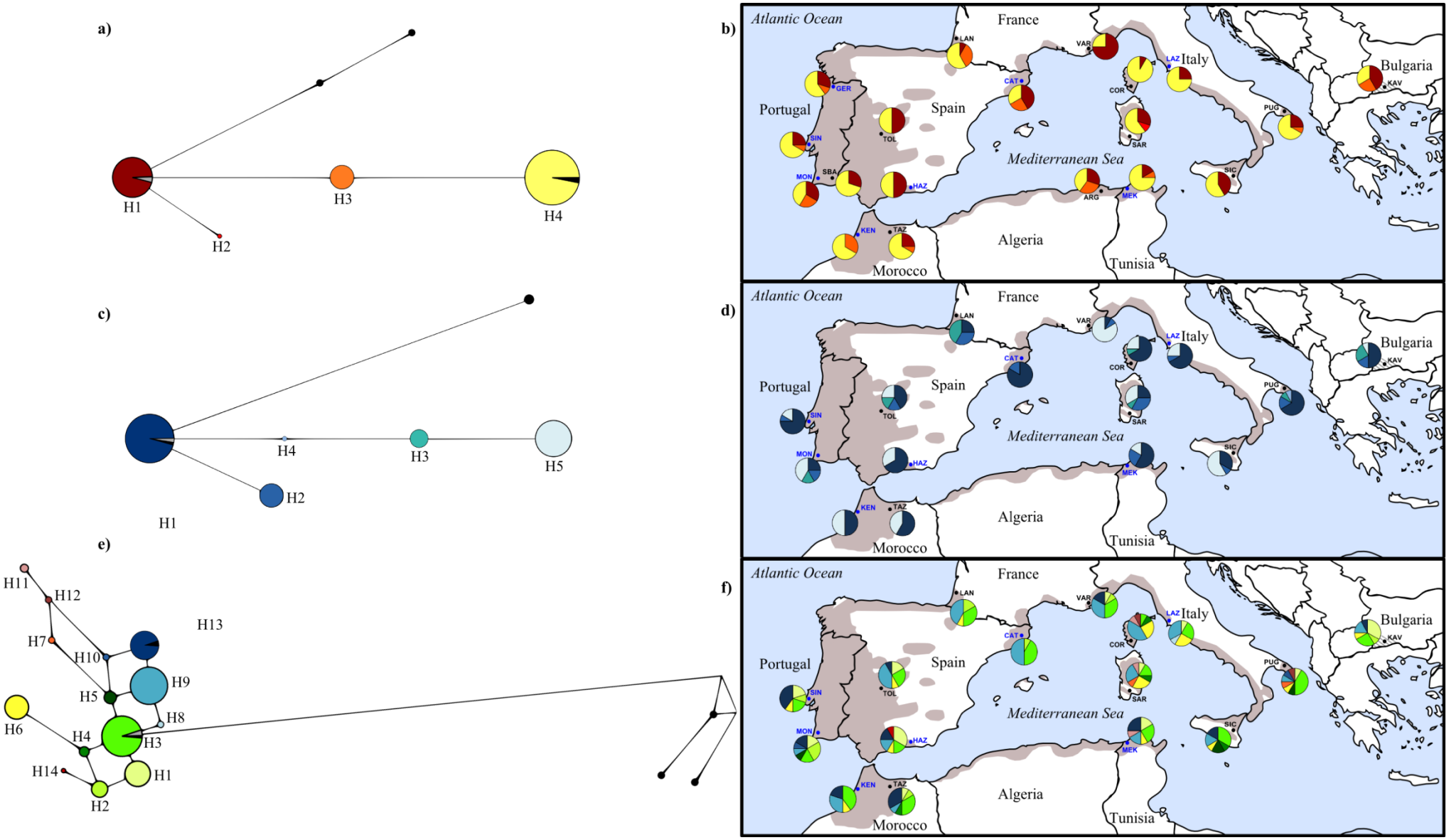



**Figure 2.2.** Median Joining haplotype networks for *NPR1* (a), *ARF16* (c) and *HSP* (d) fragments. Each haplotype is represented by a different colour. Haplotype geographical distribution is represented in maps b) *NPR1*, d) *ARF16* and f) *HSP*, with haplotype colours corresponding to the ones in the networks. *Q. rotundifolia* haplotypes are represented in black; *Q. cerris* haplotypes are represented in grey.

For *RAN3* (Figure 2.3a), two of the detected haplotypes are common to almost every population (Table 2.4, and supplementary Figure S1a, Supporting Information) and other rare haplotypes are spread throughout the populations without any apparent geographic structure.

Likewise, two of the *PR1* haplotypes are very common and were found in all the analyzed populations (haplotypes 1 and 2) (Table 2.4, supplementary Figure S1b, Supporting Information). Haplotype 1, the most common, is the most ancestral one and haplotype 2, the second most common, diverges from it by two nonsynonymous and non-conservative mutations. Haplotype 5 was also found in almost every population and is shared with *Q. cerris* (Figure 2.3b), while the other haplotype detected in this species differs from it by one mutation. The remaining haplotypes are rare and more scarcely distributed throughout cork oak populations, without any evident geographic structure (supplementary Figure S1b, Supporting Information).

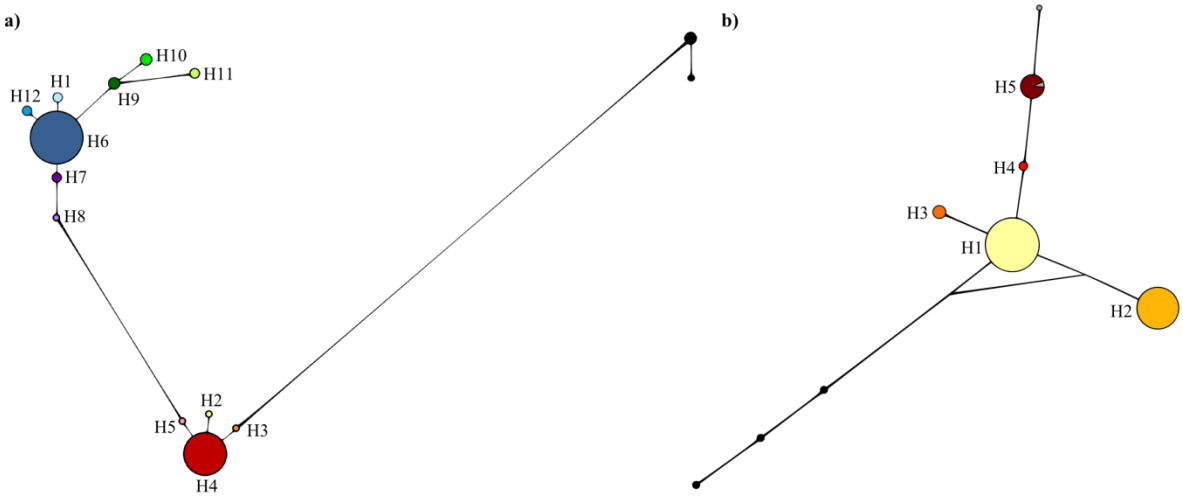



**Figure 2.3.** Median Joining haplotype networks for *RAN3* (a) and *PR1* (b) fragments. Each haplotype is represented by a different color. Haplotype geographical distribution is represented in maps in supplementary Figure S1, Supporting Information, with haplotype colors corresponding to the ones in the networks. *Q. rotundifolia* haplotypes are represented in black; *Q. cerris* haplotypes are represented in grey.

The same lack of genetic structure seems to be evident for all 5 fragments from the analysis of molecular variance (AMOVA), at least for the tested groups of population (Table 2.5). For all analyses, the overall source of variation was within populations and not among populations within groups or among groups.

**Table 2.5.** Results of the analysis of molecular variance (AMOVA) for each fragment. Four groups of populations were tested in the following order: East and West populations; populations grouped according to the genetic structure described in Magri *et al.* (2007); populations grouped by latitudes; and populations grouped by longitudes. Negative values must be interpreted as zero (Long 1986) in the AMOVA model.

| Gene | Groups | Variation within populations | | | Variation among populations | | | Variation among groups | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | % var | $F_{ST}$ | *P* | % var | $F_{SC}$ | *P* | % var | $F_{CT}$ | *P* |
| *RAN3* | [VAR,MEK,COR,PUG,LAZ,SIC][TAZ,CAT,HAZ,TOL,SIN,MON,KEN,LAN,GER,SBA] | 96.94 | 0.031 | 0.1971 | -0.16 | -0.002 | 0.5336 | 3.22 | 0.032 | 0.0025 |
| | [TAZ,CAT,HAZ][VAR,MEK,COR][PUG,LAZ,SIC][TOL,SIN,GER,MON,KEN,LAN,SBA] | 98.58 | 0.014 | 0.1858 | 1.42 | 0.014 | 0.1900 | 0.01 | 0.000 | 0.5188 |
| | [KAV,PUG,COR,CAT,GER][SIN,TOL][MEK,HAZ,SIC,MON,SBA][KEN,TAZ][LAZ][LAN,VAR] | 98.67 | 0.013 | 0.3483 | -3.46 | -0.036 | 0.9654 | 4.79 | 0.048 | 0.0000 |
| | [SIN,MON,GER,SBA][TAZ,HAZ,KEN,TOL][LAN,CAT][VAR,LAZ,COR,MEK][SIC][PUG][KAV] | 99.66 | 0.003 | 0.3503 | 1.88 | 0.018 | 0.2016 | -1.54 | -0.015 | 0.8992 |
| *NPR1* | [VAR,MEK,COR,SAR,PUG,LAZ,SIC,ARG][TAZ,CAT,HAZ,TOL,SIN,MON,KEN,LAN,GER,SBA] | 94.04 | 0.060 | 0.0004 | 6.83 | 0.068 | 0.0000 | -0.87 | -0.009 | 0.7722 |
| | [TAZ,CAT,HAZ][VAR,MEK,COR,SAR,ARG][PUG,LAZ,SIC][TOL,SIN,GER,MON,KEN,LAN,SBA] | 94.04 | 0.060 | 0.0000 | 7.77 | 0.076 | 0.0000 | -1.82 | -0.018 | 0.9254 |
| | [KAV,PUG,COR,CAT,GER][SAR,SIN,TOL][MEK,HAZ,SIC,MON,SBA][KEN,TAZ][LAZ][LAN,VAR] | 94.05 | 0.060 | 0.0005 | 7.52 | 0.074 | 0.0000 | -1.57 | -0.016 | 0.8439 |
| | [SIN,MON,GER,SBA][TAZ,HAZ,KEN,TOL][LAN,CAT][VAR,LAZ,COR,SAR,MEK][SIC][PUG][KAV] | 94.19 | 0.058 | 0.0005 | 8.77 | 0.085 | 0.0005 | -2.96 | -0.030 | 0.9348 |
| *PR1* | [VAR,MEK,COR,SAR,PUG,LAZ,SIC][TAZ,CAT,HAZ,TOL,SIN,MON,KEN,LAN] | 93.48 | 0.065 | 0.0079 | 3.29 | 0.034 | 0.0376 | 3.23 | 0.032 | 0.0267 |
| | [TAZ,CAT,HAZ][VAR,MEK,COR,SAR][PUG,LAZ,SIC][TOL,SIN,MON,KEN,LAN] | 94.88 | 0.051 | 0.0084 | 4.98 | 0.050 | 0.0074 | 0.14 | 0.001 | 0.4126 |
| | [KAV,PUG,COR,CAT][SAR,SIN,TOL][MEK,HAZ,SIC,MON][KEN,TAZ][LAZ][LAN,VAR] | 94.46 | 0.055 | 0.0030 | 1.60 | 0.017 | 0.1912 | 3.94 | 0.039 | 0.0366 |
| | [SIN,MON][TAZ,HAZ,KEN,TOL][LAN,CAT][VAR,LAZ,COR,SAR,MEK][SIC][PUG][BUL] | 94.46 | 0.055 | 0.0030 | 1.92 | 0.020 | 0.1685 | 3.62 | 0.036 | 0.0771 |
| *ARF16* | [VAR,MEK,COR,SAR,PUG,LAZ,SIC][TAZ,CAT,HAZ,TOL,SIN,MON,KEN,LAN] | 95.77 | 0.042 | 0.0000 | 5.70 | 0.056 | 0.0020 | -1.47 | -0.015 | 0.9335 |
| | [TAZ,CAT,HAZ][VAR,MEK,COR,SAR][PUG,LAZ,SIC][TOL,SIN,MON,KEN,LAN] | 95.32 | 0.047 | 0.0049 | 5.66 | 0.056 | 0.0064 | -0.98 | -0.010 | 0.7194 |
| | [KAV,PUG,COR,CAT][SAR,SIN,TOL][MEK,HAZ,SIC,MON][KEN,TAZ][LAZ][LAN,VAR] | 95.52 | 0.045 | 0.0099 | 2.45 | 0.025 | 0.1215 | 2.03 | 0.020 | 0.1349 |
| | [SIN,MON][TAZ,HAZ,KEN,TOL][LAN,CAT][VAR,LAZ,COR,SAR,MEK][SIC][PUG][BUL] | 95.36 | 0.046 | 0.0109 | 1.61 | 0.017 | 0.2614 | 3.03 | 0.030 | 0.0721 |
| *HSP* | [VAR,MEK,COR,SAR,PUG,LAZ,SIC][TAZ,CAT,HAZ,TOL,SIN,MON,KEN,LAN] | 100.57 | -0.006 | 0.8375 | -1.99 | -0.020 | 0.9496 | 1.42 | 0.014 | 0.0148 |
| | [TAZ,CAT,HAZ][VAR,MEK,COR,SAR][PUG,LAZ,SIC][TOL,SIN,MON,KEN,LAN] | 101.22 | -0.012 | 0.8434 | -1.31 | -0.013 | 0.8058 | 0.09 | 0.001 | 0.4308 |
| | [KAV,PUG,COR,CAT][SAR,SIN,TOL][MEK,HAZ,SIC,MON][KEN,TAZ][LAZ][LAN,VAR] | 101.10 | -0.011 | 0.7846 | -0.21 | -0.002 | 0.5366 | -0.89 | -0.009 | 0.8646 |
| | [SIN,MON][TAZ,HAZ,KEN,TOL][LAN,CAT][VAR,LAZ,COR,SAR,MEK][SIC][PUG][BUL] | 100.69 | -0.007 | 0.7826 | -2.55 | -0.026 | 0.9793 | 1.86 | 0.019 | 0.0272 |

### 2.4.4. Neutrality tests

Both Tajima's *D* and Fu's $F_S$ neutrality tests rejected the null neutral model on two of the fragments, *NPR1* and *ARF16* (Table 2.3). Positive *D* values indicate an excess of intermediate frequency alleles, consistent with balancing selection or population decline. Positive values of Fu's $F_S$ indicate a deficit of alleles, suggesting also balancing selection or a bottleneck.

PAML analyses were performed for all fragments except for *RAN3*, as no nonsynonymous mutations were found in this fragment. The selection model (M2) was not significantly more adjusted to the data than the neutral model (M1) for any of the five fragments investigated. However, for *HSP*, M2 likelihood was higher than M1, even though not significantly, and the selection model detected three positions potentially under positive selection, one of them with $P<0.05$ (amino acid position 17) (Figure 2.4). For *NPR1* and *PR1* similar results were obtained, as M2 likelihood was not significantly higher than M1 likelihood and a few positions were detected as possibly being under selection, although none of them had significant *P*-values. For *NPR1*, the position detected was amino acid position 29, corresponding to the conservative mutation found in this fragment, while for *PR1* four positions were detected (amino acid positions 1, 56, 68, 70), corresponding to all four nonsynonymous mutations comprised in this fragment.

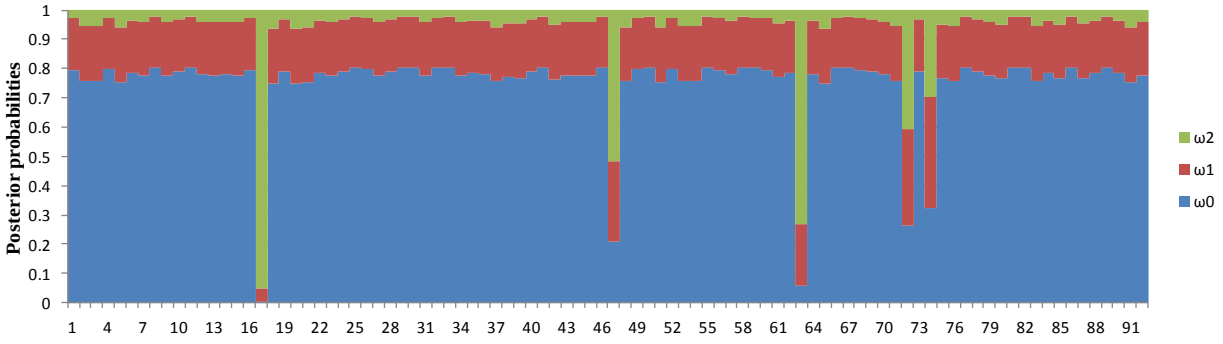



**Figure 2.4.** Posterior probability distribution of three classes of ω ($\omega_0$, $\omega_1$ and $\omega_2$) across the transcribed *HSP* fragment inferred from M2 model of PAML. The proportion of each site class with overall $\omega_0 = 0.0$, $\omega_1 = 1.00$ and $\omega_2 = 7.65$, are $p_0 = 0.914$, $p_1 = 0.000$ and $p_2 = 0.086$, respectively.

### 2.4.5. Association with environmental variables

Although several genetic data sets were used for the environmental associations, significant associations were found for the amino acid data set only. Two associations were found in *HSP* fragment: a positive correlation between aspartate frequency in the amino acid position 17 and latitude (Figure 2.5a), found to be significant by the dynamic null hypothesis analysis for G test ($P<0.05$), and also a positive correlation between aspartate frequency in position 17 and precipitation in September (Figure 2.5b), found to be significant both by dynamic null hypothesis analysis for G test and Wald Beta 1 test ($P<0.1$). Therefore, the probability of position 17 having aspartate increases with latitude and precipitation in September. It is also evident from the data that precipitation in September increases with latitude (Figure 2.5c) and populations under higher precipitation tend to have higher aspartate frequency (Figure 2.5c and d).

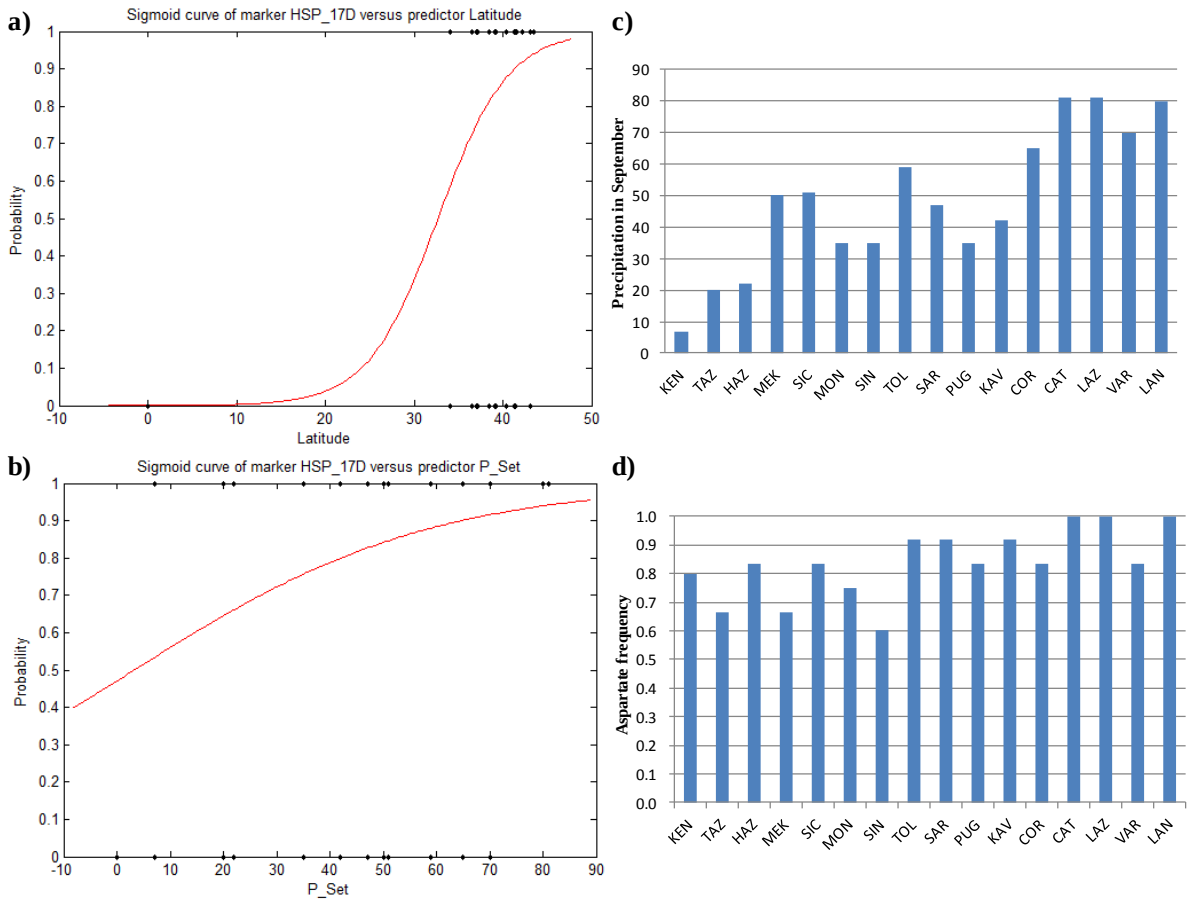



**Figure 2.1.** Results from MatSAM logistic regressions. On the left are the graphs of the significant correlations detected by the program: a) correlation of the probability of finding aspartate (D) on position 17 and latitude; b) correlation of the probability of finding aspartate (D) on position 17 and precipitation in September (P_Set). On the right, bar graphs representing precipitation in September (mm) (c) and aspartate frequency (d) in sampled populations sorted by increasing latitudes. KEN, Kenitra; TAZ, Taza; HAZ, Haza de Lino; MEK, Mekna; SIC, Sicilia; MON, Monchique; SIN, Sintra; TOL, Montes de Toledo; SAR, Sardegna; PUG, Puglia; KAV, Kavrakirovo; COR, Corse; CAT, Cataluña; LAZ, Lazio; VAR, Var; LAN, Landes.

## 2.5. Discussion

### 2.5.1. SNP validation

For many of the putative SNPs selected for the validation process, the attempts to amplify the corresponding fragments were unsuccessful. Low levels of success have been reported in other studies when amplifying genomic DNA with primers designed from expressed sequence tags (ESTs) (e.g. Zhu *et al.* 2003, Coles *et al.* 2005). However, in the present study higher levels of success were obtained than in the referred studies (59% versus 43%). Amplification failure can be attributed mostly to variation at primer annealing site and the presence of large introns within PCR amplicons. Accordingly, the amplified fragments for some of the investigated putative genes were too long to be sequenced, probably due to the presence of large introns.

Three of the SNPs analyzed could not be validated (Table 2.2), since the expected variation was not found even after sequencing all the samples used for the transcriptome sequencing, suggesting that these SNPs do not correspond to biological variation, resulting rather from 454 technical errors. In previous works (Huse *et al.* 2007, Kunin *et al.* 2010), the presence of homopolymers, i.e. repetitive sequences of identical bases (e.g. AAAA), have been reported as being associated with some errors in 454 pyrosequencing, resulting from difficulties to resolve the intensity of luminescence produced when a homopolymer is encountered. Errors caused by homopolymer effects include extensions (insertions), incomplete extensions (deletions) and carry-forward errors (insertions and substitutions). Carry-forward errors occur when an incomplete flush of base flow results in a premature incorporation of a base, usually near but not adjacent to the homopolymer (Margulies *et al.* 2005). The three putative SNPs that were not validated were located upstream from short homopolymers composed by a repeat of the base that was not found by Sanger sequencing. Therefore, these putative SNPs are most likely the result of carry-forward errors. Nonetheless, most of the successfully amplified and sequenced SNPs were validated (79%), results that are not very different from those obtained in another study where validation was carried out with Sanger sequencing (85%) (Barbazuk *et al.* 2007), supporting that transcriptome pyrosequencing is a useful tool to find and develop new genetic markers, particularly for non-model organisms, for which genetic tools are scarce.

### 2.5.2. Diversity, plylogeography and genetic structure

The low number of haplotypes found for *NPR1, PR1* and *ARF16* is comparable to what was found in another study involving cork oak nuclear DNA (ITS) (Simeone *et al.* 2009). However, in the present study, higher numbers of haplotypes were found in fragments that included non-coding regions (*RAN3* and *HSP*), revealing that the low number of haplotypes found for *NPR1, PR1* and *ARF16* might be due to mutational constraints in coding regions, as this may disrupt protein function.

For all analyzed fragments, low population genetic structure was detected, although in previous works using chloroplastidial and nuclear neutral markers geographical structure was detected (e.g. Lumaret *et al.* 2005, Magri *et al.* 2007, Simeone *et al.* 2009). Cork oaks are long-lived organisms, experiencing varying environmental conditions throughout its lifespan, and are outcrossing trees, with long-distance pollen dispersal. These life-history traits usually result in low differentiation among populations at nuclear markers (Austerlitz *et al.* 2000).

In the haplotype network analysis for *NPR1* fragment, cork oak haplotype 4 was shared with one *Q. rotundifolia* individual, although the other *Q. rotundifolia* haplotypes were very distant from this one (Figure 2.2a). Likewise, for *ARF1* and for *HSP* two haplotypes (haplotypes 1 and 5 and haplotypes 3 and 13 respectively) (Figure 2.2c and e) were shared with *Q. rotundifolia*, even though the other *Q. rotundifolia* were separated by several mutations. These results suggest the occurrence of hybridization events between cork oak and lineages from *Q. rotundifolia*. Hybridization events and haplotype sharing between these two species were not surprising, as they are well described in the literature (e.g. Jimenez *et al.* 2004, Lopez-de-Heredia *et al.* 2007, Magri *et al.* 2007, Costa 2011, Costa *et al.* 2011).

In *NPR1, ARF16* and *HSP*, the ancestral haplotypes were shared with *Q. cerris* (Figure 2.2a, c and d). Since this species are closely related to cork oak this is probably due to incomplete lineage sorting, although the occurrence of hybridization events cannot be completely excluded. For *PR1*, haplotype 5 was also shared with *Q. cerris* (Figure 2.3b), although the other *Q. cerris* haplotype differed from it for one mutation. Therefore, it is not clear in this case if this is due to incomplete lineage sorting or a hybridization event between cork oak and a lineage of *Q. cerris*.

### 2.5.3. Fragments under putative balancing selection

Neutrality tests for *NPR1* (Tajima's D and Fu's *Fs*) were significant and positive, suggesting that either this fragment is under balancing selection or cork oak populations are declining. However, demographic effects should affect all the tested fragments. Since this signal was detected for only two of them, *NPR1* is likely to be under balancing selection. Moreover, the few haplotypes found for this fragment presented no geographical structure. PAML results provide an indication that one nonsynonymous conservative mutation may be under positive selection, which suggests that even conservative mutations may have an impact on fitness. However, the values were not significant, which suggests that this result is a false positive or that positive selection signal is weak.

*NPR1* fragment has three nonsynonymous mutations in the amplified region corresponding to the ANK domain, two of which are non-conservative. Non-conservative mutations can affect the structure and therefore also the function of the protein, as the original amino acid is replaced by one with different physicochemical properties. Conservative mutations are expected to have smaller effects, although when combined with other nonsynonymous mutations, significant effects may occur. NPR1 is a well studied protein in *A. thaliana*, being a critical signalling protein for the systemic acquired resistance (SAR), which is involved in the response to several pathogens, such as *Phytophthora* sp., and in the cross-communication between all of the plant defence pathways (Pieterse and Van Loon 2004). Therefore, changes in this protein are expected to have an impact in the plant capacity to respond to pathogens. The domain in which the mutations were found in cork oak orthologous *NPR1* is responsible for the interaction with TGA transcription factors and consequently their activation. This interaction enhances the DNA binding activity of TGA factors to the promoter elements of *Pathogenesis-related* (*PR*) genes and is, therefore, thought to be critical for defence gene activation. In this way, natural variation within ANK 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. In Caldwell and Michelmore (2009), evidences of balancing selection were also revealed in ANK domain in *A. thaliana NPR1*, as reported here for the cork oak orthologous *NPR1*, suggesting that this gene may be under balancing selection in several plant species.

Polymorphism can be maintained in populations through several evolutionary scenarios, such as heterozygote advantage, fluctuating environment conditions and inverse frequency-dependent selection. No deviations from Hardy-Weinberg equilibrium were found in the data, excluding the heterozygote advantage hypothesis. As selective pressure acting on this gene is probably maintained by pathogens, both fluctuating environment conditions or inverse frequency-dependent selection might be the cause of polymorphism maintenance observed in *NPR1*gene in cork oak populations.

No environmental associations were detected for this fragment, as it should be expected, since variation is probably more dependent on biotic interactions than abiotic factors.

*NPR1* haplotypes can have different effects that need to be further investigated. Different haplotypes may induce the expression of different *PR* genes, as they may have different affinities with different TGA transcription factors, which might allow a specialized defence against specific pathogens. On the other hand, pathogen effector proteins may interfere with defence pathways signalling proteins directly, exerting direct selective pressure in *NPR1* (Caldwell and Michelmore 2009). Furthermore, the mutations in ANK domain may have some effect in the cross-talk between SAR and the other defence pathways, although it is not known which *NPR1* domain is involved in this cross-talk (Pieterse and Van Loon 2004).

*ARF16* (*Auxin Response Factor*) is the second gene fragment showing signs of balancing selection in cork oak. This gene codes for a transcription factor involved in root cap cell differentiation and is regulated by the microRNA mir160 (Axtell and Bartel 2005, Ding and Friml 2010). In *Quercus robur* L., this gene has been identified as a candidate gene involved in drought resistance (Spiess *et al.* 2012). As in *NPR1*, Tajima's *D* and Fu's *Fs* were significantly positive for *ARF16* sequences, which suggest that this fragment is also probably under balancing selection. Furthermore, no genetic structure was found, supporting the idea of polymorphism maintenance among cork oak populations. No environmental associations were detected for this gene, probably because this fragment is under balancing selection.

Three nonsynonymous mutations were found in the non-conserved domain of the *ARF16* gene, which is essential for its transcription factor function. Therefore, these mutations may have an effect on transcription regulation of the genes activated by ARF16 protein.

As mentioned before, there are several types of balancing selection. As no deviations from Hardy-Weinberg equilibrium were found for *ARF16*, it is unlikely that heterozygote advantage is the type of balancing selection acting on this gene. ARF16 protein function is not well characterized yet, although in Spiess et al. (2012) *ARF16* gene expression was associated with response to drought. In this way, drought might be the selective pressure acting upon this gene, maintaining polymorphism due to fluctuating environment conditions. The results here obtained are in accordance with other studies in which drought is pointed out as an important selective agent in the modulation of cork oak adaptive genetic variation (Ramirez-Valiente *et al.* 2009a, 2009b, 2011). However, definite conclusions cannot be taken until ARF16 protein function is better studied.

### 2.5.4. Putative positively selected fragment

*HSP* encodes for a small Heat Shock Protein class I, which are generally involved in response to stress. The amplified fragment corresponds to the alpha crystalline domain (ACD) and contains 27 conserved residues common to *A. thaliana* HSP15.7, which is involved in heat and oxidative stress responses (Ma *et al.* 2006). The *HSP* fragment was found to be in Hardy-Weinberg disequilibrium at the haplotype level in two positions at the locus by locus level. This may indicate that this gene is under selection and specifically that positions four and eight of this fragment may have heterozygote advantage. As the mutation at position four corresponds to a nonsynonymous and non-conservative mutation, it is plausible that it is under selection. However, neither both Tajima's D and Fu's *Fs* results were significant for this fragment nor PAML analysis was significant for those positions. Other reasons can account for deviation from Hardy-Weinberg equilibrium, such as genetic drift, migration or sampling effects.

Although the haplotype network for *HSP* is more complex than the ones estimated for the other fragments, no structure was evident either. Most of the variation between haplotypes is due to SNPs in the non-coding region and synonymous mutations. Only three nonsynonymous (and non-conservative) mutations were detected in the ACD domain. Mutations in this domain have been shown to have great impact in HSP protein

structure and its chaperone function (MacRae 2000). This is indicative that these mutations can have great importance and be under the effect of selection.

The analysis with PAML revealed that one of the three nonsynonymous mutations may be under positive selection. This same position was associated with precipitation in September and latitude by MatSAM analysis, correlating positively with aspartate frequency. In fact, precipitation in September increases with increasing latitude. Therefore, it seems that precipitation in September is the variable that affects directly the amino acid frequency in this position and latitude is correlated with this variable. As *HSP* encodes for a protein involved in response to stress, drought is expected to have an impact in this gene's variability distribution. Therefore, precipitation in September is likely to influence *HSP* adaptive variation. September follows a period during which most of the populations are submitted to a certain degree of drought (July and August). It seems plausible that precipitation in this month may be important for the recovery from the drought period and that different alleles may confer varying degrees of long-term drought resistance. In this way, individuals with aspartate in the amino acid position detected by both PAML and MatSAM are probably maladapted in southern populations, characterized by longer drought periods and low levels of precipitation in September. In Ramírez-Valliente *et al.* (2009a) northern cork oak populations were shown to be poorly adapted to drought, through the analysis of population divergence in quantitative traits in a common garden experiment, which is concordant with the results here reported.

The fact that the same position was detected by two different approaches, despite the low significance values, indicates that it is probably under positive selection. These significance values can be explained by low selective pressure.

### 2.5.5. Fragments that revealed no selection signals

*RAN3* encodes a Ras-related nuclear protein involved in mRNA export from the nucleus and protein import into the nucleus and may also be involved in cell cycle progression (Haizel *et al.* 1997, Meier and Brkljacic 2010). Two positions were found to be in Hardy-Weinberg disequilibrium, by a deficit of heterozygotes. This may indicate that those positions are under selection, although other reasons that seem more plausible in

this case can account for a deviation from Hardy-Weinberg equilibrium, such as genetic drift, migration or sampling effects (Wahlund effect). None of the neutrality tests were significant and no associations were found with spatial or environmental variables. In this gene, only synonymous mutations were found in the coding region and additional mutations were found in the non-coding region. These types of mutations are expected to be neutral, so these results are not surprising. RAN3 is a protein with a crucial function in nucleocytoplasmic transport, so it is probably under strong purifying selection, as any change in its structure could disrupt its function.

*PR1* (*Pathogenesis-related gene 1*) encodes for a protein involved in salicylic acid-mediated pathogen defence and it is induced by salicylic acid defence pathway, in which *NPR1* is a key signalling protein. Its biological function is still unclear, but there is substantial evidence for PR1 acting as a defence protein in plant–pathogen interactions and several studies demonstrated that antifungal activity is associated with purified PR1 proteins (Niderman *et al.* 1995, Rauscher *et al.* 1999). *PR1* fragment also did not present any significant statistical selection signatures. PAML detected three positions that might be under positive selection (not significant values). However, PAML has been described as prone to false positive detection (Hughes 2007), so these results should be seen as merely indicative for future studies, as no conclusions can be draw from them. Since the PR1 protein is involved in defence against pathogens, it is possible that it is under some kind of selective pressure. However, this pressure may be weak, thus justifying the non-significant results, be stronger in another region of the *PR1* gene or may be acting at a regulatory level.

### 2.5.6. Neutrality tests and environmental correlations

The different approaches used in this study suggest that different genes may be under selection. This is probably associated to the specificities of each of the statistical methods used. Site-frequency based methods, such as Tajima's *D* and Fu's *Fs*, try to identify patterns consistent with positive selection. These tests can only detect single, recent selective sweeps. Therefore, if selective pressure is low and does not produce a sweep pattern or if the selective sweep is historical and the typical pattern begins to be less clear, these tests will fail to detect positive selection. Moreover, this type of tests is

prone to be affected by demography, although it does not seem to be the case in this study as no genetic structure was detected (Jensen *et al.* 2007).

Unlike the two statistical methods mentioned, divergence-based methods, as implemented in PAML, are site specific, which is an advantage, allowing to detect specifically which positions are under positive selection (Jensen *et al.* 2007). However, these methods are based on phylogeny and assume that the inferred phylogeny is completely accurate when phylogenies are difficult to reconstruct under a selection scenario. Moreover, a $dN>dS$ pattern can be achieved by chance and codons identified as being under positive selection can in this way be false positives. Therefore, small $dN>dS$ values should be interpreted with special caution (Hughes 2007).

None of these selection detecting methods provide insights into selection drivers. Methods based on correlation between genetic and environmental variables are of great interest, as they allow the identification of loci under selection and the establishment of hypotheses about the ecological factors that may be driving the species adaptation (Joost *et al.* 2008). However, correlations have also limitations, as it may be difficult to find the environmental factors that are relevant for each species adaptation. Furthermore, environmental variables can be influenced by each other, which can lead to false associations of environmental variables with a certain allele frequency when in fact the relevant environmental variable is simply correlated to the detected one.

Therefore, it is important to adopt a pluralistic approach, since neutrality tests and environmental associations complement each other by looking at different evolutionary scales and types of selection.

## 2.6. Conclusion

In this study, different selection signatures were detected for three of the analyzed gene fragments, which support the idea that studying SNPs in functional genes can be a good approach for detecting natural selection. The findings here reported are relevant to understand cork oak adaptation to conditioning biotic and abiotic factors, as one gene involved in plant defence response to pathogens, *NPR1*, and two genes putatively implicated in drought response, *ARF16* and *HSP*, were detected as being under selective

pressure. However, further studies are needed to draw more definitive conclusions about the adaptive value of the mutations identified in these genes. Therefore, it would be relevant to further investigate the functional and ecological implications of the detected variation. This knowledge is essential to delineate sustainable management practices and conservation strategies for cork oak.

## 2.6. References


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

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 Physiology* **25**, 1085-1090.

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**, 1309-1321.

Axtell MJ, Bartel DP (2005) Antiquity of microRNAs and their targets in land plants. *Plant Cell* **17**, 1658-1673.

Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular biology and evolution* **16**, 37-48.

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. *Plant Journal* **51**, 910-918.

Biswas S, Akey JM (2006) Genomic insights into positive selection. *Trends in Genetics* **22**, 437-446.

Brasier CM (1996) *Phytophthora cinnamomi* and oak decline in southern Europe. Environmental constraints including climate change. *Annales Des Sciences Forestieres* **53**, 347-358.

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

Campos P, Aronson J (2009) Economic Analysis. In: *Cork Oak Woodlands on the Edge* (eds. Aronson J, Pereira JS, Pausas JG), pp. 127-128. Island Press, Washington DC, USA.

Coelho AC, Horta M, Neves D, Cravador A (2006) Involvement of a cinnamyl alcohol dehydrogenase of *Quercus suber* in the defence response to infection by *Phytophthora cinnamomi*. *Physiological and Molecular Plant Pathology* **69**, 62-72.

Coles ND, Coleman CE, Christensen SA, Jellen EN, Stevens MR, Bonifacio A, Rojas-Beltran JA, Fairbanks DJ, Maughan PJ (2005) Development and use of an expressed sequenced tag library in quinoa (*Chenopodium quinoa* Willd.) for the discovery of single nucleotide polymorphisms. *Plant Science* **168**, 439-447.

Costa J (2011) *Differentiation and genetic variability in cork oak populations (Quercus suber L.).* Msc thesis, Faculdade de Ciências da Universidade de Lisboa.

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 Proceedings* **5**, (Suppl 7), 13.

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

Ding Z, Friml J (2010) Auxin regulates distal stem cell differentiation in *Arabidopsis* roots. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 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. *Molecular Ecology Resources* **10**, 564-567.

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

Gandour M, Khouja ML, Toumi L, Triki S (2007) Morphological evaluation of cork oak (*Quercus suber*): Mediterranean provenance variability in Tunisia. *Annals of Forest Science* **64**, 549-555.

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

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 Journal* **11**, 93-103.

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

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

Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* **99**, 364-373.

Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**, 143.

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, Switzerland.

Jensen JD, Wong A, Aquadro CF (2007) Approaches for identifying targets of positive selection. *Trends in Genetics* **23**, 568-577.

Jimenez P, de Heredia UL, Collada C, Lorenzo Z, Gil L (2004) High variability of chloroplast DNA in three Mediterranean evergreen oaks indicates complex evolutionary history. *Heredity* **93**, 510-515.

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. *Molecular Ecology* **16**, 3955-3969.

Joost S, Kalbermatten M, Bonin A (2008) Spatial analysis method (SAM): a software tool combining molecular and environmental data to identify candidate loci for selection. *Molecular Ecology Resources* **8**, 957-960.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology* **12**, 118-123.

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

Long JC (1986) The allelic correlation structure of Gainj and Kalam speaking people. I. The estimation and interpretation of Wright's F-statistics. *Genetics* **112**, 629-647.

Lopez-de-Heredia U, Carrion JS, Jimenez 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**, 1505-1517.

Lumaret R, Tryphon-Dionnet M, Michaud H, Sanuy A, Ipotesi E, Born C, Mir C (2005) Phylogeographical variation of chloroplast DNA in cork oak (*Quercus suber*). *Annals of Botany* **96**, 853-861.

Ma CL, Haslbeck M, Babujee L, Jahn O, Reumann S (2006) Identification and characterization of a stress-inducible and a constitutive small heat-shock protein targeted to the matrix of plant peroxisomes. *Plant Physiology* **141**, 47-60.

MacRae TH (2000) Structure and function of small heat shock/alpha-crystallin proteins: established concepts and emerging ideas. *Cellular and Molecular Life Sciences* **57**, 899-913.

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

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu PG, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

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

Meier I, Brkljacic J (2010) The *Arabidopsis* nuclear pore and nuclear envelope. *The Arabidopsis book / American Society of Plant Biologists* **8**, 139.

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 Physiology* **108**, 17-27.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics* **39**, 197-218.

Pausas JG, Pereira JS, Aronson J (2009) The tree. In: *Cork Oak Woodlands on the Edge* (eds. Aronson J, Pereira JS, Pausas JG), pp. 11-21. Island Press, Washington DC, USA.

Pieterse CM, Van Loon L (2004) NPR1: the spider in the web of induced resistance signaling pathways. *Current Opinion in Plant Biology* **7**, 456-464.

Pina-Martins F, Paulo OS (2008) CONCATENATOR: sequence data matrices handling made easy. *Molecular Ecology Resources* **8**, 1254-1255.

Quartau JA, Mathias ML (2010) Insects of the understorey in Western Mediterranean forest landscapes: a rich biodiversity under threat. In: *Insect Habitats: Characteristics, Diversity and Management* (eds. Harris EL, Davies NE), pp. 133-142. Nova Science Publishers, Hauppauge NY, USA.

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 Genetics & Genomes* **7**, 285-295.

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. *Molecular Ecology* **18**, 3803-3815.

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. *Evolutionary Ecology* **24**, 1031-1044.

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.). *Forest Ecology and Management* **257**, 1676-1683.

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 Journal* **19**, 625-633.

Serrasolses I, Pérez-Devesa M, Vilagrosa A, Pausas JG, Sauras T, Cortina J, Vallejo VR (2009) Soil properties constraining cork oak distribution. In: *Cork Oak*

*Woodlands on the Edge* (eds. Aronson J, Pereira JS, Pausas JG), pp. 89-99. Island Press, Washington DC, USA.

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**, 236-252.

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**, 601-607.

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. *Environmental and Experimental Botany* **77**, 117-126.

Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics* **76**, 449-462.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978-989.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595.

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 Research* **22**, 4673-4680.

Toumi L, Lumaret R (1998) Allozyme variation in cork oak (*Quercus suber* L.): the role of phylogeography and genetic introgression by other Mediterranean oak species and human activities. *Theoretical and Applied Genetics* **97**, 647-656.

Valladares F, Gianoli E, Gomez JM (2007) Ecological limits to plant phenotypic plasticity. *New Phytologist* **176**, 749-763.

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

Wiedmann RT, Smith TPL, Nonneman DJ (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genetics* **9**, 81.

Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**, 1586-1591.

Yang ZH, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular biology and evolution* **22**, 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. *Molecular biology and evolution* **26**, 273-283.

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. *Genetics* **163**, 1123-1134.

# CHAPTER 3

# Final Remarks and Future Prospects

## 3.1. Final remarks and future prospects

This work emphasizes the importance of whole genome scans when assessing genomic variability in non-model species, for which no genomic resources are available, as highlighted in previous studies (e.g. Namroud *et al.* 2008, Poncet *et al.* 2010). More recently, NGS methods have greatly impacted genomics research, as these techniques allow experiments that were previously impossible or economically unviable. 454 pyrosequencing, the NGS method in which the present work is based, is particularly useful when no reference genome is available, since this NGS technique produces the longest reads (Glenn 2011), facilitating downstream work. Furthermore, 454 pyrosequencing has been successfully used to discover and develop molecular markers, such as SNPs and SSRs (e.g. Grattapaglia *et al.* 2011, Lai *et al.* 2012).

In this work, the utility of 454 transcriptome sequencing as a tool to identify useful and putatively functional markers in a non-model species was confirmed. In spite of the problems in amplifying genomic DNA with primers designed from ESTs data, a good percentage of amplification success was obtained comparing to previous studies (e.g. Zhu *et al.* 2003, Coles *et al.* 2005). Furthermore, regardless the technical errors that are known to occur in 454 pyrosequencing (Margulies *et al.* 2005, Huse *et al.* 2007), a high percentage of the putative SNPs tested (79%) was successfully validated. Mutations within coding regions are in general particularly insightful, as they can affect the amino acid composition of the protein and therefore its structure and/or function, which means that the discovery of SNP markers in transcriptome sequences can facilitate the identification of genes involved in several processes, particularly in adaptive change (Renaut *et al.* 2010, Zhou *et al.* 2012). Accordingly, in the present study 454 transcriptome sequencing allowed for the development of useful markers to investigate cork oak adaptive genetic variation.

A variety of selection signatures was detected in some of the gene fragments here studied. After detecting signals of selection, biochemical, physiological and ecological studies should be carried out to test if the detected mutations are in fact adaptive and confer fitness advantages (Hughes 2007). For instance, for *ARF16* it would be interesting to assess if plants with different genotypes present differences in development and differentiation of root cap. Moreover, since it was previously identified as a candidate gene for drought resistance in *Quercus robur* (Homolka *et al.*

2012, Spiess *et al.* 2012), it should be of particular interest to carry out drought stress assays and analyze if different genotypes show different levels of tolerance or drought recovery. For *HSP* gene, drought stress assays should also be performed, namely long-term drought resistance tests, to assess if plants with different amino acids in the position associated with precipitation in September during this study have different responses towards drought conditions. Differences in drought resistance should be also tested for *HSP* complete genotypes. Since cork oak stands have been facing a significant decline much associated with severe drought periods (Toumi and Lumaret 1998, Soto *et al.* 2007), and in the face of the major climate changes expected to occur in the Mediterranean Basin, with more severe and long drought periods (IPCC 2007), studying these two genes would be of major interest.

Several pests and diseases have also been implicated in cork oak decline, especially the fungus *Phytophthora cinnamomi* (Brasier 1996, Cabral and Ferreira 1999). The hereby reported findings suggesting that *NPR1* is under balancing selection may help to better understand this species' defence responses to several pathogens, including *P. cinnamomi*, and host-pathogen interactions. For this gene, it should be investigated if different alleles lead to different defence response intensities or if individuals with different genotypes have different levels of resistance to pathogens that commonly attack cork oak. A possible involvement in the response to herbivory is not to exclude either, as *NPR1* gene is implicated in the cross-talk between defence responses to pathogens (salicylic acid dependent signalling pathway) and defence responses to wounds inflicted by herbivorous insects (jasmonic acid dependent signalling pathway) (Pieterse and Van Loon 2004). Although no significant selection signal was detected for *PR1*, it would also be interesting to study this gene and its putative role in response to *P. cinnamomi* (and other pathogens) infections, as PR1 protein has been associated with antifungal activity (Niderman *et al.* 1995, Rauscher *et al.* 1999). Assessing if different *PR1* alleles confer differential pathogen resistance could enlighten its functional and adaptive role in cork oak.

Furthermore, expression studies should be carried out for all genes, since detecting different levels of fitness in the suggested assays do not necessarily mean that these are caused by differences in genotypes. It should therefore be ruled out that variations in gene expression levels are the cause of the possibly detected differences (Hughes 2007).

Finally, it would be interesting to continue the SNP validation process and subsequent exploitation of these markers, since useful, putatively functional markers that may be involved in cork oak adaptation can be developed in this way.

In conclusion, in the present work important steps were taken towards gaining new knowledge concerning adaptive genetic variation in cork oak, which is of major importance for the definition of management and conservation strategies for this relevant species.

## 3.2. References


Brasier CM (1996) *Phytophthora cinnamomi* and oak decline in southern Europe. Environmental constraints including climate change. *Annales Des Sciences Forestieres* **53**, 347-358.

Cabral MT, Ferreira MC (1999) *Pragas dos Montados*. Estação Florestal Nacional, Lisbon, Portugal.

Coles ND, Coleman CE, Christensen SA, Jellen EN, Stevens MR, Bonifacio A, Rojas-Beltran JA, Fairbanks DJ, Maughan PJ (2005) Development and use of an expressed sequenced tag library in quinoa (*Chenopodium quinoa* Willd.) for the discovery of single nucleotide polymorphisms. *Plant Science* **168**, 439-447.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11**, 759-769.

Grattapaglia D, Silva-Junior OB, Kirst M, de Lima BM, Faria DA, Pappas GJ, Jr. (2011) High-throughput SNP genotyping in the highly heterozygous genome of *Eucalyptus*: assay success, polymorphism and transferability across species. *BMC Plant Biology* **11**, 65.

Homolka A, Eder T, Kopecky D, Berenyi M, Burg K, Fluch S (2012) Allele discovery of ten candidate drought-response genes in Austrian oak using a systematically informatics approach based on 454 amplicon sequencing. *BMC research notes* **5**, 175-175.

Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* **99**, 364-373.

Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**, 143.

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, Switzerland.

Lai K, Duran C, Berkman PJ, Lorenc MT, Stiller J, Manoli S, Hayden MJ, Forrest KL, Fleury D, Baumann U, Zander M, Mason AS, Batley J, Edwards D (2012) Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnology Journal* **10**, 743-749.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu PG, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

Namroud M-C, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology* **17**, 3599-3613.

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 Physiology* **108**, 17-27.

Pieterse CM, Van Loon L (2004) NPR1: the spider in the web of induced resistance signaling pathways. *Current Opinion in Plant Biology* **7**, 456-464.

Poncet BN, Herrmann D, Gugerli F, Taberlet P, Holderegger R, Gielly L, Rioux D, Thuiller W, Aubert S, Manel S (2010) Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Molecular Ecology* **19**, 2896-2907.

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 Journal* **19**, 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). *Molecular Ecology* **19**, 115-131.

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**, 601-607.

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. *Environmental and Experimental Botany* **77**, 117-126.

Toumi L, Lumaret R (1998) Allozyme variation in cork oak (*Quercus suber* L.): the role of phylogeography and genetic introgression by other Mediterranean oak species and human activities. *Theoretical and Applied Genetics* **97**, 647-656.

Zhou Y, Gao F, Liu R, Feng J, Li H (2012) De novo sequencing and analysis of root transcriptome using 454 pyrosequencing to discover putative genes associated with drought tolerance in *Ammopiptanthus mongolicus*. *BMC Genomics* **13**, 266.

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. *Genetics* **163**, 1123-1134.

# APPENDIX

# 4.1. Supporting Information

**Table S1.** SNPs for which the corresponding DNA fragment was successfully amplified, primers (5'-3') used for their amplification, annealing temperatures used (Ta) and length (bp) of the amplified fragments.

| SNP | Putative gene | Primer F | Primer R | Ta (ºC) | Length |
|---|---|---|---|---|---|
| SNP1 | *SQD1* | TCCAATGAACCATCCAAGCC | TAGAGTAAGGAGCAGACCGC | 59 | 336 |
| SNP4 | *NPR1* | ACAGAGCTCCTTGATCTTGC | GAGATCATCACCTGCCATAGC | 53 | 270 |
| SNP6 | *RAN3* | TATCTTGCCAGGAAGCTTGC | GGTCTATGGTCAATAGCCGAC | 53 | 627 |
| SNP7 | *ATP binding protein* | GCAGAAGGAGAGGATGTTACTG | GAACCCAGGTACTCTGTTACG | 53 | 500 |
| SNP8 | *Annexin 8* | GGTCATCAATGCCTCGAATGG | TTGGCCTAGTGACTACTTACCG | 54 | 500 |
| SNP13 | *ATPase* | CAGCTTGTAGGAAAGGATGC | CATCCATACCAGCTCCTCTC | 65 | 240 |
| SNP15 | *PR1* | CAACCGATGAATGTGCCTCC | TGGACCTATAACATGGGACGC | 64 | 257 |
| SNP17 | *MYC2* | GCTTAACCAGAGGTTCTACG | CATCCCATCCGATTATCTTCAC | 56 | 300 |
| SNP20 | *HSP* | GTGTTCAAAGCTGATCTTCC | ACCTTCTGACAAGTAAACCC | 56 | 374 |
| SNP22 | *GLN* | GCCCTTCTGTTGGTATATCTGC | GTTTCATGTCGGCCAGTGAG | 62 | 600 |
| SNP25 | *Triacylglycerol lipase* | ATGCTCCAGATCTTGTTCCT | TCCTTCAAACCATCCATTGTC | 56 | 237 |
| SNP28 | *NFYA7* | GCATGAATCTCGACATTTGC | GAGAAATCCATCGGAGAAGC | 58 | 397 |
| SNP29 | *ARF16* | GAATATCTTCAGAAGATCTCCACC | CATTTAGAAATCTGCTCCTCAGTG | 65 | 234 |
| SNP30 | *Transcription factor bHLH51* | ATGTGAAGGATCTCAAGCGA | CCCAACACTTGCTATGTCAG | 63 | 256 |

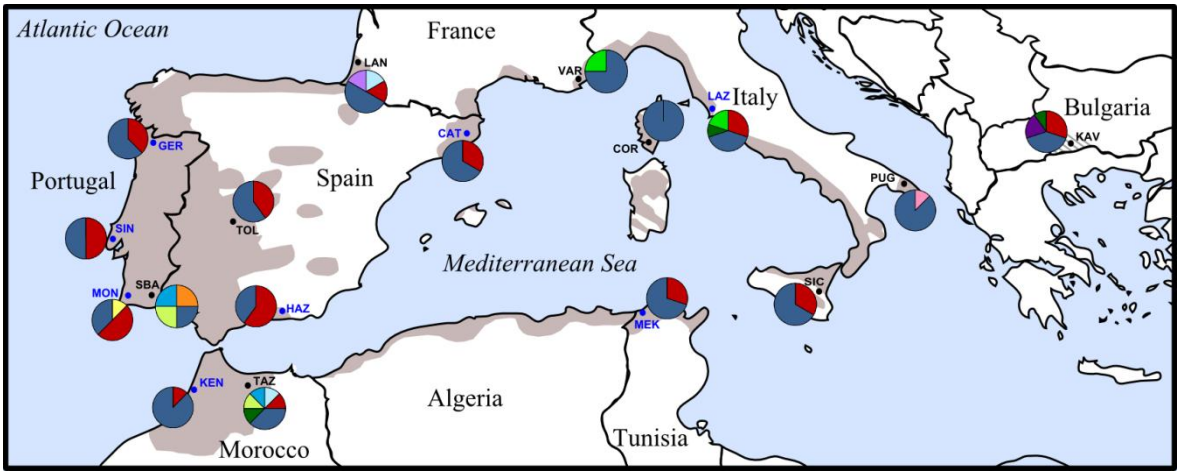
**Table S2.** Species, populations and number of individuals amplified for each fragment for the population genetics study.

| Species | Populations | Gene | | | | |
|---|---|---|---|---|---|---|
| | | *RAN3* | *NPR1* | *PR1* | *ARF16* | *HSP* |
| *Q. suber* | Sintra (SIN) | 5 | 6 | 6 | 6 | 5 |
| | Monchique (MON) | 4 | 6 | 6 | 6 | 6 |
| | Gerês (GER) | 4 | 5 | - | - | - |
| | S. Bras de Alportel (SBA) | 2 | 5 | - | - | - |
| | Lazio, Toscany (LAZ) | 5 | 6 | 6 | 6 | 6 |
| | Puglia, Brindisi (PUG) | 4 | 6 | 5 | 6 | 6 |
| | Sicilia, Catania (SIC) | 3 | 6 | 6 | 6 | 6 |
| | Sardegna, Cagliari (SAR) | - | 5 | 6 | 6 | 6 |
| | Var, Bomes les Mimoses (VAR) | 2 | 6 | 6 | 6 | 6 |
| | Landes, Soustons (LAN) | 3 | 6 | 6 | 6 | 6 |
| | Corse, Sartene (COR) | 2 | 6 | 6 | 6 | 6 |
| | Montes de Toledo, Cañamero (TOL) | 5 | 6 | 6 | 6 | 6 |
| | Cataluña, Sta Coloma Farnes (CAT) | 3 | 6 | 6 | 6 | 6 |
| | Haza de Lino (HAZ) | 5 | 6 | 6 | 6 | 6 |
| | Kenitra, Ain Johra (KEN) | 4 | 6 | 6 | 6 | 5 |
| | Taza, Bab Azhar (TAZ) | 4 | 6 | 6 | 6 | 6 |
| | Mekna, Tabarka (MEK) | 5 | 6 | 5 | 6 | 6 |
| | Kavrakirovo (KAV) | 5 | 6 | 6 | 6 | 6 |
| | Guerbès (ARG) | - | 5 | - | - | - |
| | **Total** | 65 | 110 | 94 | 96 | 94 |
| *Q. rotundifolia* | - | 2 | 3 | 3 | 3 | 3 |
| *Q. cerris* | - | - | 1 | 1 | 1 | 1 |

**Table S3.** Hardy-Weinberg equilibrium analysis, both locus by locus and at the haplotype level, for each of the studied fragments. Significant values are at bold.

| Fragment | SNP position | Obs. Het. | Exp. Het. | *P*-value |
|---|---|---|---|---|
| *RAN3* | 1 | 0.38462 | 0.44615 | 0.27514 |
| | 2 | 0.38462 | 0.43518 | 0.39300 |
| | 3 | 0.38462 | 0.43518 | 0.39393 |
| | 4 | 0.12308 | 0.11640 | 1.00000 |
| | 5 | 0.36923 | 0.42934 | 0.37965 |
| | 6 | 0.38462 | 0.43518 | 0.39413 |
| | 7 | 0.44615 | 0.47335 | 0.79132 |
| | 8 | 0.38462 | 0.43518 | 0.39350 |
| | **9** | **0.00000** | **0.06011** | **0.00019** |
| | 10 | 0.03077 | 0.03053 | 1.00000 |
| | 11 | 0.04615 | 0.04544 | 1.00000 |
| | 12 | 0.35385 | 0.43518 | 0.15490 |
| | **13** | **0.32308** | **0.44615** | **0.04671** |
| | haplotype | - | - | 0.17565 |
| *NPR1* | 1 | 0.37273 | 0.43250 | 0.18409 |
| | 2 | 0.37273 | 0.43250 | 0.18320 |
| | 3 | 0.47273 | 0.49166 | 0.70087 |
| | 4 | 0.47273 | 0.49166 | 0.70171 |
| | 5 | 0.00909 | 0.00909 | 1.00000 |
| | haplotype | - | - | 0.31341 |
| *PR1* | 1 | 0.21277 | 0.20776 | 1.00000 |
| | 2 | 0.06383 | 0.06212 | 1.00000 |
| | 3 | 0.18085 | 0.18267 | 1.00000 |
| | 4 | 0.45745 | 0.44072 | 0.81513 |
| | 5 | 0.45745 | 0.44072 | 0.81381 |
| | haplotype | - | - | 0.72841 |
| *ARF16* | 1 | 0.46875 | 0.46853 | 1.00000 |
| | 2 | 0.36458 | 0.41094 | 0.31803 |
| | 3 | 0.23958 | 0.22769 | 1.00000 |
| | 4 | 0.45833 | 0.46575 | 1.00000 |
| | haplotype | - | - | 0.35428 |
| *HSP* | 1 | 0.01064 | 0.01064 | 1.00000 |
| | 2 | 0.27660 | 0.28399 | 0.72583 |
| | 3 | 0.31915 | 0.26966 | 0.11651 |
| | **4** | **0.74468** | **0.46991** | **0.00000** |
| | 5 | 0.07447 | 0.07208 | 1.00000 |
| | 6 | 0.07447 | 0.07208 | 1.00000 |
| | 7 | 0.19149 | 0.17408 | 1.00000 |
| | **8** | **0.84043** | **0.48987** | **0.00000** |
| | 9 | 0.19149 | 0.17408 | 1.00000 |
| | 10 | 0.32979 | 0.29099 | 0.28852 |
| | **haplotype** | - | - | **0.00000** |

**a)**




**b)**

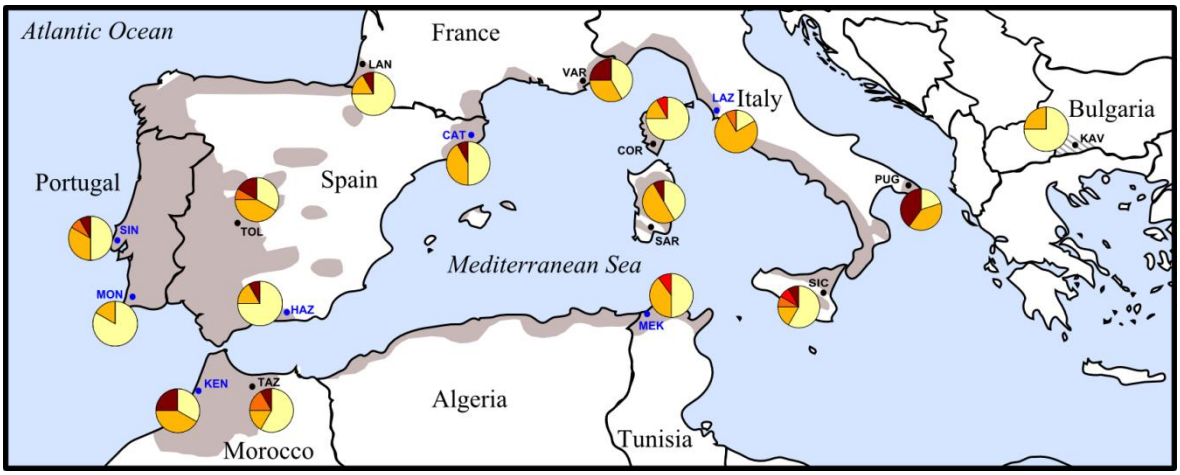



**Figure S1.** Haplotype geographical distribution of *RAN3* (a) and *PR1* (b) fragments. Haplotype colours correspond to the ones in the networks (Figure 2.3 a and b respectively).