UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

Faculdade de Ciências

Departamento de Biologia

Analyses of the genomic variation to study cork oak evolution and adaptation: from past to future climatic changes

Doutoramento em Biologia e Ecologia das Alterações Globais
Especialidade em Biologia do Genoma e Evolução

Francisco Rente de Pina Martins

Tese orientada por:
Prof. Doutor Octávio S. Paulo
Prof. Doutor Georgios Joannis Pappas

Documento especialmente elaborado para a obtenção do grau de doutor

2017

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

Faculdade de Ciências

Departamento de Biologia

Analyses of the genomic variation to study cork oak evolution and adaptation: from past to future climatic changes

Doutoramento em Biologia e Ecologia das Alterações Globais
Especialidade em Biologia do Genoma e Evolução

Francisco Rente de Pina Martins

Tese orientada por:
Prof. Doutor Octávio S. Paulo
Prof. Doutor Georgios Joannis Pappas

Júri:

Presidente:

- Doutor Henrique Manuel Roque Nogueira Cabral, Professor Catedrático
  Faculdade de Ciências de Universidade de Lisboa

Vogais:

- Doutor Christian Rellstab, Senior Researcher / Scientific Staff Member
  Swiss Federal Research Institute WSL (Suiça);
- Doutor Daniel Vieira Noro e Silva Sobral, Investigador Principal
  Instituto Gulbenkian Ciência;
- Doutora Maria da Luz da Costa Pereira Mathias, Professora Catedrática
  Faculdade de Ciências da Universidade de Lisboa;
- Doutora Margarida Matos Demony de Carneiro Pacheco de Matos, Professora Associada com Agregação
  Faculdade de Ciências de Universidade de Lisboa;
- Doutor Octávio Fernando de Sousa Salgueiro Godinho Paulo, Professor Auxiliar
  Faculdade de Ciências de Universidade de Lisboa (orientador).

Documento especialmente elaborado para a obtenção do grau de doutor

SFRH/BD/51411/2011, from Fundação para a Ciência e Tecnologia (FCT) – Portugal.

2017

Abstract

Current scientific literature indicates that climate change will cause an average world temperature increase between 1 and 4ºC, along with changes in precipitation patterns and extreme weather events in the next 50 years. These are likely to have a negative impact for biodiversity in general, and forest ecosystems should be particularly affected, especially those in Mediterranean areas, like the cork oak (*Quercus suber* L.) "montados". In order to understand how species can respond to such alterations, it is important to know their evolutionary history and genetic architecture of adaptive traits. Advances in sequencing technologies have relatively recently brought down the cost of sequencing per base pair to a point where even small research facilities can obtain genomic information of non-model organisms. These advances made SNP markers become the most abundant type of genetic variation in eukaryotic genomes, especially with the advent of Reduced Representation libraries such as RAD-Seq and GBS. Yet, despite their widespread use, SNP data analyses still bore its own set of bioinformatics challenges. While most of these are related with the practical aspects of the process, such as being able to handle very large datasets, or discriminate between neutral and non-neutral markers, some fundamental problems, like reproducibility are also important issues affecting research in this area.

In this thesis, genomic and transcriptomic data from *Q. suber* was used to assess the evolutionary history of the species, detect the effects of natural selection across the cork oak's distribution range and find any associations between the obtained markers and environmental variables.

The main methodological contributions of this thesis are in the form of three software suites: (1) *4Pipe4*, a software for automatically mining SNP markers from NGS data when no reference genome nor strain information is present, (2) *NCBI Mass Sequence Downloader*, a program to automate the downloading of large datasets from the NCBI databases, and (3) *Structure_threader*, a software to automate and parallelize analyses using several popular clustering analyses programs. All of these programs were developed with the intent to improve the automation and reproducibility value of the analysis processes they are meant to be part of.

The main findings of this thesis are that (1) the evolutionary history and population structure of *Q. suber* is not as neatly structured as chloroplastidial markers indicate, (2) local adaptation plays and important role in the distribution of the species' genetic variability, and (3) the cork oak may be better equipped, from a genetic point of view, to adapt to climate change than what previous studies based solely on ecological modelling indicated.

Keywords: Bioinformatics, Evolutionary Biology, Genotyping by Sequencing, Natural Selection, Software Development

Resumo

A esmagadora maioria da literatura científica atual indica que durante os próximos 50 anos, a temperatura mundial irá aumentar em média entre 1 e 4ºC dependendo do local e modelo utilizado para a previsão. A acompanhar estas alterações de temperatura preveem-se também alterações aos padrões e frequência de precipitação e eventos climáticos extremos (como furacões, secas extremas, chuvas torrenciais). Os efeitos destas mudanças na biodiversidade são ainda desconhecidos, mas pesquisas recentes indicam que terão um impacto geralmente negativo. Espera-se que este impacto seja particularmente sentido em sistemas florestais devido a fatores como a longevidade das árvores, que dificultam a resposta a alterações rápidas como as que estamos a vivenciar. As florestas mediterrâneas serão possivelmente as mais afetadas devido à sua sensibilidade aos ciclos de precipitação. Tentar prever a reposta das espécies a alterações ambientais exige conhecimento da história evolutiva das mesmas, bem como da arquitetura genética das respetivas características adaptativas. Uma espécie emblemática que aparece frequentemente associada a estes sistemas florestais mediterrânicos é o sobreiro (*Quercus suber* L.). Esta árvore da família Fagacea é caracterizada pelo casca do seu tronco, composta por um tipo de suberina também designada por "cortiça". A sua distribuição estende-se por diversos países da margem Norte do Mediterrâneo (Portugal, Espanha, França e Itália) e também em países da margem Sul do mesmo (Marrocos, Argélia e Tunísia) *Q. suber* aparece ainda na costa Atlântica de alguns países Europeus (Portugal, Espanha e França). No locais onde ocorre forma tipicamente um sistema denominado *montado* (ou *dehesa* em castelhano) frequentemente em associação com outras árvores dos géneros *Quercus* e *Pinus*.

A história evolutiva do sobreiro tem sido alvo de diversos estudos, baseados em diversas técnicas de análise. No entanto, trabalhos sobre esta espécie que recorrem a uma amostragem representativa de todo o espetro de distribuição da mesma têm sido efetuados com recurso a marcadores cloroplastidiais e a principal conclusão destes é que *Q. suber* é uma espécie essencialmente dividida em quatro linhagens distintas, geograficamente segregadas. Apesar disto, estudos efetuados com marcadores nucleares, de abrangência geográfica menor sugerem um baixo nível de diferenciação populacional, altos índices de diversidade e estrutura populacional não evidente. Relativamente à potencial resposta do sobreiro às alterações climáticas globais, estudos de modelação ecológica, tanto de âmbito localizado como generalizado foram efetuados, mas apresentando resultados contraditórios. Enquanto alguns indicam que a área de distribuição de *Q. suber* será reduzida nos próximos 70 anos, outros preveem que a mesma aumente em igual período de tempo.

Apenas com o aparecimento e democratização das tecnologias de sequenciação de segunda geração, normalmente denominadas de "Next Generation Sequencing" (NGS) se tornou possível a centros de investigação de pequenas dimensões trabalharem com dados de DNA nuclear em larga escala. Esta tendência ganhou ainda mais relevância com

o surgimento das técnicas de “Reduced Representation Libraries” (RRLs) que permitem amostrar a variação de todo o genoma de um organismo. Uma das consequências destas técnicas foi um aumento de várias ordens de grandeza na quantidade de dados gerados e que passou a ser necessário analisar pelos investigadores. Métodos que eram adequados para analisar dados pré NGS deixaram de ser possíveis de aplicar por simplesmente não serem capazes de escalar com a quantidade de informação com que agora eram obrigados a lidar. Este “novo tipo” de dados trouxe também outros novos desafios bioinformáticos associados requerendo novo software, aumentando também os padrões de qualidade que a comunidade científica passou a exigir do mesmo.

Se antes do surgimento das tecnologias NGS a bioinformática e a biologia computacional eram já disciplinas em ascensão, os problemas gerados por este novos dados deram-lhes tal impulso que levou alguns autores a sugerir que no futuro, toda a biologia será biologia computacional. Um dos grandes desafios inerentes a estas áreas da ciência prende-se com a reprodutibilidade. Apesar de estar intrinsecamente ligada a todos os campos científicos, ganha especial relevo nas análises bioinformáticas de dados, por um lado porque repetir análises computacionais é mais simples e barato que o equivalente laboratorial, e por outro porque a partilha de dados digitais é mais simples e frequente entre investigadores do que por exemplo amostras físicas, o que aumenta a visibilidade do problema. Este problema de reprodutibilidade em ciência (ou falta dela) levou inclusive que fosse decretada uma “crise de reprodutibilidade” que afeta principalmente pesquisa publicada em revistas de alto fator de impacto. Apesar de esta “crise” estar longe de ultrapassada, as práticas que a podem resolver são já conhecidas.

Três objetivos principais foram definidos para esta tese. Foram eles: (1) estudar a história evolutiva de *Q. suber*, baseado em SNPs minados de dados transcritómicos de NGS, (2) estudar a ação da seleção natural ao longo da área de distribuição do sobreiro com recurso a dados genómicos, e (3) realizar estudos de associação entre SNPs de sobreiro e um conjunto de variáveis ambientais. Todos os objetivos foram cumpridos durante a realização da tese, no entanto, no decorrer da mesma surgiu a oportunidade de tentar responder a outras questões científicas relacionadas, razão pela qual os produtos da tese não se prendem apenas com estes três objetivos.

Do ponto de vista técnico, este tese contribuiu para a ciência com três suites de software. O primeiro, denominado *4Pipe4*, é uma pipeline de análise de dados otimizada para minar automaticamente SNPs em conjuntos de dados de 454 (apesar de funcionar com outras tecnologias NGS) quando não estão disponíveis nem sequências de referência, nem informação de estirpe. À data da sua publicação, este método demonstrou uma performance superior a todos os outros métodos testados na métrica de deteção de falsos positivos. O segundo, um programa chamado *NCBI Mass Sequence Downloader*, permite automaticamente descarregar das bases de dados do NCBI grandes quantidades de dados, que tendem a ser problemáticas quando efetuadas a partir do navegador web. Este programa simplesmente disponibiliza duas interfaces (gráfica e linha de comandos) para a API do próprio site e acrescenta verificação de sequências e

correção de erros. O terceiro é um programa chamado *Structure_threader*, que permite automatizar e paralelizar software popular de estimação de estrutura populacional (STRUCTURE, FASTSTRUCTURE e *Maverick*). Além disso tem também a capacidade de desenhar os gráficos de "Q-values", que são o produto típicos destes programas, tanto numa forma estática, com qualidade de publicação, como numa versão dinâmica e interativa, ideal para a fase exploratória da análise. Finalmente, este tese contribui ainda com o processo de automatização da análise dos dados de GBS utilizados, que não sendo um programa formal, por um lado serve para garantir a reprodutibilidade da análise e por outro é suficientemente flexível para que possa ser usado com outros conjuntos de dados em análises similares. Todos estes produtos metodológicos têm em vista a automatização e a reprodutibilidade de processos analíticos.

Do ponto de vista biológico, os principais resultados desta tese são também três. Em relação à história evolutiva, os dados de GBS indicam que *Q. suber* não é uma espécie tão estruturada do ponto de vista genético como os marcadores cloroplastidiais indicam. Em relação aos efeitos da seleção natural, os resultados das análises de outliers e associações, sugerem que a adaptação local desempenha um papel importante na distribuição da variabilidade genética da espécie. Finalmente os resultados das análises de risco de não adaptação indicam que o sobreiro está provavelmente melhor equipado geneticamente para resistir às alterações climáticas do que o indicado por previsões efetuadas em estudos anteriores baseados exclusivamente em dados ecológicos.

Palavras-chave: Bioinformática, Biologia Evolutiva, Desenvolvimento de Software, *Genotyping by Sequencing*, Seleção Natural.

Acknowledgements

Professor Octávio Paulo, my main advisor, for having introduced me to both Evolutionary Biology and Bioinformatics. And for everything else along these years.

All the members (and ex-members) of the CoBiG² research group, you all had an impact on this thesis, but you are just too many to mention each and every one of you. Still, I'd like to particularly mention:

- Telma Laurentino – most of all, for making a better scientist out of me.
- Diogo Silva – for being my go-to bioinformatics peer.
- Eduardo Marabuto – for the perspectives only a field work expert can provide. And for the Statler and Waldorf moments.
- Joana Fino – for talking me into jump-starting the *Structure_threader* project.
- Bruno Vieira – for *really* introducing me to the problem of reproducibility.
- Sofia Seabra – for peer reviewing.
- *Yana* Vieira – fer helping me keep things in prespective.
- Sara Silva – for making sure the wet-lab was stocked up whenever I needed it.

Paulo Bastos – for all the assistance with FCUL's IT infrastructure.

My parents, for all they have provided.

Miguel, my brother, for always showing me a different life perspective.

Ana, my wife ♡. None of this would have been possible without your unconditional support.

And finally, Gonçalo & Beatriz, my offspring. For making me proud every single day.

Table of Contents

Author contributions

At the time of submission, three chapters (2, 3 and 4) are published in international peer reviewed journals. Two more are submitted for review. I am the largest contributor and first author in all of them. However, science is a team effort and credit should be attributed to the other co-authors. Below is a citation for each of the chapters that have co-authors, along with a description of my contribution to each of them.

Chapter 2. Pina-Martins, F., Vieira, B. M., Seabra, S. G., Batista, D., & Paulo, O. S. (2016). 4Pipe4 – A 454 data analysis pipeline for SNP detection in datasets with no reference sequence or strain information. *BMC Bioinformatics*, *17*, 41. doi:10.1186/s12859-016-0892-1.

FPM – participated in the pipeline's design and concept, developed the software, drafted and implemented the SNP detection routines and wrote the manuscript.

Chapter 3. Pina-Martins, F., & Paulo, O. S. (2016). NCBI Mass Sequence Downloader–Large dataset downloading made easy. *SoftwareX*, *5*, 80–83. doi:10.1016/j.softx.2016.04.007.

FPM – conceptualized the problem, drafted the implementation, developed the software and wrote the manuscript.

Chapter 4. Pina-Martins, F., Silva, D. N., Fino, J., & Paulo, O. S. (2017). *Structure_threader*: An improved method for automation and parallelization of programs STRUCTURE, FASTSTRUCTURE and *MavericK* on multicore CPU systems. *Molecular Ecology Resources*, n/a-n/a. doi:10.1111/1755-0998.12702

FPM – conceptualized the problem, drafted the implementation, developed the software (except the advanced plotting options), and wrote the manuscript.

Chapter 5. Pina-Martins, F., Batista, D., & Paulo, O. S. (n.d.). New SNPs mined from Cork Oak (*Quercus suber* L.) EST data provide preliminary insights on the species' potential response to climatic change. (submitted to Frontiers in Plant Science).

FPM – participated in the study's conception, performed fieldwork, performed wet-lab work, performed the entire data analyses and wrote the manuscript.

Chapter 6. Pina-Martins, F., Baptista, J., Pappas G., & Paulo, O. S. (n.d.). New insights on adaptation and population structure of Cork Oak using genotyping by sequencing. (submitted to Molecular Ecology).

FPM – participated in the study's conception, performed fieldwork, performed wet-lab work, performed the entire data analyses and wrote the manuscript.

Bibliography style

Throughout this thesis, the chosen bibliographic style is "American Psychological Association 6th edition ("doi:" DOI prefix)".

Chapter 1

General Introduction

1 A changing world

1.1 The fast pace of change

The year is 2017. Global climate change is one of the hottest topics currently being debated, both in academia and by the general public. It also happens to be a quite controversial and politicized subject. Regardless of what some factions claim in media centred debates (Walther et al. 2005), an overwhelming majority of scientific literature suggests an increase in average temperature will happen until 2100 (Walther et al. 2002; Oreskes 2004; IPCC 2014). Estimates of this increase vary with the used model and localization, but an average increase of 1.0°C - 4.0°C by 2090 (Figure 1.1), relative to 1990 values is what most current estimates point towards (IPCC 2014). Climatic changes are not just bound to temperature, as precipitation patterns are also very likely to change, along with the frequency of extreme meteorologic events (Beniston et al. 2007; IPCC 2014).

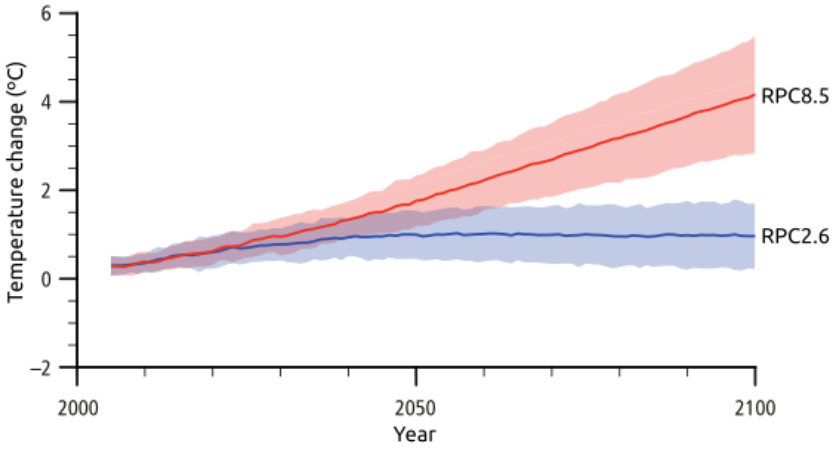

Figure 1.1: Global average surface temperature change from 2006 to 2100 as determined by multi-model simulations. All changes are relative to 1986–2005. Time series of projections and a measure of uncertainty (shading) are shown for scenarios RCP2.6 (blue) and RCP8.5 (red). Adapted from (IPCC 2014).

These changes are likely to have a negative impact for biodiversity in general (IPCC 2014), but it should be particularly felt on forest ecosystems (Alley 2003) due to the long life-span of trees, which is an obstacle to coping with rapid environmental changes (Lindner et al. 2010). This is especially true for European forest systems, which have a particularly high anthropogenic imprint in their compositions (Milad et al. 2011).

Even though the expected temperature shift may bring some benefits to forests in temperate and boreal areas due to a longer growth season (Saxe et al. 2002; Briceño-

Elizondo et al. 2006), Mediterranean forests, where precipitation is expected to shift almost exclusively to winter seasons, are more likely to experience detrimental effects (Loustau et al. 2005). Also accompanying these climatic changes, other factors, such as diseases, parasites and herbivores that afflict forest species will change with the local conditions, in ways that are currently unpredictable (Lindner et al. 2010; Lindner et al. 2014).

Species affected by the changing environment will therefore inevitably respond to the new environmental conditions. This response can be enacted by altering one or more of the following traits: (1) the phenology and physiology of organisms, (2) the range and distribution of species, (3) the composition of and interactions within communities, and (4) the structure and dynamics of ecosystems (Walther et al. 2002). In order to estimate in which way a species is capable of responding to environmental change (Kremer et al. 2012), it is of extreme importance to understand the genetic architecture of its adaptive traits (Alberto et al. 2013) and evolutionary history (Kremer et al. 2014). Such studies have been previously enacted in European forest trees such as *Pinus* (Alberto et al. 2013) or *Populus* (Olson et al. 2013) genera, where both evolutionary history and adaptive traits are assessed to help predict the potential response of these tree species to global climatic alterations. Despite their differences, distinguishing the effects in marker evolution of evolutionary history from adaptation is not an easy task, especially when migration is high or the divergent selection pressure is weak (Thibert-Plante and Hendry 2010).

1.2 The *montado* system

A typically Mediterranean forest system is the *montado* in Portuguese, or *dehesa* in Spanish (Figure 1.2), which is known to play an important role in water, soil and biodiversity conservation (Joffre et al. 1999).

Oak *dehesas* can be of either natural or artificial origin, and have been described as multi-functional systems, mainly of the foresty-pastoral type, consisting of multiple, integrated and interdependent systems and subsystems (Lauw et al. 2013). *Montados* can be composed of several *Quercus* species (*Q. suber*, *Q. rotundifolia*, *Q. faginea* and *Q. pyrenaica*), in either 'pure' or 'mixed' settings. When cork oak is the dominant species of a *montado* it can be associated with trees of the *Pinus* genus (*Pinus pinea* and *Pinus pinaster*) (Lauw et al. 2013). *Montados* maintain very high levels of biodiversity and provide "ecosystem services", such as carbon sequestration (Moreno et al. 2016). These high levels of animal and plant diversity have been attributed to the mixture of forest and open habitat at several scales, typical of the *montado* systems (Moreno et al. 2016). These traits, make the species that compose it is particularly interesting to study under a climate change scenario.

Figure 1.2: Example of a *montado* forest. © Eduardo Marabuto 2017.

One such species is *Quercus suber* L., commonly known as "cork oak" (Figure 1.3). This evergreen species of the Fagacea family is mainly distributed across the Mediterranean Basin (Figure 1.4) (Coelho et al. 2006), being naturally found in countries like Algeria, France, Italy, Morocco, Portugal, Spain and Tunisia (Pereira-Leal et al. 2014). Cork oak trees are characterized by their uncommon bark, which is composed of a peculiar type of suberin (Vishwanath et al. 2015), commonly referred to as "cork", which makes it impermeable to liquids and gases. "Cork" is the main reason *Q. suber* has such a high economic importance (Costa et al. 2015), representing approximately US$1147.5 Million each year in exports from the Iberian Peninsula (Sierra-Pérez et al. 2015). This makes this tree species (and the respective *montado* system) quite important for the economies of both Portugal and Spain, which encompass 34% and 27% of the world's cork oak forests and 50% and 30% of the world's cork production respectively (Sierra-Pérez et al. 2015). Cork extraction is known to be compatible with the typically high biodiversity levels of the *montado* systems (Sierra-Pérez et al. 2015), which makes the economic exploitation of these systems a sustainable activity (Lauw et al. 2013).

Figure 1.3: A cork oak tree. © Eduardo Marabuto 2017.

1.3 Evolutionary history of *Q. suber*

The evolutionary history of *Q. suber* has been the subject of various studies, using several molecular biology techniques. The first study of *Q. suber* with molecular data resorted to allozyme data, revealing that introgression with holm oak (*Quercus ilex*) was occurring, and that two genetically distinct groups were geographically segregated – one group in the Iberian Peninsula and adjacent French regions, and another comprising the remaining distribution range (Toumi and Lumaret 1998). The idea of genetically differentiated groups of *Q. suber* was further reinforced by PRC-RFLP techniques with cpDNA of three Oak species – *Q. ilex*, *Q. coccifera* and *Q. suber* (Jiménez et al. 2004). The same technique, applied to the whole chloroplast DNA, indicated that cork oak populations could be further divided into three groups, instead of the previous two, corresponding to potential glacial refuges in Italy, North Africa and Iberia (Lumaret et al. 2005; López de Heredia et al. 2007).

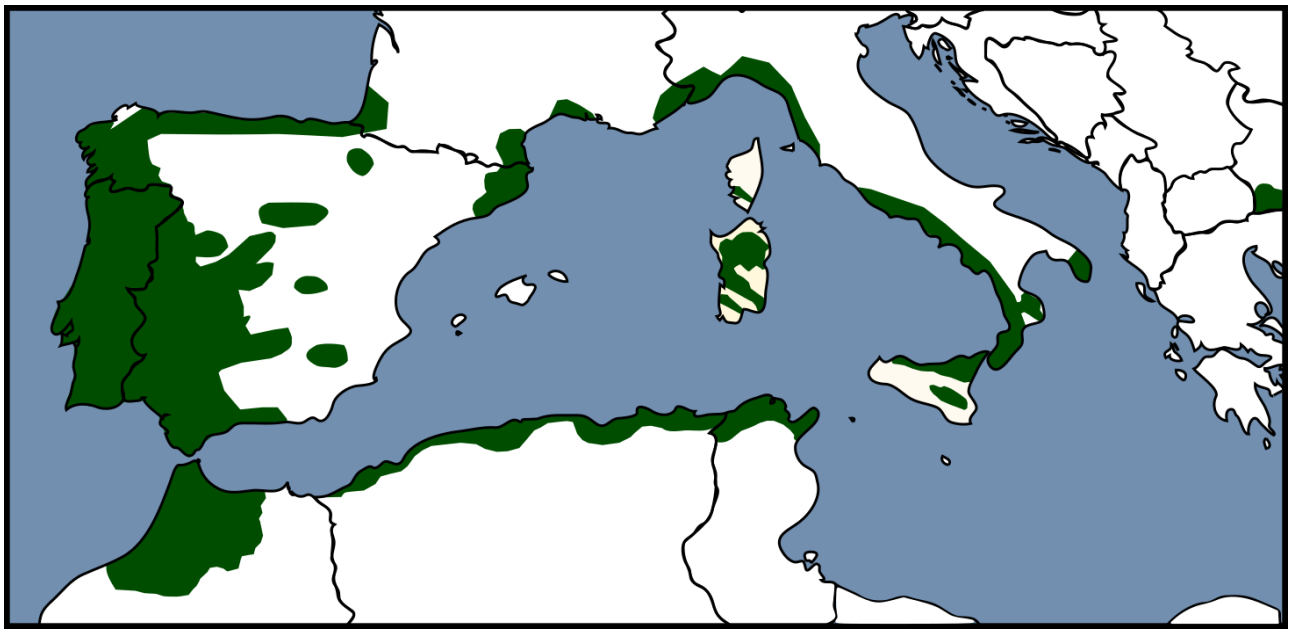

Figure 1.4: Mercator projection map of *Quercus suber*'s distribution. Cork oak trees occur in the green land areas. Adapted from EUFORGEN 2009 (www.euforgen.org).

A study with eight chloroplast microsatellite loci further increased the number of identified lineages of *Q. suber* to four – two lineages in the Iberian Peninsula, a third in Southern France, Corsica, Sardinia and Tunisia, and a fourth in the Italian Peninsula, which are considered to have originated due to plate tectonic dynamics (Magri et al. 2007) instead of glacial refugia as stated in previous studies (Figure 1.5).

The first published study of *Q. suber* using nuclear DNA resorted to AFLP markers in an attempt to associate loci with cork quality in Portuguese populations (Coelho et al. 2006). Nuclear markers were, for the first time, used to assess cork oak population structure (nine microsatellite loci, eight of which polymorphic in cork oak) and suggested a lack of population structuring in this species, although, the study was confined to a relatively small area of 11.3 ha (Soto et al. 2007), which, combined with low polymorphism levels, led the authors to maintain the question of whether or not population structuring exists in the analyzed population.

Another study focusing on *Q. suber* using nuclear DNA resorted to six microsatellite loci to assess the effects of natural selection along a temperature cline in a common garden experiment, using only samples collected in Spain (Ramírez-Valiente et al. 2009). Despite the fact that its focus was not on the species' evolutionary history, it was nevertheless an important development due to the used markers, which were later applied to search for associations between these loci, a temperature cline and tree physiological traits (Ramírez-Valiente et al. 2010). One locus revealed associations with temperature, and with leaf growth and size, hinting at effects of putative local adaptation. These same markers were further used to investigate the impact of neutral evolutionary processes on the genetic variance and functional diversity within 3 populations of *Q. suber*

(Ramírez-Valiente et al. 2014). This study revealed less genetic differentiation between populations than within populations.

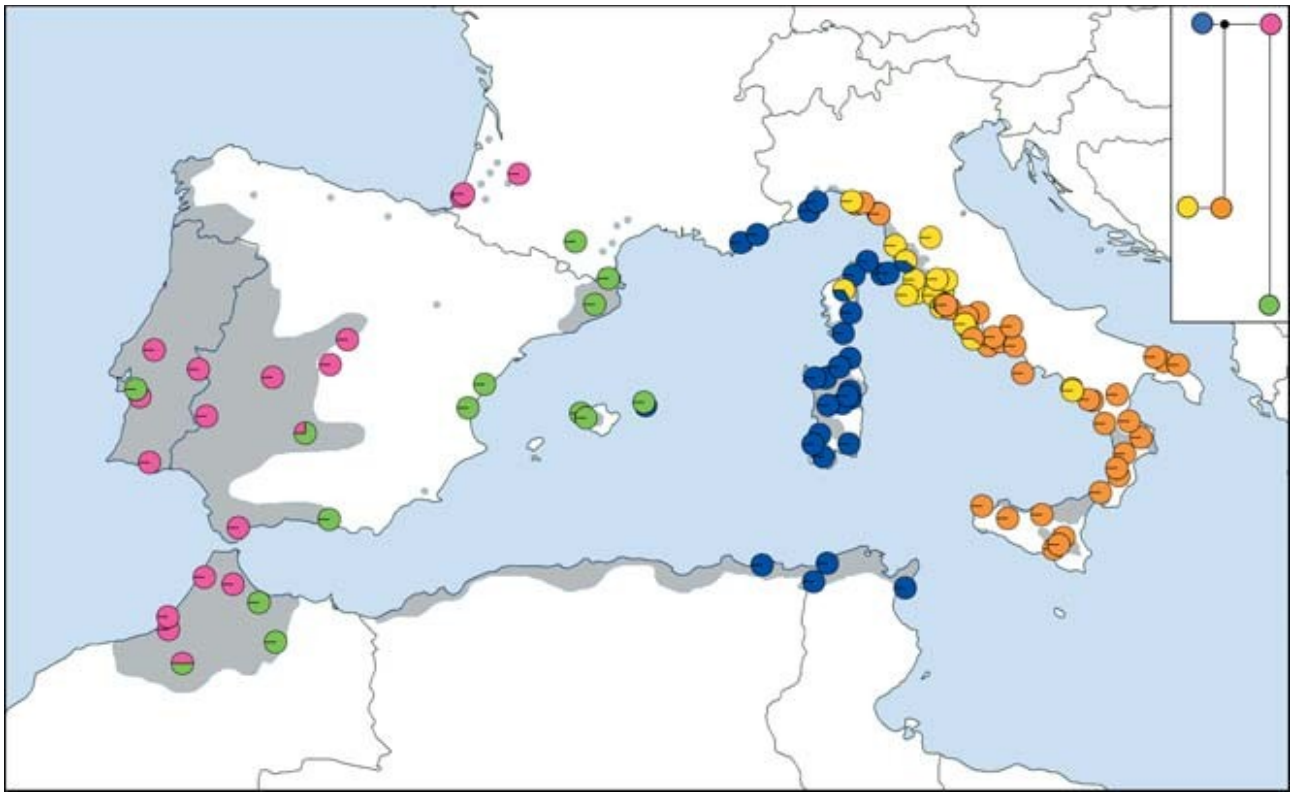

Figure 1.5: Cork oak segregation in four lineages as proposed in Magri et al. (2007).

Another study (Simeone et al. 2009) that used nuclear DNA markers (nuclear ribosomal DNA), revealed poorly supported phylogenetic differentiation between *Q. suber* populations, which was not consistent with the previously described plastidial lineages. The discrepancy, however, is attributed to hybridization between *Q. suber* and other *Quercus* species.

The evolutionary history of *Q. suber* as detailed above, is consistent with that of other European oaks, such as *Q. petraea*, (Siegismund and Jensen 2001; Bruschi et al. 2003; Muir et al. 2004) or *Q. robur* (Streiff et al. 1998), for which plastidial markers reveal population structuring, but nuclear DNA, indicates high molecular diversity, low population differentiation and weak or nonexistent structuring, suggestive of high rates of gene flow.

Despite all these studies regarding *Q. suber*'s evolutionary history, this species is far from being a model organism, with no publicly available genomic resources, and relatively few publicly available genetic sequences at the time of writing.

Another phenomenon that may be adding complexity to the evolutionary history of *Q. suber* is the hybridization with other oaks, such as *Q. ilex* and *Q. cerris* (Petit et al. 1997;

Belahbib et al. 2001; Costa et al. 2011) which can happen differentially across the species' range.

Q. suber has also been studied regarding its potential response to climate change. Some studies, performed in small scale, provide contradicting results, indicating both a likely increase in distribution (Attorre et al. 2011) and a likely decline (Benito Garzón et al. 2008). Recently, two new studies have modeled this response on a large scale (Correia et al. 2017; Vessella et al. 2017). Once again, these studies show contradictory results, one predicting that *Q. suber* will approximately maintain the same suitable distribution area (or even increase it), albeit with a shift to the North, and the other indicating that the distribution area of *Q. suber* will drastically decrease in the next 60 years. All of these studies are, however, focused on present and future environmental conditions, and do not consider the species' ability to adapt to a changing environment.

Despite both its ecological and economical importance, and relatively well studied evolutionary history, no studies regarding how *Q. suber* may respond to climate change from a genetic point of view are published.

Trying to understand the response of cork oak to climate change on a large scale, from an adaptation point of view was what motivated this work. For that, as stated above, it is important to understand the species' evolutionary history and the genetic architecture of its adaptive traits, preferably on as large a scale as possible, which is the focus of this thesis.

2 A World of Change

2.1 A new kind of data generation

It has long been known that organelle genomes tell a different evolutionary history from that of nuclear genomes, especially in plants (Govindarajulu et al. 2015), where the male gamete may migrate in both the pollen and in the seed, whereas the female gamete is restricted to the seed (Petit et al. 1993). This means that in order to know an organism's full evolutionary history, nuclear DNA (nuDNA) information is required in addition to that of monoparental inheritance, such as mitochondrial or chloroplast DNA (mtDNA and cpDNA respectively).

However, studying non-model organisms' nuDNA was typically an expensive and time consuming task until Second Generation Sequencing technologies, often referred to as Next Generation Sequencing (NGS) emerged (Schuster 2008). This situation led to non-model species being studied for many years resorting to organelle DNA, with nuDNA only being frequently used much later. The previous section regarding cork oak's evolutionary history is a clear example of this trend.

Only when access to NGS technologies became more affordable (Kumar et al. 2012) (Figure 1.6) did Single Nucleotide Polymorphisms (SNPs), the most abundant type of genetic variation in eukaryotic genomes (Rafalski 2002), became a viable solution to answer evolutionary biology questions in non-model organisms. Strategies such as SNP mining from Expressed Sequence Tags (EST) databases were slowly made possible and gained the necessary traction for widespread use (Orsini et al. 2011).

Another, more direct way to obtain SNP markers in large scale from organisms with few or no genomic resources is the use of Reduced Representation Libraries (RRLs) combined with NGS. Techniques such as Genotyping by Sequencing (GBS) (Elshire et al. 2011), or Restriction-site Associated DNA (RAD) (Baird et al. 2008), have done for the development of nuclear SNP markers what NGS technologies did for genome sequencing – a drastic lowering of the entry barrier (Lowry et al. 2017). These RRL methodologies are able to produce thousands of SNP markers per application (Van Tassell et al. 2008), which makes the development of these SNP libraries relatively cheap and fast. These are ideal qualities for a technique to be widely applied to organisms for which not many genetic/genomic resources are available.

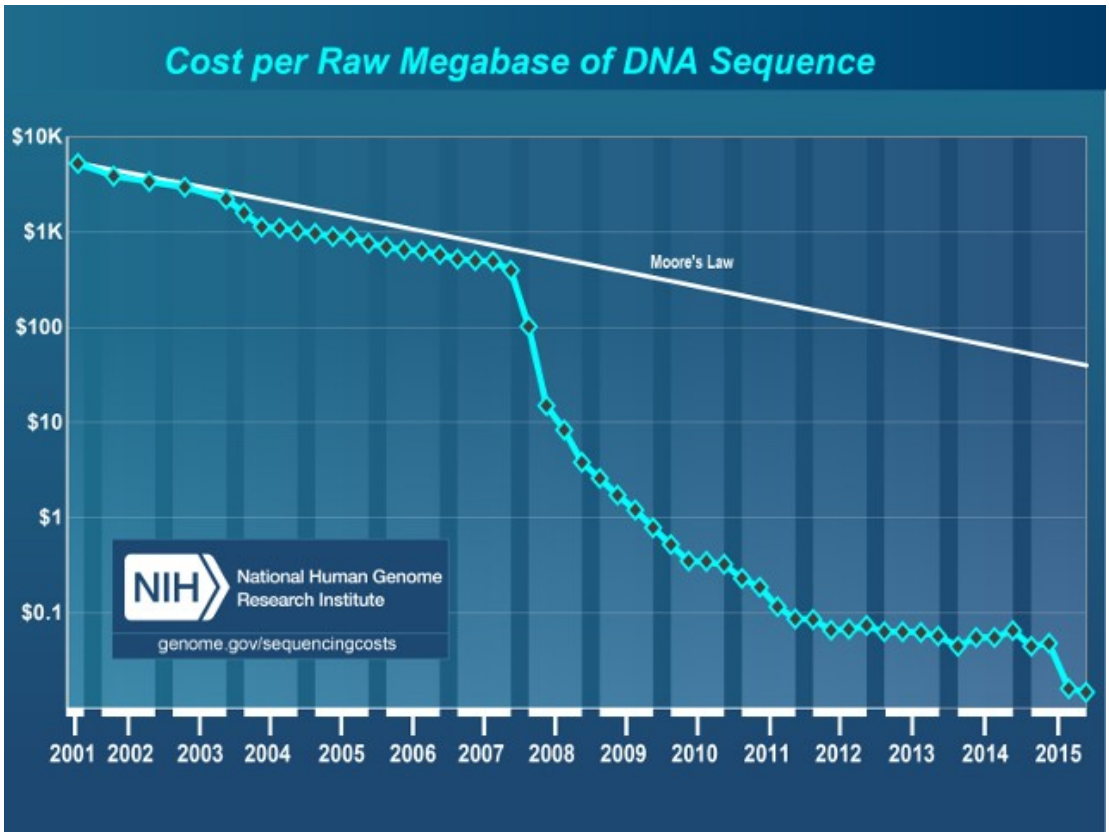

Figure 1.6: Average cost of sequencing from 2001 to 2015. Reproduced from (Wetterstrand 2016).

It is the sheer number of SNP markers that are nowadays possible to obtain for relatively low operation costs that represents their great strength (Kirk and Freeland 2011). The use of RRLs was so disruptive regarding previous techniques, that the RAD technique was considered by the journal *Science* as one of the breakthroughs of the year (Science 2010). Some authors go as far as considering these techniques as revolutionary in the field of genetics (Andrews et al. 2016).

It was the availability of NGS libraries and RRL techniques (in this case, GBS and Roche's 454) that allowed the generation of almost all the data for this thesis. In this aspect, these methodologies were of paramount importance, since without them (or their equivalents), it would not have been possible to reach the results obtained here.

2.2 A new kind of data handling

Despite the fact that studies using SNPs are becoming more common, they still bore associated bioinformatics challenges that must be overcome when developing and analysing them, especially for non-model organisms (Kumar et al. 2012).

These challenges are present in every step of the data analysis process, starting with the SNP calling methodologies (Nielsen et al. 2011), to distinguishing neutral from non-neutral markers (Porcher et al. 2006), to disentangling historical population structure

from the adaptation process. The task becomes especially difficult when migration is high or the divergent selection pressure is weak (Thibert-Plante and Hendry 2010). Although methodologies are available to deal with these issues (Foll and Gaggiotti 2008; Vitalis et al. 2014), applying them correctly is rarely a straightforward exercise.

SNP markers whose allelic frequencies differ from expectations under a neutral scenario (outlier SNPs) are not as frequent as their neutral counterparts, being represented as 1-3% of the total genotyped SNPs in some studies (Chen et al. 2012; De Kort et al. 2014; Berdan et al. 2015). However, they are known to provide a better insight into genetic diversity, local adaptation and evolutionary potential than neutral markers (Kirk and Freeland 2011). Despite the low frequency of these "special case" markers, with a large enough SNP pool it is possible to make biological inferences based on them, while still retaining significant statistical power (Andrews et al. 2016).

Many methods have been proposed to address the issue of selection detection, but even today it remains a challenging task (Vitalis et al. 2014). Early methods attempted to test markers for neutrality relying on the comparison of site-frequency spectrum summary statistics to their expected distribution from diffusion theory (Bustamante et al. 2001). Many other methods for performing this task exist, and all have known problems and caveats. Two methods are currently in widespread use due to their peculiar approaches to the problem. These methods are implemented in the software *Bayescan* (Foll and Gaggiotti 2008) and *SelEstim* (Vitalis et al. 2014). *Bayescan* is based on an island model in which subpopulation allele frequencies are correlated through a common migrant gene pool. The difference in allele frequency between this common gene pool and each subpopulation is measured by a subpopulation specific F_{ST} coefficient. This F_{ST} coefficient is then decomposed into a population specific and a general component which are used to infer departures from neutrality. *SelEstim*, on the other hand, is based on the same island model with migration as *Bayescan*, but relies on allelic frequencies and not on F_{ST} to infer which markers are under strong selective pressures (since *SelEstim* further assumes that all marker loci respond to selection in some extent). However, despite their sophistication, both these methods are affected by a relatively high rate of false positives (albeit *SelEstim* is less affected than *Bayescan*). A common empirical "work around" for this issue is to use more than one program to perform the outlier loci detection and consider only the intersection of the marker sets as being under selection (Pais et al. 2017).

Like outlier markers, SNPs that can provide insights regarding the adaptation process (whose allelic frequencies are associated with environmental or phenotypic variables) are also not very common (Eckert et al. 2015; Rellstab et al. 2016) (albeit their frequency cannot be compared between studies as in the case of outlier markers, due to the multivariate nature of the analyses). They do provide the kind of interesting information

regarding a species' biology that evolutionary history alone cannot explain (Kirk and Freeland 2011).

Environmental association analyses can be as simple as finding a direct correlation between a specific genotype and an environmental variable (Joost et al. 2007), to methodologies like *Baypass* that incorporate data transformations to account for eventual biases due to underlying population structure (Gautier 2015). Albeit these methods can provide large numbers of markers under potential selection, it is important to underline that these are still based on correlation analyses, which may not imply causation (Gautier 2015). In order to imply causation in the found correlations it is important to identify the putative function of each associated marker and relate it with the correlated variable.

SNPs are thus, very suitable markers for population genomics studies, since they can easily be used to identify and separate locus-specific effects, such as selection, mutation, assortative mating or recombination (Evans et al. 2014) from genome-wide effects, such as bottlenecks, gene flow or inbreeding (Keller et al. 2010) and consequently improve our understanding on microevolution (Black et al. 2001).

Despite the fact that this thesis consisted of field, laboratory and desktop work, the bioinformatics analyses were by far the greatest challenge. These challenges, are also the motivation behind the creation of all the software presented here as chapters 2, 3 and 4. On a personal note, it was also the most rewarding part of the PhD, even trumping some of the biological insights discovered as a result of the results interpretation.

3 Changing the World

3.1 The rise of bioinformatics

With the increase in data throughput that NGS technologies brought to the field of biology, so did the difficulty in analysing such large quantities of data (Markowetz 2017). This phenomenon led to an increase in importance and relevance of fields such as bioinformatics and computational biology (Atwood et al. 2015; Markowetz 2017). This trend can be seen in the increase of publications in PubMed containing the term “Bioinformatics” and “Computational Biology” (Figure 1.7).

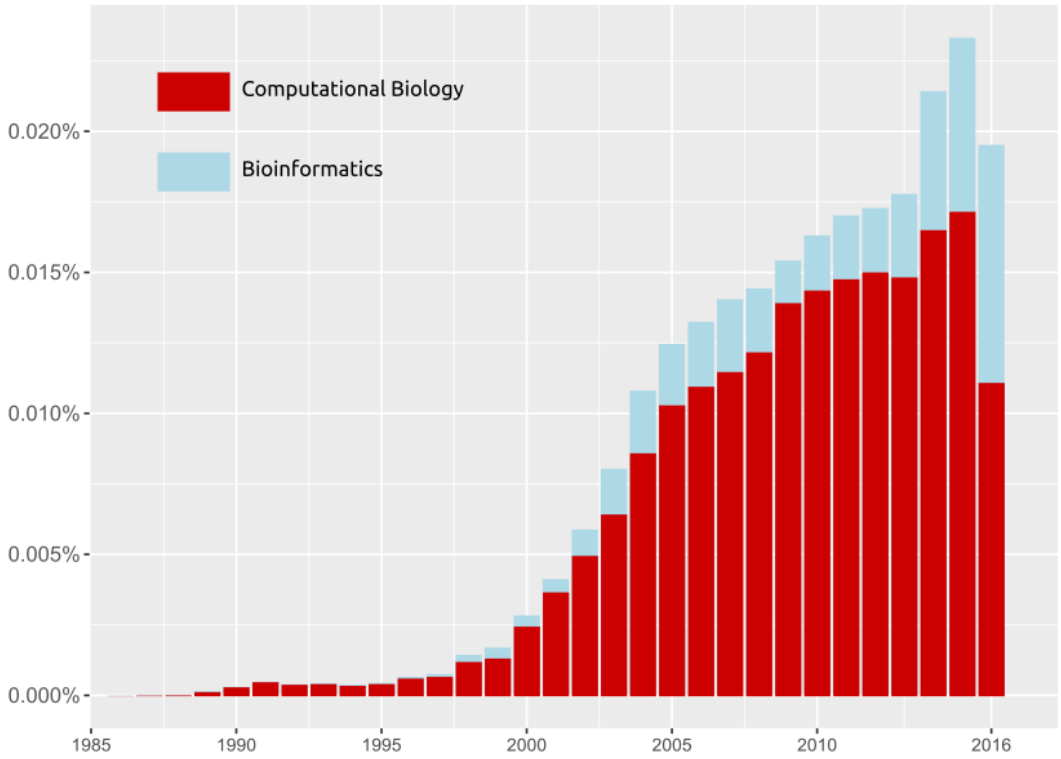

Figure 1.7: Relative number of publications per year in PubMed containing the terms "Computational Biology" and "Bioinformatics".

Defining the interdisciplinary fields of bioinformatics and computational biology, is not straightforward, and currently there is no clear consensus on the subject. However, loose definitions are put forward by some authors: both disciplines target to develop methodology and analysis tools to explore, store, organize, systematize, annotate, visualize, query, mine, understand and interpret biological data (Abdurakhmonov 2016). For this purpose, they resort to tools from the domain of computer science, statistics and mathematics (Abdurakhmonov 2016). Bioinformatics is usually considered a tool kit, therefore of a more technical nature, whereas computational biology is viewed as a

"science", of a more theoretical nature (Searls 2010; Abdurakhmonov 2016). The boundaries between these two disciplines are usually intertwined, a fact which frequently binds them together (Searls 2010).

Bioinformatics and computational biology are relatively new fields in science. Although they mostly gained notoriety in the mid-2000's, with the advent of NGS technologies, it has arguably started in the early 20th century when statistics and numerical analysis began to establish a foothold in biology (Searls 2010). These disciplines have gained notorious relevance in recent years, and biology, as a major field of science has made strides in becoming closer to other, longer established, exact sciences such as physics and mathematics (Markowetz 2017). Because of this, some would argue that the days of computational biology are numbered, since it is only a matter of time until all biology is synonymous with computational biology, and bioinformatics become just another tool specific to biology, just like microscopy, or cell culture (Markowetz 2017).

3.2 Reproducibility crisis

Reproducibility, the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator (Bollen et al. 2015) is a foundational characteristic of scientific research. Result consistency from independent research laboratories is ultimately the primary way to gather evidence for or against a formulated hypothesis (Leek and Peng 2015). The modern use of "reproducible research" was originally applied not to corroboration, but to transparency, with application in the computational sciences (Goodman et al. 2016). The term was coined by Jon Claerbout who associated it with a set of procedures that permit the reader of a scientific paper to transparently see the entire analysis process from the raw data and code to final figures and tables (Claerbout and Karrenbach 1992). This concept was quickly exported to many other data intensive science domains, such as biology, clinical trials, or economics (Goodman et al. 2016).

In recent times, the scientific community as a whole was shaken by reports that a troubling proportion of peer-reviewed studies are not reproducible (McNutt 2014). The problem continued to escalate until it was considered a "reproducibility crisis" (Schooler 2014; Baker 2016; Scannell and Bosley 2016; Voelkl and Würbel 2016).

Many reasons have been put forward as an explanation for the current "reproducibility crisis", such as measurement errors (Plant et al. 2014), sample size (Button et al. 2013), data availability (Rung and Brazma 2013; Van Noorden 2015), and closed source software (Ince et al. 2012).

Consequences of this problem are threefold: monetary – with one study estimating the costs of irreproducible research at US$28,000,000,000 per year in the USA alone (Freedman et al. 2015), scientific – the waste in resources it causes to other members of

the scientific community (Russell 2013), and social – by triggering a decrease in confidence in science from the general public (Pulverer 2015; Sarewitz 2015).

Proposals exist to solve, or at least mitigate this "crisis", such as standardized lab practices (Plant et al. 2014), more thorough planning (Button et al. 2013), open data (Ince et al. 2012; Rung and Brazma 2013; Van Noorden 2015), and open source (Ince et al. 2012). Despite the existence of these potential solutions, implementing and making them widespread takes effort, time, training and experience. Albeit the crisis is not yet surpassed, it can be argued that the path out of it has already been traced.

3.3 Hands on

Although reproducibility can usually be expensive and time consuming to achieve in wet-lab experiments (Sadowski et al. 2016), the same is not true regarding dry lab data analyses, which is comparatively orders of magnitude cheaper per retrial (Sarewitz 2015).

That being said, it is still very frequent to observe scientific papers with expressions similar to "... the analysis was performed using an in-house script...". Since these publications do not provide their full analyses environment, their results, are by default non-reproducible. As my own research progressed during the course of this work, I became more and more aware of the problem. Although this is already noticeable in chapter 5, where the traditional model of analysis description is followed, albeit with more detail than usual, and all the code is provided, it is in chapter 6 that the importance given to the reproducibility of the study is emphasised, with a fully automated and portable analysis environment. As such, chapter 6 contains supplementary material that will allow any researcher to reproduce the entire data analyses. This means that the input data, code, configuration files and parameters for every single analysis step is provided with the manuscript.

Most of the work performed in this thesis is clearly inserted in the computational biology and bioinformatics fields, and the bulk of the analyses are dry-lab, rather than wet-lab. In an attempt to pioneer the way out of the aforementioned reproducibility crisis, a great deal of effort was made to ensure all the dry-lab analyses presented are reproducible.

The objective of this endeavour is twofold – on one hand, it allows any reader to scrutinize, understand and build upon what was performed for the data analyses steps. On the other hand, it works as a way to increase confidence in the obtained results, by making sure every step can be automated (reducing human error), that the results of multiple subsequent runs are the same and to keep track of what was tried during the exploratory phase.

3.4 Objectives & short description

Three main objectives were envisioned when planning this PhD:

1. Assess the evolutionary history of *Quercus suber* based on genomic SNP markers, resorting to a distribution wide sampling strategy;
2. Assess the action of natural selection on the cork oak, based on outlier and association analyses and infer its role in shaping the species' genetic background;
3. Improve data analyses methodological reliability by automating and streamlining the process, consequently increasing the reproducibility value of the work performed for the thesis.

Chapters 2, 3 and 4 are mostly technical, dedicated to fulfilling the third objective. These describe the most important software projects designed and written specifically for the objectives of this work, but that are likely to have an impact on the broader scientific community. Chapters 5 and 6 represent the biological meaningful analyses that were performed to meet the first and second thesis objectives.

4 References

Abdurakhmonov IY. 2016. Bioinformatics: Basics, Development, and Future. Available from: http://www.intechopen.com/books/bioinformatics-updated-features-and-applications/bioinformatics-basics-development-and-future

Alberto FJ, Aitken SN, Alia R, Gonzalez-Martinez SC, Hanninen H, Kremer A, Lefevre F, Lenormand T, Yeaman S, Whetten R, et al. 2013. Potential for evolutionary responses to climate change - evidence from tree populations. Glob. Change Biol. 19:1645–1661.

Alley RB. 2003. Abrupt Climate Change. Science 299:2005–2010.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17:81–92.

Attorre F, Alfò M, De Sanctis M, Francesconi F, Valenti R, Vitale M, Bruno F. 2011. Evaluating the effects of climate change on tree species abundance and distribution in the Italian peninsula. Appl. Veg. Sci. 14:242–255.

Atwood TK, Bongcam-Rudloff E, Brazas ME, Corpas M, Gaudet P, Lewitter F, Mulder N, Palagi PM, Schneider MV, Gelder CWG van, et al. 2015. GOBLET: The Global Organisation for Bioinformatics Learning, Education and Training. PLOS Comput. Biol. 11:e1004143.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376.

Baker M. 2016. 1,500 scientists lift the lid on reproducibility. Nat. News 533:452.

Belahbib N, Pemonge MH, Ouassou A, Sbay H, Kremer A, Petit RJ. 2001. Frequent cytoplasmic exchanges between oak species that are not closely related: Quercus suber and Q. ilex in Morocco. Mol. Ecol. 10:2003–2012.

Beniston M, Stephenson DB, Christensen OB, Ferro CAT, Frei C, Goyette S, Halsnaes K, Holt T, Jylhä K, Koffi B, et al. 2007. Future extreme events in European climate: an exploration of regional climate model projections. Clim. Change 81:71–95.

Benito Garzón M, Sánchez de Dios R, Sainz Ollero H. 2008. Effects of climate change on the distribution of Iberian tree species. Appl. Veg. Sci. 11:169–178.

Berdan EL, Mazzoni CJ, Waurick I, Roehr JT, Mayer F. 2015. A population genomic scan in Chorthippus grasshoppers unveils previously unknown phenotypic divergence. Mol. Ecol. 24:3918–3930.

Black WC 4th, Baer CF, Antolin MF, DuTeau NM. 2001. Population genomics: genome-wide sampling of insect populations. Annu. Rev. Entomol. 46:441–469.

Bollen K, Cacioppo JT, Kaplan RM, Krosnick JA, Olds JL. 2015. Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. Available from: https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf
Briceño-Elizondo E, Garcia-Gonzalo J, Peltola H, Matala J, Kellomäki S. 2006. Sensitivity of growth of Scots pine, Norway spruce and silver birch to climate change and forest management in boreal conditions. For. Ecol. Manag. 232:152–167.
Bruschi P, Vendramin GG, Bussotti F, Grossoni P. 2003. Morphological and Molecular Diversity Among Italian Populations of Quercus petraea (Fagaceae). Ann. Bot. 91:707–716.
Bustamante CD, Wakeley J, Sawyer S, Hartl DL. 2001. Directional Selection and the Site-Frequency Spectrum. Genetics 159:1779–1788.
Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR. 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14:365–376.
Chen J, Källman T, Ma X, Gyllenstrand N, Zaina G, Morgante M, Bousquet J, Eckert A, Wegrzyn J, Neale D, et al. 2012. Disentangling the Roles of History and Local Selection in Shaping Clinal Variation of Allele Frequencies and Gene Expression in Norway Spruce (Picea abies). Genetics 191:865–881.
Claerbout J, Karrenbach M. 1992. Electronic documents give reproducible research a new meaning. In: SEG Technical Program Expanded Abstracts 1992. SEG Technical Program Expanded Abstracts. Society of Exploration Geophysicists. p. 601–604. Available from: http://library.seg.org/doi/abs/10.1190/1.1822162
Coelho AC, Lima MB, Neves D, Cravador A. 2006. Genetic diversity of two evergreen oaks [Quercus suber (L.) and Quercus ilex subsp rotundifolia (Lam.)] in Portugal using AFLP markers. SILVAE Genet. 55:105–118.
Correia RA, Bugalho MN, Franco AMA, Palmeirim JM. 2017. Contribution of spatially explicit models to climate change adaptation and mitigation plans for a priority forest habitat. Mitig. Adapt. Strateg. Glob. Change:1–16.
Costa A, Oliveira G, Shackleton C, Pandey A, Ticktin T. 2015. Cork oak (Quercus suber L.): A case of sustainable bark harvesting in Southern Europe. Ecol. Sustain. Non-Timber For. Prod. Dyn. Case Stud. Harvest.:179–198.
Costa J, Miguel C, Almeida H, Oliveira MM, Matos JA, Simões F, Veloso M, Ricardo PC, Paulo OS, Batista D. 2011. Genetic divergence in Cork Oak based on cpDNA sequence data. BMC Proc. 5:P13.
De Kort H, Vandepitte K, Bruun HH, Closset-Kopp D, Honnay O, Mergeay J. 2014. Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species Alnus glutinosa. Mol. Ecol. 23:4709–4721.
Eckert AJ, Maloney PE, Vogler DR, Jensen CE, Mix AD, Neale DB. 2015. Local adaptation at fine spatial scales: an example from sugar pine (Pinus lambertiana, Pinaceae). Tree Genet. Genomes 11:42.
Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLOS ONE 6:e19379.
Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J-G, et al. 2014. Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. Nat. Genet. [Internet] advance online publication. Available from: http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.3075.html
Foll M, Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180:977–993.
Freedman LP, Cockburn IM, Simcoe TS. 2015. The Economics of Reproducibility in Preclinical Research. PLOS Biol. 13:e1002165.
Gautier M. 2015. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. Genetics:genetics.115.181453.
Goodman SN, Fanelli D, Ioannidis JPA. 2016. What does research reproducibility mean? Sci. Transl. Med. 8:341ps12-341ps12.
Govindarajulu R, Parks M, Tennessen JA, Liston A, Ashman T-L. 2015. Comparison of nuclear, plastid, and mitochondrial phylogenies and the origin of wild octoploid strawberry species. Am. J. Bot. 102:544–554.
Ince DC, Hatton L, Graham-Cumming J. 2012. The case for open computer programs. Nature 482:485–488.
IPCC. 2014. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC AR5 Synth. Rep. Website:151 pp.
Jiménez 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.

Joffre R, Rambal S, Ratte JP. 1999. The dehesa system of southern Spain and Portugal as a natural ecosystem mimic. Agrofor. Syst. 45:57–79.
Joost S, Bonin A, Bruford MW, Després L, Conord C, Erhardt G, Taberlet P. 2007. A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. Mol. Ecol. 16:3955–3969.
Keller SR, Olson MS, Silim S, Schroeder W, Tiffin P. 2010. Genomic diversity, population structure, and migration following rapid range expansion in the Balsam Poplar, Populus balsamifera. Mol. Ecol. 19:1212–1226.
Kirk H, Freeland JR. 2011. Applications and Implications of Neutral versus Non-neutral Markers in Molecular Ecology. Int. J. Mol. Sci. 12:3966–3988.
Kremer A, Potts BM, Delzon S. 2014. Genetic divergence in forest trees: understanding the consequences of climate change. Funct. Ecol. 28:22–36.
Kremer A, Ronce O, Robledo-Arnuncio JJ, Guillaume F, Bohrer G, Nathan R, Bridle JR, Gomulkiewicz R, Klein EK, Ritland K, et al. 2012. Long-distance gene flow and adaptation of forest trees to rapid climate change. Ecol. Lett. 15:378–392.
Kumar S, Banks TW, Cloutier S. 2012. SNP Discovery through Next-Generation Sequencing and Its Applications. Int. J. Plant Genomics [Internet] 2012. Available from: http://www.hindawi.com/journals/ijpg/2012/831460/abs/
Lauw A, Gonçalves-Ferreira A, Gomes A, Moreira AC, Fonseca A, Azul A, Mira A, Murilhas A, Pinheiro AC, Costa A, et al. 2013. Livro Verde dos Montados. ICAAM Available from: https://dspace.uevora.pt/rdpc/handle/10174/10116?mode=full&submit_simple=Show+full+item+record
Leek JT, Peng RD. 2015. Opinion: Reproducible research can still be wrong: Adopting a prevention approach. Proc. Natl. Acad. Sci. 112:1645–1646.
Lindner M, Fitzgerald JB, Zimmermann NE, Reyer C, Delzon S, van der Maaten E, Schelhaas M-J, Lasch P, Eggers J, van der Maaten-Theunissen M, et al. 2014. Climate change and European forests: What do we know, what are the uncertainties, and what are the implications for forest management? J. Environ. Manage. 146:69–83.
Lindner M, Maroschek M, Netherer S, Kremer A, Barbati A, Garcia-Gonzalo J, Seidl R, Delzon S, Corona P, Kolström M, et al. 2010. Climate change impacts, adaptive capacity, and vulnerability of European forest ecosystems. For. Ecol. Manag. 259:698–709.
López de Heredia U, Carrión JS, Jiménez P, Collada C, Gil L. 2007. Molecular and palaeoecological evidence for multiple glacial refugia for evergreen oaks on the Iberian Peninsula. J. Biogeogr. 34:1505–1517.
Loustau D, Bosc A, Colin A, Ogee J, Davi H, Francois C, Dufrene E, Deque M, Cloppet E, Arrouays D, et al. 2005. Modeling climate change effects on the potential production of French plains forests at the sub-regional level. Tree Physiol. 25:813–823.
Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, Storfer A. 2017. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. Mol. Ecol. Resour. 17:142–152.
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). Ann. Bot. 96:853–861.
Magri D, Fineschi S, Bellarosa R, Buonamici A, Sebastiani F, Schirone B, Simeone MC, Vendramin GG. 2007. The distribution of Quercus suber chloroplast haplotypes matches the palaeogeographical history of the western Mediterranean. Mol. Ecol. 16:5259–5266.
Markowetz F. 2017. All biology is computational biology. PLOS Biol. 15:e2002050.
McNutt M. 2014. Reproducibility. Science 343:229–229.
Milad M, Schaich H, Bürgi M, Konold W. 2011. Climate change and nature conservation in Central European forests: A review of consequences, concepts and challenges. For. Ecol. Manag. 261:829–843.
Moreno G, Gonzalez-Bornay G, Pulido F, Lopez-Diaz ML, Bertomeu M, Juárez E, Diaz M. 2016. Exploring the causes of high biodiversity of Iberian dehesas: the importance of wood pastures and marginal habitats. Agrofor. Syst. 90:87–105.
Muir G, Lowe AJ, Fleming CC, Vogl C. 2004. High Nuclear Genetic Diversity, High Levels of Outcrossing and Low Differentiation Among Remnant Populations of Quercus petraea at the Margin of its Range in Ireland. Ann. Bot. 93:691–697.
Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12:443–451.
Olson MS, Levsen N, Soolanayakanahally RY, Guy RD, Schroeder WR, Keller SR, Tiffin P. 2013. The adaptive potential of Populus balsamifera L. to phenology requirements in a warmer global climate. Mol. Ecol. 22:1214–1230.

Oreskes N. 2004. BEYOND THE IVORY TOWER: The Scientific Consensus on Climate Change. Science 306:1686–1686.
Orsini L, Jansen M, Souche EL, Geldof S, De Meester L. 2011. Single nucleotide polymorphism discovery from expressed sequence tags in the waterflea Daphnia magna. BMC Genomics 12:309.
Pais AL, Whetten RW, Xiang Q-Y (Jenny). 2017. Ecological genomics of local adaptation in Cornus florida L. by genotyping by sequencing. Ecol. Evol. 7:441–465.
Pereira-Leal JB, Abreu IA, Alabaça CS, Almeida MH, Almeida P, Almeida T, Amorim MI, Araújo S, Azevedo H, Badia A, et al. 2014. A comprehensive assessment of the transcriptome of cork oak (Quercus suber) through EST sequencing. BMC Genomics 15:371.
Petit R, Kremer A, Wagner D. 1993. Heredity - Abstract of article: Finite island model for organelle and nuclear genes in plants. Heredity 71:630–641.
Petit RJ, Pineau E, Demesure B, Bacilieri R, Ducousso A, Kremer A. 1997. Chloroplast DNA footprints of postglacial recolonization by oaks. Proc. Natl. Acad. Sci. 94:9996–10001.
Plant AL, Locascio LE, May WE, Gallagher PD. 2014. Improved reproducibility by assuring confidence in measurements in biomedical research. Nat. Methods 11:895–898.
Porcher E, Giraud T, Lavigne C. 2006. Genetic differentiation of neutral markers and quantitative traits in predominantly selfing metapopulations: confronting theory and experiments with Arabidopsis thaliana. Genet. Res. 87:1–12.
Pulverer B. 2015. Reproducibility blues. EMBO J. 34:2721–2724.
Rafalski A. 2002. Applications of single nucleotide polymorphisms in crop genetics. Curr. Opin. Plant Biol. 5:94–100.
Ramírez-Valiente JA, Lorenzo Z, Soto A, Valladares F, Gil L, Aranda I. 2009. Elucidating the role of genetic drift and natural selection in cork oak differentiation regarding drought tolerance. Mol. Ecol. 18:3803–3815.
Ramírez-Valiente JA, Lorenzo Z, Soto A, Valladares F, Gil L, Aranda I. 2010. Natural selection on cork oak: allele frequency reveals divergent selection in cork oak populations along a temperature cline. Evol. Ecol. 24:1031–1044.
Ramírez-Valiente JA, Valladares F, Aranda I. 2014. Exploring the impact of neutral evolution on intrapopulation genetic differentiation in functional traits in a long-lived plant. Tree Genet. Genomes:1–10.
Rellstab C, Zoller S, Walthert L, Lesur I, Pluess AR, Graf R, Bodénès C, Sperisen C, Kremer A, Gugerli F. 2016. Signatures of local adaptation in candidate genes of oaks (*Quercus* spp.) in respect to present and future climatic conditions. Mol. Ecol. [Internet]. Available from: http://doi.wiley.com/10.1111/mec.13889
Rung J, Brazma A. 2013. Reuse of public genome-wide gene expression data. Nat. Rev. Genet. 14:89–99.
Russell JF. 2013. If a job is worth doing, it is worth doing twice. Nat. News 496:7.
Sadowski MI, Grant C, Fell TS. 2016. Harnessing QbD, Programming Languages, and Automation for Reproducible Biology. Trends Biotechnol. 34:214–227.
Sarewitz D. 2015. Reproducibility will not cure what ails science. Nat. News 525:159.
Saxe H, Cannell MGR, Johnsen Ø, Ryan MG, Vourlitis G. 2002. Tree and forest functioning in response to global warming: Tansley review no. 123. New Phytol. 149:369–399.
Scannell JW, Bosley J. 2016. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. PLOS ONE 11:e0147215.
Schooler JW. 2014. Metascience could rescue the 'replication crisis.' Nat. News 515:9.
Schuster SC. 2008. Next-generation sequencing transforms today's biology. Nat. Methods 5:16–18.
Science AA for the A of. 2010. Areas to Watch. Science 330:1608–1609.
Searls DB. 2010. The Roots of Bioinformatics. PLOS Comput. Biol. 6:e1000809.
Siegismund HR, Jensen JS. 2001. Intrapopulation and Interpopulation Genetic Variation of Quercus in Denmark. Scand. J. For. Res. 16:103–116.
Sierra-Pérez J, Boschmonart-Rives J, Gabarrell X. 2015. Production and trade analysis in the Iberian cork sector: Economic characterization of a forest industry. Resour. Conserv. Recycl. 98:55–66.
Simeone, Cosimo M, Papini A, Vessella F, Bellarosa R, Spada F, Schirone B. 2009. Multiple genome relationships and a complex biogeographic history in the eastern range of Quercus suber L. (Fagaceae) implied by nuclear and chloroplast DNA variation. Caryologia 62:236–252.
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.

Streiff R, Labbe T, Bacilieri R, Steinkellner H, Glössl J, Kremer A. 1998. Within-population genetic structure in Quercus robur L. and Quercus petraea (Matt.) Liebl. assessed with isozymes and microsatellites. Mol. Ecol. 7:317–328.
Thibert-Plante X, Hendry AP. 2010. When can ecological speciation be detected with neutral loci? Mol. Ecol. 19:2301–2314.
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. Theor. Appl. Genet. 97:647–656.
Van Noorden R. 2015. Sluggish data sharing hampers reproducibility effort. Nat. News [Internet]. Available from: http://www.nature.com/news/sluggish-data-sharing-hampers-reproducibility-effort-1.17694
Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5:247–252.
Vessella F, López-Tirado J, Simeone MC, Schirone B, Hidalgo PJ. 2017. A tree species range in the face of climate change: cork oak as a study case for the Mediterranean biome. Eur. J. For. Res.:1–15.
Vishwanath SJ, Delude C, Domergue F, Rowland O. 2015. Suberin: biosynthesis, regulation, and polymer assembly of a protective extracellular barrier. Plant Cell Rep. 34:573–586.
Vitalis R, Gautier M, Dawson KJ, Beaumont MA. 2014. Detecting and Measuring Selection from Gene Frequency Data. Genetics 196:799–817.
Voelkl B, Würbel H. 2016. Reproducibility Crisis: Are We Ignoring Reaction Norms? Trends Pharmacol. Sci. 37:509–510.
Walther G-R, Hughes L, Vitousek P, Stenseth NC. 2005. Consensus on climate change. Trends Ecol. Evol. 20:648–649.
Walther G-R, Post E, Convey P, Menzel A, Parmesan C, Beebee TJC, Fromentin J-M, Hoegh-Guldberg O, Bairlein F. 2002. Ecological responses to recent climate change. Nature 416:389–395.
Wetterstrand KA. 2016. DNA Sequencing Costs: Data. DNA Seq. Costs Data NHGRI Genome Seq. Program GSP [Internet]. Available from: https://www.genome.gov/27541954/DNA-Sequencing-Costs-Data

Chapter 2

4Pipe4 – A 454 data analysis pipeline for SNP detection in datasets with no reference sequence or strain information

1 Abstract

Background: Next-generation sequencing datasets are becoming more frequent, and their use in population studies is becoming widespread. For non-model species, without a reference genome, it is possible from a panel of individuals to identify a set of SNPs that can be used for further population genotyping. However the lack of a reference genome to which the sequenced data could be compared makes the finding of SNPs more troublesome. Additionally when the data sources (strains) are not identified (e.g. in datasets of pooled individuals), the problem of finding reliable variation in these datasets can become much more difficult due to the lack of specialized software for this specific task.

Results: Here we describe 4Pipe4, a 454 data analysis pipeline particularly focused on SNP detection when no reference or strain information is available. It uses a command line interface to automatically call other programs, parse their outputs and summarize the results. The variation detection routine is built-in in the program itself. Despite being optimized for SNP mining in 454 EST data, it is flexible enough to automate the analysis of genomic data or even data from other NGS technologies. 4Pipe4 will output several HTML formatted reports with metrics on many of the most common assembly values, as well as on all the variation found. There is also a module available for finding putative SSRs in the analysed datasets.

Conclusions: This program can be especially useful for researchers that have 454 datasets of a panel of pooled individuals and want to discover and characterize SNPs for subsequent individual genotyping with customized genotyping arrays. In comparison with other SNP detection approaches, 4Pipe4 showed the best validation ratio, retrieving a smaller number of SNPs but with a considerably lower false positive rate than other methods.

4Pipe4's source code is available at https://github.com/StuntsPT/4Pipe4.

2 Background

With the democratization of NGS technologies, large amounts of genomic and transcriptomic data became available to scientists in a short time span (Schuster, 2008). However, this magnitude of sequence data has brought most researchers a new bioinformatics challenge: to analyse and mine very large datasets (Papanicolaou, Stierli, Ffrench-Constant, & Heckel, 2009). One of the areas of particular interest of NGS data analysis is the detection of sequence polymorphisms. This task, however, becomes particularly difficult when no reference genome is available, which is common in non model organisms. This problem is somewhat mitigated when the samples can be accurately identified (strain information is present) (Peterlongo, Schnel, Pisanti, Sagot, &

Lacroix, 2010). However, if neither of these is accessible – such as in datasets with pools of individuals, looking for reliable variation can be a real problem. It was for this purpose that 4Pipe4 was developed: to find variation in 454 EST datasets where no reference sequence or strain information is available. This is especially useful for researchers who wish to find reliable variation in a panel dataset of pooled individuals to use as a starting point for designing genotyping arrays to further explore their data. The pipeline can provide very high quality SNPs as well as the flanking region sequence, necessary for the design of customized genotyping arrays, currently the most efficient way to extend SNP genotyping from those found in a panel of samples to a larger set of individuals for population genomic studies (Modesto et al., 2014; Savage, Kiemnec-Tyburczy, Ellison, Fleischer, & Zamudio, 2014).

Due to the nature of NGS data, any automated pipeline has to be strict enough as to follow a work-flow but, at the same time, flexible enough to serve the different purposes of each investigator. This is the role that 4Pipe4 intends to take. Although 4Pipe4 is tuned for EST data, it can also be used with genomic data and, to some extent, to help automate the process of gene discovery.

3 Implementation

4Pipe4 is written in Python 3 and is licensed under the GPLv3. It is written in a modular manner that allows for relatively simple expansion of functionality.

Most of the functions present in 4Pipe4 result from the automation of already existing programs and the integration of their respective outputs. However, the variation detection routines are of original design and are based on three criteria, all of which can be adjusted by the user:

Base coverage – The minimum required coverage (C); the default value is 15;

Base variants – The minimum number of equal base variants required in a position (v_{min}); the default value is 20% of the minimum required coverage;

Base quality – The average minimum quality of each of the base variants (Q_{min}); the default value is 70.

This means that in order to consider a position of the alignment as a putative SNP, the below condition must be verified:

$$\sum R \geqslant C \wedge \sum V2 \geqslant v_{min} \wedge -Q_{V1} \geqslant Q_{min} \wedge -Q_{V2} \geqslant Q_{min}$$

Where “R” is the number of reads in the considered position, “C” is the minimum coverage as defined by the user, “V1” is the most frequent variant base type in the considered position, “V2” is the second most frequent variant base type in the considered position and “Q” is the quality value.

4Pipe4 uses a configuration file, called “4Pipe4rc” with a simple and self documented syntax for setting variables such as the location of programs, the SNP detection criteria and the parameters that should be passed to the external software. How the program uses this configuration file is explained in detail in the program documentation.

The analysis process is divided in 9 steps, each of which can be excluded from the run by issuing the appropriate arguments at run time. In step 1, 4Pipe4 takes an SFF file and, if all the steps are run, step 9 outputs a series of HTML formatted reports, compressed in *7zip*. Steps 7 (Gene Ontology) and 8 (SSR detection) are considered optional since they are not required for the SNP detection routines. 4Pipe4 requires the use of external programs, which can all be installed locally without root privileges (except Blast2GO which requires a MySQL database). The distribution comes with a set of helper scripts to automatically download and install all of the required software. All of the required programs are available under open-source licenses and are free to use (except Blast2GO which is not open source, but is free to use).

4 Results and Discussion

4.1 The analysis process

The above mentioned 9 steps can be described as follows (See Figure 2.1 for a more graphical overview):

Step 1 – Extraction of the “FASTA” and “FASTA.QUAL” files from the original “SFF” file. This step can be skipped if not dealing with 454 data.

Step 2 – “Cleaning” the sequences, by discarding low overall quality and short reads, as well as reads that contain contaminants matched against the “UNIVEC” database (“The UniVec Database,” n.d.) or any other contaminant database at the user's discretion. This step uses the “Sequence Cleaner” program (“Sequence Cleaner,” n.d.) and can also be skipped if dealing with Illumina data.

Step 3 – Assembling. This step uses mira (Chevreux et al., 2004). A set of optimized parameters for SNP calling is contained in the example configuration file.

Steps 4 and 5 – SNP gathering. Resorting to the “MAF” output from step 3 (which is converted into the “SAM” (Li et al., 2009) format), potential SNPs are identified in the assembly. The result is a summary intermediate “TCS” file and a “FASTA” file including all the “contigs” that contain putative SNPs (which are identified in the sequence title). The software “pysam” (“pysam-developers/pysam,” n.d.) (Li et al., 2009) is used in this step.

Step 6 – Characterization of the detected SNPs, by attempting to fit them into Open Reading Frames (ORFs). The result is a “FASTA” file containing the ORFs with the SNPs identified in the sequence title, as well as the ORF frame allowing the quick assessment

of the length and level of conservation of the SNP's flanking region. This step uses the "EMBOSS getorf" program (Rice, Longden, & Bleasby, 2000). Also in this step, BLASTx (Altschul, Gish, Miller, Myers, & Lipman, 1990) is run with the resulting ORFs against a large protein database, such as NCBI's "nr". Lastly, this step will produce an HTML formatted report with the characterized SNPs for easy referencing. The report is formatted as a table and can easily be transferred to any spreadsheet software for further data exploring. Another output of this step is an additional HTML report with a compilation of various dataset metrics.

Step 7 (optional) – Blast2GO annotation; this step queries the contigs that contain SNPs against a large protein database such as NCBI's 'nr' using BLASTx; these are then run through Blast2GO (Conesa et al., 2005) using Blast2Go4Pipe, resulting in an annotation file that can be further analysed with Blast2GO itself.

Step 8 (optional) – SSR detection, by using "EMBOSS etandem" to detect potential SSRs in the assembly. The required quality of the putative SSRs is defined in the configuration file.

Step 9 – Compression of all the relevant result files into a 7zip archive which simplifies the transfer of (often large) results.

4.2 Example usage

A test dataset with documentation on example usage is provided with the software package. An example resulting report is also provided for the test dataset (run with default values on all settings).

4.3 Validation

In order to assess the efficiency of SNP detection and the rate of false positives, and assess the best default values to use, an approach using reference data was used.

For this goal, two reference sequences of two *E. coli* strains were used (http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=ADWQ01 - Strain 85 and http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=ADWR01 - Strain 79). Two 454 datasets were also downloaded from the NCBI Sequence Read Archive (SRA) (Leinonen, Sugawara, & Shumway, 2011) (http://www.ncbi.nlm.nih.gov/sra/SRX036805 and http://www.ncbi.nlm.nih.gov/sra/SRX036804) for the same strains as the references.

To assess the number of SNPs between the strains that could be found on the 454 datasets, the 454 reads of one strain were mapped against the reference sequences of the other strain using bowtie2 (Langmead & Salzberg, 2012). Atlas-SNP2 (Shen et al., 2010) reported 29673 SNPs between the reference sequence '85' and the 454 reads of

the strain '79', and 28525 SNPs between the reference sequence '79' and the 454 reads of the strain '85'.

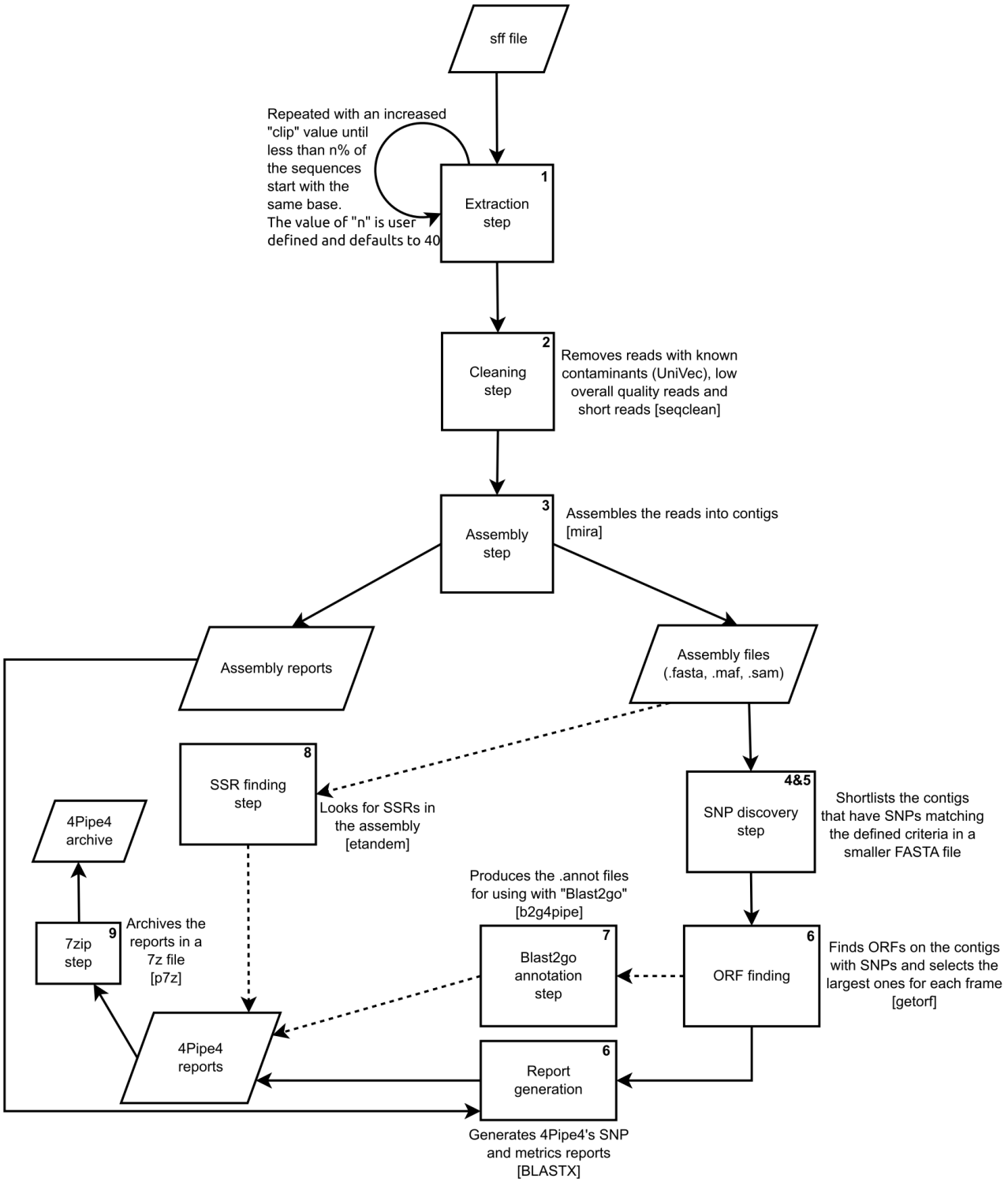

Figure 2.1: 4Pipe4 flowchart. The rectangular shapes represent processes, the rhomboid shapes represent input/output files. The dashed arrows represent optional steps. The names inside square brackets are the names of the used external programs. The digits on the top right corner of each rectangle represent the step number of each process.

4Pipe4 was then run on the two merged 454 datasets, discarding all strain information.

Although this validation method is not as good as true wet-lab genotyping, it is likely to be a good proxy, since Atlas-SNP2 is known to have very high sensitivity and specificity when dealing with 454 datasets (Shen et al., 2010).

The results varied with the different tested parameters (Table 2.1), but the best output was obtained with the default values of minimum coverage of 15 and minimum average quality of 70 per variant. This setup retrieved 114 SNPs, of which 32 did not match to any of those detected by Atlas-SNP2, being thus, considered false positives (28.07% false discovery rate).

Table 2.1: Obtained and validated SNPs per parameter set. Of the six tested parameter combinations, the lowest false positive rate was retrieved with the default values: 15 Minimum coverage and 70 Minimum average quality.

	Parameters used (Min. Coverage|Min. Average Quality)					
	10|60	10|70	15|60	15|70 (Default)	15|75	20|70
Total SNPs retrieved	234	169	155	114	107	89
Confirmed SNPs	97	86	88	82	69	57
False Positive rate (%)	58.55	49.11	43.23	28.07	35.51	35.96

Although the number of provided SNPs is relatively low, due to the restrictive assembly and filtering parameters, we find this a good trade-off relative to the high confidence of the retrieved SNPs.

The task of SNP calling in 454 data has been performed before on organisms without a reference sequence or strain information, with varying degrees of false positives. One such study, conducted using custom scripts for SNP calling provided a false positive rate of 80% on 4200 retrieved SNPs (Tollenaere et al., 2012). Another example, where the contigs of 283 SNPs were manually screened and selected, had a slightly better false positive rate of 45% (Broders, Woeste, San Miguel, Westerman, & Boland, 2011).

The above mentioned studies are not directly comparable to the results of the benchmark performed here, since they are performed on different datasets, nevertheless they can be used to infer that, in general, 4Pipe4 retrieves a smaller number of SNPs than other methods, but with a considerably lower false positive rate. Since the main goal of this pipeline is to provide the user with high confidence SNPs for genotyping arrays, a rate of 28.07% false positives is a considerable improvement relative to the other mentioned approaches.

4.4 4Pipe4 compared to other software

Although 4Pipe4 is specifically designed for the purpose of detecting variation when no strain information or reference sequence is available, other software exists that can be used for the same purpose, but which differs from 4Pipe4 in some aspects:

QualitySNP (Tang, Vosman, Voorrips, van der Linden, & Leunissen, 2006) – Relies on CAP3 for clustering the reads (which is optimized for Sanger sequences, while 4Pipe4 uses mira, which is optimized for NGS data). Requires perl, PHP, a configured webserver and a MySQL database for SNP retrieval. This means that root access to the machine in which the program is being run on is required. Furthermore, QualitySNP has been superseded by the simpler and faster QualitySNPng (Nijveen, Kaauwen, Esselink, Hoegen, & Vosman, 2013).

AGSNP (You et al., 2011) – Relies on Newbler assembler for clustering, and if strain information is not available, it further requires combining 454 data with *Illumina* or *SOliD* data (4Pipe4 does not require multiple technologies data for SNP calling).

Still other programs exist for SNP detection, but they usually require either a reference sequence, such as Atlas-SNP2 or SAMtools, or strain information, such as discoSnp++ (Uricaru et al., 2014) (formerly kisSnp (Peterlongo et al., 2010)) or DIAL (Ratan, Zhang, Hayes, Schuster, & Miller, 2010).

There is, however, another program that can be used for the same purpose as 4Pipe4 – QualitySNPng. This program, however is not an analysis pipeline, but rather a SNP caller for read alignments. It has a graphical user interface, which can be disabled for use in servers, but still requires "Qt4" to be installed in the server (which is not frequent). In order to compare it with 4Pipe4, we have modified the program to be usable without "Qt4" installed (https://github.com/StuntsPT/QualitySNP) and provide a branch of 4Pipe4 which is ready to use QualitySNPng (https://github.com/StuntsPT/4Pipe4/tree/new_snp_caller), without requiring any further dependencies.

Benchmarking the results of 4Pipe4 with QualitySNPng as the SNP caller, more SNPs were returned (513|147 SNPs found with the default|tuned values) than with our SNP caller, but with a larger rate of false positives (only 60|22 SNPs were a match to those found by AtlasSNP2, meaning a false positive rate of 88.3%|85%). Therefore, the builtin SNP caller was kept as default, but QualitySNPng can still be used from its own git branch if desired.

For the sake of completeness, we also made the SNP calling on the benchmark dataset using the software discoSnp++ (which requires strain identification) with the most restrictive parameters, to minimize the number of false positives. This program retrieved 9226 SNPs, of which 5967 were considered true positives (false positive rate of 35.3%).

As expected, this method retrieves more SNPs than both 4Pipe4 and QualitySNPng since it takes advantage of strain information, but it still provides a somewhat higher false positive rate than 4Pipe4.

5 Conclusions

We present here an automated analysis process specifically designed for SNP detection from 454 pyrosequencing transcriptome reads, which we named 4Pipe4. This is the first program specifically built to automate the whole process of finding putative SNPs in NGS datasets that lack both information regarding the origin of each read and a reference sequence. In-silico validation of 4Pipe4 results using previously analysed reference data revealed good performance in the calling of high confidence SNPs.

The 4Pipe4 pipeline, at the cost of retrieving a relatively low number of SNPs, has provided a lower rate of false positive SNPs than both an alternative SNP caller (QualitySNPng) and an alternative software that uses strain information (discoSnp++), as well as those obtained in previous studies that used different approaches for a similar type of data and goal.

Since the main purpose of this software is to retrieve high confidence SNPs for further exploring, we expect the incremental contributions it brings to improve, speed up and facilitate research on the field of population genomics.

Furthermore, we expect to implement new features in 4Pipe4, such as: graphics in the metrics report; indel variation finding; integration of alternative software (such as newbler for assembling instead of mira); process optimization for NGS technologies besides 454; switch from FASTA + FASTA.QUAL format to FASTQ. These are some of the planned features, but others can be requested and implemented, should there be demand for them.

6 Authors' contributions

FPM has developed the software, drafted the SNP detection routines and written the manuscript, BMV has extensively reviewed the code and assisted in the initial data analyses, SGS has provided valuable insights on SNP data interpretation and proofread the manuscript, DB has provided the samples and the datasets for the initial data analyses, and proofread the manuscript and OSP has conceived of the study, and participated in its design and coordination.

7 Acknowledgements

This work was fully supported by projects SOBREIRO/0036/2009 (under the framework of the Cork Oak ESTs Consortium), PTDC/BIA-BEC/098783/2008, and PTDC/AGR-GPL/119943/2010 from Fundação para a Ciência e Tecnologia (FCT) – Portugal. F. Pina-Martins was funded by FCT grant SFRH/BD/51411/2011, under the PhD program "Biology and Ecology of Global Changes", Univ. Aveiro & Univ. Lisbon, Portugal. D. Batista was funded by FCT grant SFRH/BPD/104629/2014. We would furthermore like to thank the reviewers of the manuscript, whom have provided very constructive criticism, resulting in great improvements to both the manuscript and the software.

8 References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). BASIC LOCAL ALIGNMENT SEARCH TOOL. *Journal of Molecular Biology*, *215*(3), 403–410.

Broders, K. D., Woeste, K. E., San Miguel, P. J., Westerman, R. P., & Boland, G. J. (2011). Discovery of single-nucleotide polymorphisms (SNPs) in the uncharacterized genome of the ascomycete Ophiognomonia clavigignenti-juglandacearum from 454 sequence data. *Molecular Ecology Resources*, *11*(4), 693–702. doi:10.1111/j.1755-0998.2011.02998.x

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Muller, W. E. G., Wetter, T., & Suhai, S. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, *14*(6), 1147–1159.

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, *21*(18), 3674–3676.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–9.

Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(Database issue), D19-21.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–9.

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

Nijveen, H., Kaauwen, M. van, Esselink, D. G., Hoegen, B., & Vosman, B. (2013). QualitySNPng: a user-friendly SNP detection and visualization tool. *Nucleic Acids Research*, *41*(W1), W587–W590. doi:10.1093/nar/gkt333

Papanicolaou, A., Stierli, R., Ffrench-Constant, R. H., & Heckel, D. G. (2009). Next generation transcriptomes for next generation genomes using est2assembly. *BMC Bioinformatics*, *10*, 447.

Peterlongo, P., Schnel, N., Pisanti, N., Sagot, M. F., & Lacroix, V. (2010). Identifying SNPs without a Reference Genome by Comparing Raw Reads, *6393*, 147–158.

pysam-developers/pysam. (n.d.). Retrieved March 13, 2015, from https://github.com/pysam-developers/pysam

Ratan, A., Zhang, Y., Hayes, V. M., Schuster, S. C., & Miller, W. (2010). Calling SNPs without a reference sequence. *BMC Bioinformatics*, *11*, 130.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European molecular biology open software suite. *Trends in Genetics : TIG*, *16*(6), 276–277.

Savage, A. E., Kiemnec-Tyburczy, K. M., Ellison, A. R., Fleischer, R. C., & Zamudio, K. R. (2014). Conservation and divergence in the frog immunome: pyrosequencing and de novo assembly of immune tissue transcriptomes. *Gene*, *542*(2), 98–108. doi:10.1016/j.gene.2014.03.051

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, *5*(1), 16–18.

Sequence Cleaner. (n.d.). Retrieved March 13, 2015, from http://sourceforge.net/projects/seqclean/

Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E. A., … Yu, F. (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Research*, *20*(2), 273–80.

Tang, J., Vosman, B., Voorrips, R. E., van der Linden, C. G., & Leunissen, J. A. M. (2006). QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics*, *7*, 438.

The UniVec Database. (n.d.). Retrieved March 13, 2015, from http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/

Tollenaere, C., Susi, H., Nokso-Koivisto, J., Koskinen, P., Tack, A., Auvinen, P., … Laine, A.-L. (2012). SNP Design from 454 Sequencing of Podosphaera plantaginis Transcriptome Reveals a Genetically Diverse Pathogen Metapopulation with High Levels of Mixed-Genotype Infection. *PLoS ONE*, *7*(12), e52492. doi:10.1371/journal.pone.0052492

Uricaru, R., Rizk, G., Lacroix, V., Quillery, E., Plantard, O., Chikhi, R., … Peterlongo, P. (2014). Reference-free detection of isolated SNPs. *Nucleic Acids Research*, *43*(2), e11–e11. doi:10.1093/nar/gku1187

You, F. M., Huo, N., Deal, K. R., Gu, Y. Q., Luo, M.-C., McGuire, P. E., … Anderson, O. D. (2011). Annotation-based genome-wide SNP discovery in the large and complex Aegilops tauschii genome using next-generation sequencing without a reference genome sequence. *BMC Genomics*, *12*, 59.

Chapter 3

NCBI Mass Sequence Downloader – Large dataset downloading made easy

1 Abstract

Sequence databases, such as NCBI are a very important resource in many areas of science. Downloading small amounts of sequences to local storage can easily be performed using any recent web browser, but downloading tenths of thousands of sequences is not as simple.

NCBI Mass Sequence Downloader is an open source program aimed at simplifying obtaining large amounts of sequence data from NCBI databases to local storage. It is written in python (can be run under both python 2 and python 3), and uses PyQt5 for the GUI. The program can be run in either graphical or command line mode.

Source code is licensed under the GPLv3, and is supported on Linux, Windows and Mac OSX. Available at https://github.com/StuntsPT/NCBI_Mass_Downloader.

2 Introduction

National Center for Biotechnology Information (NCBI) sequence databases are nowadays a resource of unquestionable importance for researchers in many areas of science (Miller, Norton, & Sarkar, 2009). The current count of sequences available in this database as of 15 December 2015 ascends to over $18.9x10^7$ sequences and $20.3x10^{10}$ base pairs (ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt), representing roughly 2.5 Tb of compressed data and keeps growing. Due to advances in sequencing technology, the amount of sequence data required by investigators has increased by orders of magnitude in the last few years. This has naturally led to increased use of the NCBI databases by investigators for retrieving sequence data.

NCBI Mass Sequence Downloader provides a user friendly interface and automated error checking for downloading large sets of sequence data.

3 Problems and Background

Although downloading sequences from NCBI can be done in a simple fashion using any standards compliant web browser via the *Entrez* (Sayers et al., 2010) web portal, this method does not scale well, and downloading large amounts of sequences (in the order of the tenths of thousand) from these databases can cause problems when performed this way (https://www.biostars.org/p/43970/). Furthermore, manually performing this type of tasks is time consuming and error prone, which may hamper the reproducibility of scientific work.

For retrieving large sets of data, NCBI provides the *E-utilities* API (Sayers et al., 2010), although it can be difficult to use by investigators without an IT related background,

despite it's through documentation. Frameworks exist, written in various popular languages, such as *Python*, *Perl* or *Ruby*, that provide some level of abstraction for using this API, such as *Biopython* (Cock et al., 2009), *BioPerl* (Stajich et al., 2002), or *BioRuby* (Goto et al., 2010), respectively. However, these too, require some degree of programming knowledge to use, rather than providing end-user packages ready to use for a specific purpose. This leaves investigators without a simple, ready made solution. Although this is not much of a problem for someone with a bioinformatics background, it poses a serious issue for someone with a molecular biology background, who may frequently require this kind of data, but lack the programming skills to use one of the mentioned frameworks or the API.

By using NCBI's API, our program intends to solve the problem of retrieving large datasets, in a user friendly, automated, and reproducible way. The tool is therefore aimed at molecular biologists that don't have an IT related background, but need to download large datasets from the NCBI databases.

4 Software Framework

4.1 Software Architecture

NCBI Mass Sequence Downloader is written in python (http://www.python.org) and can be run under both python 2 and python 3. The command line interface (CLI) version of the program can be run on any OS that has python available. The GUI version further requires PyQt5 (http://www.riverbankcomputing.com/software/pyqt/intro) available, which means all major currently used operating systems such as GNU/Linux, MS Windows and Mac OSX are supported. The program uses a slightly altered (changed the import statements) module from *Biopython* (Cock et al., 2009) – *Entrez* (Sayers et al., 2010), which is included with the software. This avoids needing to have *Biopython* installed, which, despite being a popular library in the bioinformatics community, is not usually so for molecular biologists. The consequence of this convenience for the user is a higher maintenance requirement since it makes it necessary to keep up with the upstream *Entrez* module. However, this *Biopython* module has not had many recent changes, and merging them into *NCBI Mass Sequence Downloader* has so far, been trivial.

The program consists of essentially three modules – a back-end, a front-end and the *Entrez* module. If the program is run without arguments, the GUI version is launched (Figure 3.1), but if the program is run with arguments, the command line version will be run instead. This makes the program quite flexible to use in different environments.

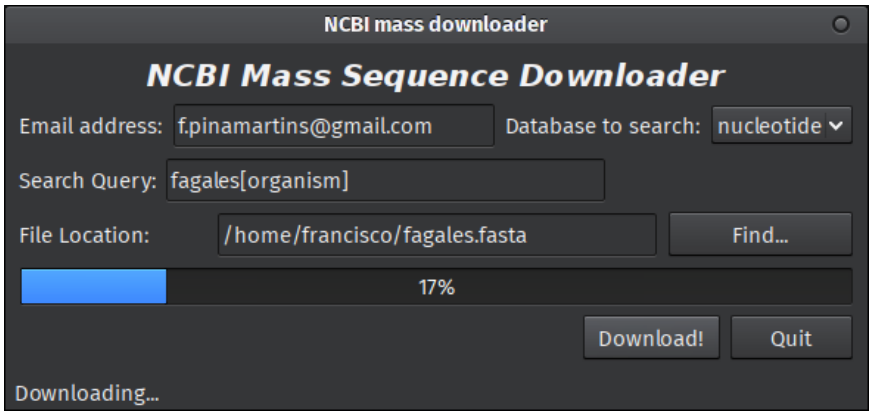

Figure 3.1: A screenshot of NCBI Mass Sequence Downloader running under a GNU/Linux based OS in graphical user interface, downloading sequences matching the query "Fagales[organism]".

The program's source code is available on github (https://github.com/StuntsPT/NCBI_Mass_Downloader), along with binaries for GNU/Linux, MS Windows and Apple OSX.

4.2 Software Functionalities and Limitations

NCBI Mass Sequence Downloader is made to solve a single task – downloading sets of sequences from the NCBI databases. For this, the user should provide an email address for eventual contact from NCBI (which is sent only to NCBI), the database to be queried, the search query, and a path to the file for the downloaded sequences. Download progress is indicated in both user interfaces.

Currently, the program is limited to downloading sequences in the FASTA format and to NCBI databases, but data from several databases can be retrieved: *nucleotide, nuccore, nucgss, protein, genome* and *popset*.

4.3 Internal Routines and Error Handling

Once the program is requested to start the download, it queries the selected NCBI database for the provided search term. It will then store the returned sequence IDs in memory, and begin downloading the respective records in batches of 3000 sequences. Every batch is temporarily stored in memory, and once 3000 sequences are downloaded, they are immediately stored in the output file, flushed from memory, and then, the next batch is processed.

After all the records are downloaded, the output file is parsed and it's sequences' IDs matched to the originally retrieved sequence IDs. If any sequences are missing, a new pass is made, to retrieve them. This process is repeated as often as necessary until all the requested sequences are stored in the output file. Stopping the program at any time will not affect any sequences already stored in the output file.

If a pre-existing FASTA file is selected as the output file, instead of overwriting it, the file is parsed, and the sequences' IDs are retrieved and compared to those returned by NCBI for the requested query. Any already present sequences are not downloaded again, and any missing sequences are appended to the end of the file. This capability makes it possible to resume any cancelled download.

NCBI Mass Sequence Downloader will handle any server errors thrown during sequence retrieval by pausing all activity for eight seconds and then retrying. Five such failures in a row, cause a further 20 second pause before trying a retrieval operation again.

4.4 Future Plans

Several developments are expected for future releases of *NCBI Mass Sequence Downloader*, such as being able to get data in formats other than FASTA, adding an online interactive help system to the GUI or even the capability to query databases other than NCBI. We expect to keep the software maintained to work with future versions of python, Qt, and database APIs for the foreseeable future.

5 Illustrative Examples

5.1 Example use case

A molecular biologist has to analyse a hypothetical dataset of transcriptomic data of a plant-fungus system (*Castanea dentata*, *Cryphonectria parasitica*). In order to identify which sequences can be considered “plant” and which can be considered “fungus”, instead of downloading the entire “nt” database from NCBI and running BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) queries against it, by using *NCBI Mass Sequence Downloader*, it is possible to download only the sequences of the *Fagales* (plants) order and *Sordariomycetidae* (fungus) subclass, and run the required BLAST queries against the resulting files. This would considerably reduce both download and query time, provide the user with more specific results and enable a simpler downstream data filtering process.

An example study where *NCBI Mass Sequence Downloader* could have been useful, is (Haçarız, Akgün, Kavak, Yüksel, & Sağıroğlu, 2015), where the investigators performed BLAST searches against several data sets, that could have been quickly obtained and segregated with this software.

5.2 The command line interface (CLI)

In order to use the CLI version of the program for solving the problem described in 4.1 the user needs to run the program with the following arguments: “user email address”,

"query term", "database to query" and "output file". *Screenshots of NCBI Mass Sequence Downloader* performing this specific task in the CLI environment can be seen in Figure 3.2.

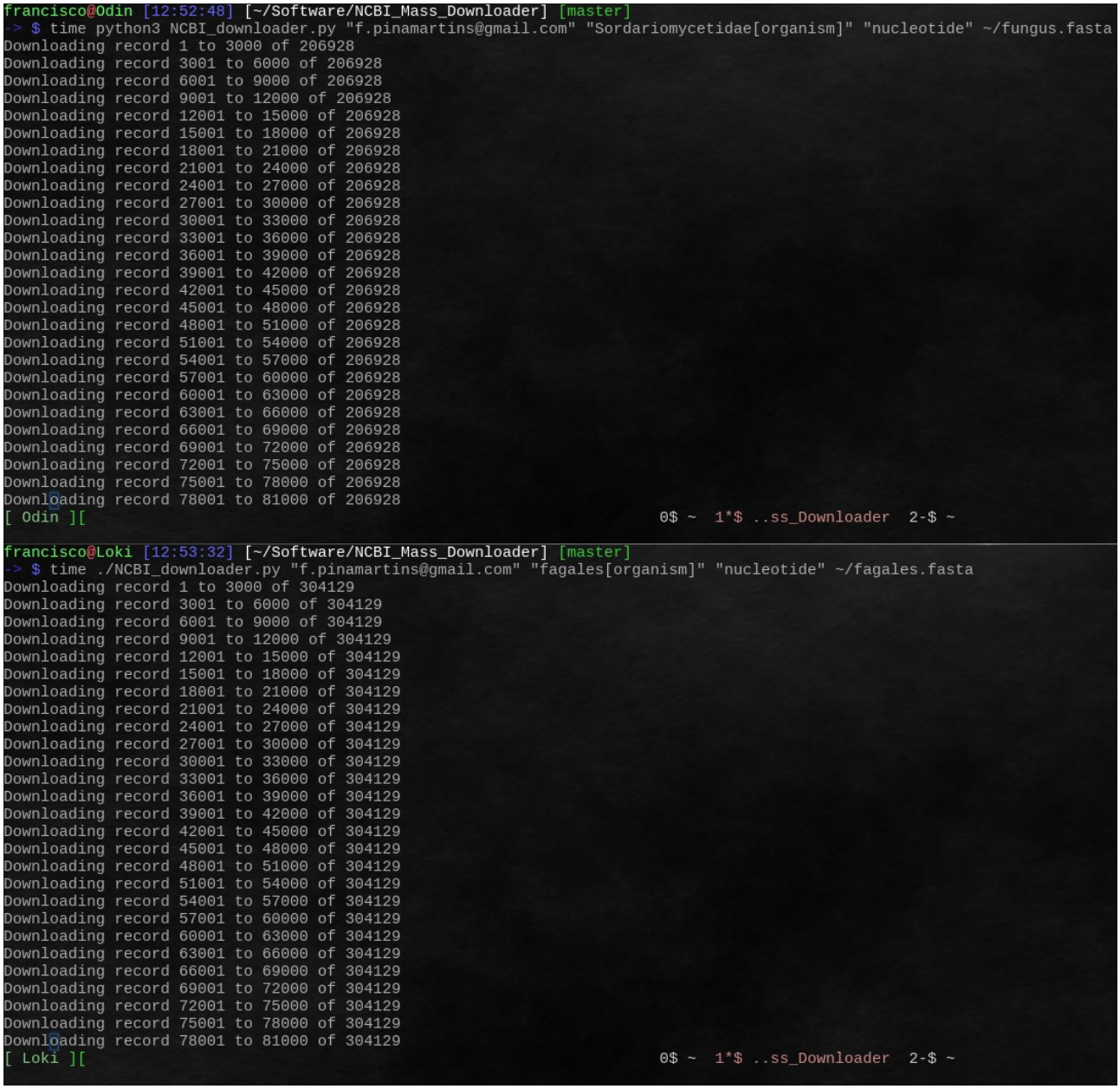

Figure 3.2: A screenshot of NCBI Mass Sequence Downloader running under a GNU/Linux based OS in command line interface, downloading sequences matching the queries "Sordariomycetidae[organism]" (above) and "Fagales[organism]" (below).

In this test, the plant query took ~48 minutes to download all 304129 records, roughly 1.2 GB of sequence data. The fungi query took ~42 minutes to download all 206928 records, amounting to approximately 2.0 GB of sequences.

5.3 The Graphical User Interface (GUI)

In order to use the GUI version of the program, the user needs to run the program without any arguments. This will bring up the interface main window (Figure 3.1), where the user can enter the required information to proceed with the downloading of the queried sequences.

The performance of the GUI version was essentially the same as the one obtained using the CLI method.

5.4 Using the alternative methods

Using the *Entrez* web portal to download the sequences mentioned in the example resulted in having to attempt each of the downloads several times until all the sequences were obtained. Furthermore, the download would simply stop without issuing any error messages, and it was thus, necessary to manually verify that all the requested sequences had been downloaded (which didn't happen in the first three tries for the plant dataset and for the first two times for the fungus dataset). This method, was thus more time consuming and required manual user intervention several times until all the requested data was locally stored.

The E-Utilities API can also be used directly. In order get the example data using this method, the following actions need to be taken:

1. Make the search query to the NCBI servers
2. Retrieve the “Query Key” and “WebEnv” variables
3. Request the sequences in blocks of up to 10^4 until all are downloaded

This can be done manually, but it is a tedious and error prone process (step 3 would have to be performed 62 times to download all sequences from the example case). Alternatively, this behaviour can be scripted to automates the process, but that requires programming skills, which may act as a barrier to molecular biologists.

6 Conclusions

Although querying the NCBI database and downloading the respective sequences can usually be done from the web browser, when it is necessary to download large amounts of sequences, this procedure becomes unreliable since the probability of download problems increases with it's size and the *Entrez* web portal does not provide a way to resume interrupted downloads. Using the alternate method – via the *E-utilities* API requires programming skills and not every molecular biologist is equipped to deal with that. These issues make the process of retrieving large datasets from NCBI an error prone and attention demanding process, unless the user has some programming skills.

NCBI Mass Sequence Downloader was designed to fill in this gap. To allow anyone without programming skills to easily download large sequence datasets from the NCBI databases, in an automated, reliable and reproducible way.

Furthermore, the possibility to choose the interface, makes *NCBI Mass Sequence Downloader* appropriate to use both on desktop and on the command line based systems.

7 Acknowledgements

This study was financed by Portuguese National Funds, through FCT – Fundação para a Ciência e a Tecnologia, within the projects UID/BIA/00329/2013, SOBREIRO/0036/2009 and SFRH/BD/51411/2011.

8 References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). BASIC LOCAL ALIGNMENT SEARCH TOOL. *Journal of Molecular Biology*, *215*(3), 403–410.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, *25*(11), 1422–3.

Goto, N., Prins, P., Nakao, M., Bonnal, R., Aerts, J., & Katayama, T. (2010). BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics (Oxford, England)*, *26*(20), 2617–2619. doi:10.1093/bioinformatics/btq475

Haçarız, O., Akgün, M., Kavak, P., Yüksel, B., & Sağıroğlu, M. Ş. (2015). Comparative transcriptome profiling approach to glean virulence and immunomodulation-related genes of Fasciola hepatica. *BMC Genomics*, *16*(1), 366. doi:10.1186/s12864-015-1539-8

Miller, H., Norton, C. N., & Sarkar, I. N. (2009). GenBank and PubMed: How connected are they?, *2*, 101.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., … Ye, J. (2010). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *38*(Database issue), D5-16.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., … Birney, E. (2002). The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, *12*(10), 1611–1618. doi:10.1101/gr.361602

Chapter 4

Structure_threader: An improved method for automation and parallelization of programs STRUCTURE, FASTSTRUCTURE and *MavericK* on multi core CPU systems

1 Abstract

Structure_threader is a program to parallelize multiple runs of genetic clustering software that does not make use of multi-threading technology (STRUCTURE, FASTSTRUCTURE and *MavericK*) on multi-core computers. Our approach was benchmarked across multiple systems and displayed great speed improvements relative to the single threaded implementation, scaling very close to linearly with the number of physical cores used.

Structure_threader was compared to previous software written for the same task - *ParallelStructure* and *StrAuto*, and was proven to be the faster (up to 25% faster) wrapper under all tested scenarios.

Furthermore, *Structure_threader* can perform several automatic and convenient operations, assisting the user in assessing the most biologically likely value of 'K' via implementations such as the "Evanno", or "Thermodynamic Integration" tests and automatically draw the "meanQ" plots (static or interactive) for each value of K (or even combined plots).

Structure_threader is written in python 3 and licensed under the GPLv3. It can be downloaded free of charge at https://github.com/StuntsPT/Structure_threader.

2 Introduction

Clustering analyses are widely used in population genetics and, nowadays, population genomics. This technique of using multilocus genotype data to infer population clusters, is frequently performed based on multiple MCMC re-sampling. One of the most popular tools for performing this type of analyses is structure (Pritchard, Stephens, & Donnelly, 2000). Despite producing robust results, this approach demands long run times, even in modern machines. This problem is aggravated as the type of analysed datasets, which gradually grow from relatively small, such as microsatellite loci (De Barro, 2005; Muchadeyi et al., 2007), to high throughput sequencing (Lamaze, Sauvage, Marie, Garant, & Bernatchez, 2012; Renaut, Grassa, Moyers, Kane, & Rieseberg, 2012), consequently increasing run times by orders of magnitude.

The process can be sped up by either running multiple instances of the used software, which is an inefficient and error prone method requiring constant attention and intervention from the user. There are faster software alternatives to STRUCTURE, which can also be used to speed up the analysis process.

One such option is analysing the data in the program FASTSTRUCTURE (Raj, Stephens, & Pritchard, 2014), which decreases run times by up to two orders of magnitude. However FASTSTRUCTURE does not support the popular "no admixture" model present in STRUCTURE,

and is not capable of handling haploid data (and several other less widely used features), which limits its application to a wide range of data.

Another option in to analyse the data in the software *MavericK* (Verity & Nichols, 2016), which is also considerably faster than STRUCTURE, but not as fast as FASTSTRUCTURE by an order of magnitude. It does, however support most of the same features as STRUCTURE and uses a built-in, improved method for helping determine the most biologically relevant value of "K" called "Thermodynamic Integration" (Verity & Nichols, 2016). Regardless of the speed gains these programs offer, they are only able to use a single CPU core for their computations, which means that these methods too, do not scale well with current multi-core IT infrastructure.

Alternatively, a method to bootstrap multiple simultaneous runs of the software can be used, such as the R (R Core Team, 2013) package *ParallelStructure* (Besnier & Glover, 2013), or *StrAuto* (Chhatre & Emerson, 2017), which does exactly that for the software STRUCTURE (Pritchard et al., 2000). *ParallelStructure*, however, has scaling problems, as described in the manuscript, considerably loosing efficiency as more CPU cores are used. *StrAuto* is another option that does indeed scale well with the number of CPU cores used, but like *ParallelStructure*, it only works as a wrapper for the software STRUCTURE, and cannot be used to speed up other popular genetic clustering programs.

Furthermore, after the clustering step is finished, it is necessary to infer the number of clusters that make most biological sense for the data (Earl & vonHoldt, 2012), using methods such as the "Evanno test" (Evanno, Regnaut, & Goudet, 2005), or the "Thermodynamic Integration" (TI) method (Verity & Nichols, 2016). After this, it is also often necessary to plot the "meanQ" values of each cluster per individual, to be able to interpret the biological significance of the data. This is usually done with software such as DISTRUCT (Rosenberg, 2004).

All of these steps typically require parsing the results files of each clustering run and manually running all the required steps until the final outcome is produced (Earl & vonHoldt, 2012). This is not only time consuming as it is also error prone due to the large number of separate steps that must be taken during the process. Neither *ParallelStructure* nor *StrAuto* provide an automated and reproducible way to perform this task.

Part of this process is largely facilitated by the program STRUCTURE HARVESTER (Earl & vonHoldt, 2012), which automates the parsing of STRUCTURE runs and uses that information to perform an "Evanno test" on the data, which uses some heuristics to predict which value of 'K' makes the most biological sense regarding the analysed data. Although this is a very convenient automation, it still relies on manual user intervention to input the data from STRUCTURE, does not provide assistance with the plotting of the "meanQ" values and only works for the software STRUCTURE. Other programs, such as

FASTSTRUCTURE include the necessary software to perform these tests, and even to plot the "meanQ" values, but still require manual intervention between these steps. *MavericK* goes further and presents the full posterior distribution for 'K' using the "Thermodynamic Integration" test as an automatic last step of the analysis and even recommends some scripts for drawing "meanQ" plots, but this last step also requires human intervention.

To address these two problems (reducing run times and automating the analyses tasks), we herein present *Structure_threader:* a program to parallelize STRUCTURE, FASTSTRUCTURE and *MavericK* runs that considerably reduces the scaling problems of previous approaches and automates the entire process, - wrapping the runs, assisting in the choice of the most biologically relevant value of K, and drawing the "meanQ" plots.

Structure_threader is available on https://github.com/StuntsPT/Structure_threader. For the stable (non development) versions, check the releases page, or get it from Pypi.

3 Materials & Methods

3.1 Program description

Structure_threader, licensed under the GPLv3, is an open source program written in python (https://www.python.org/) that automates and parallelizes genetic clustering software (STRUCTURE, FASTSTRUCTURE and *MavericK*) runs.

The software was written according the "Best Practices in the Development of Bioinformatics Software" (Leprevost, Barbosa, Francisco, Perez-Riverol, & Carvalho, 2014) and can be run on any platform where python is available, such as GNU/Linux, Mac OS X and Microsoft Windows (other platforms may also work, but were not tested). Additional details are available in the program's documentation.

All options supported by the wrapped programs can be passed to *Structure_threader* as command line arguments. These are explained in detail in both the program's online documentation and builtin help text.

Parameters are passed to the wrapped software as in their default implementations – all wrapped programs take arguments from the command line, STRUCTURE also reads settings from the files "mainparams" and "extraparams" and *MavericK* from "parameters.txt".

After performing the parallelized runs of the wrapped software, *Structure_threader* runs a slightly modified (for integrating with *Structure_threader*) version of STRUCTURE HARVESTER *chooseK.py* (Raj et al., 2014) or TI for helping identify the most biologically relevant value of 'K' for any given dataset.

Finally, *Structure_threader* parses the results files and draws the “meanQ” plots for each considered value of “K”. These are drawn in both an interactive version for visualization and in a static version, better suited for publication.

Example data files and results are provided in the program's repository.

3.2 Threading strategy

The threading strategy used in *Structure_threader* is represented in Figure 4.1. *Structure_threader* takes the provided input file, the values of “K” to test and the required number of replicates, and creates a job queue, which is sorted by decreasing complexity order. After this, *P* child processes are spawned, (where *P* is the number of threads made available to the software) each containing one independent instance of the wrapped program. Each of these child processes takes the first available job from the queue and once it is finished, its output is processed by the main process for error handling and logging. The child processes are spawned using python’s “multiprocessing” and “subprocess” modules from the standard library.

3.3 Benchmarking process

In order to assess the gains provided by *Structure_threader*, the program was benchmarked in four different systems, described in Table 4.1, with various specifications. Runs were performed twice to serve as replicates (Appendix I Table 1). Run times for STRUCTURE were assessed using both *Structure_threader* v0.4.1, *ParallelStructure* v1.0 and *StrAuto* v1.0, which were then compared. FASTSTRUCTURE and *MavericK* runs were only wrapped in *Structure_threader*, since none of the other programs supports this, and compared with the default, single threaded implementation.

Table 4.1: Characteristics of the systems where the programs were benchmarked, along with the run time of the single threaded run.

System Name	CPU				OS	STRUCTURE single thread run time (s)	FASTSTRUCTURE single thread run time (s)	MavericK single thread run time (s)
	Type	Frequency Base/Turbo (GHz)	Physical cores	Logical cores				
Haswell Laptop	i7 4700MQ	2.4/3.4	4	8	ArchLinux	9668	3140	1009
Ivy Bridge Desktop	i5 3350P	3.1/3.3	4	4	ArchLinux	10926	2854	1140
Nehalem Rack	Xeon E5520x2	2.26/2.53	8	16	Ubuntu 16.04	16000	6019	1835
Sandy Bridge Rack	Xeon E5-2609x2	2.4	8	8	Ubuntu 12.04	15805	5054	1711

Usage of RAM was monitored during the benchmarking process, and it was never detected as a bottleneck on any of the systems. None of the wrapped programs is very I/O intensive (at least as far as the tested systems were concerned), meaning that the present tests were always CPU bound.

Run times were measured using *zsh's time* builtin method (wall time was used), and then normalized to a "speed up" factor (Besnier & Glover, 2013) by dividing the time of the multi-core runs by the time of the single core runs, which were performed in the measured programs' default implementations.

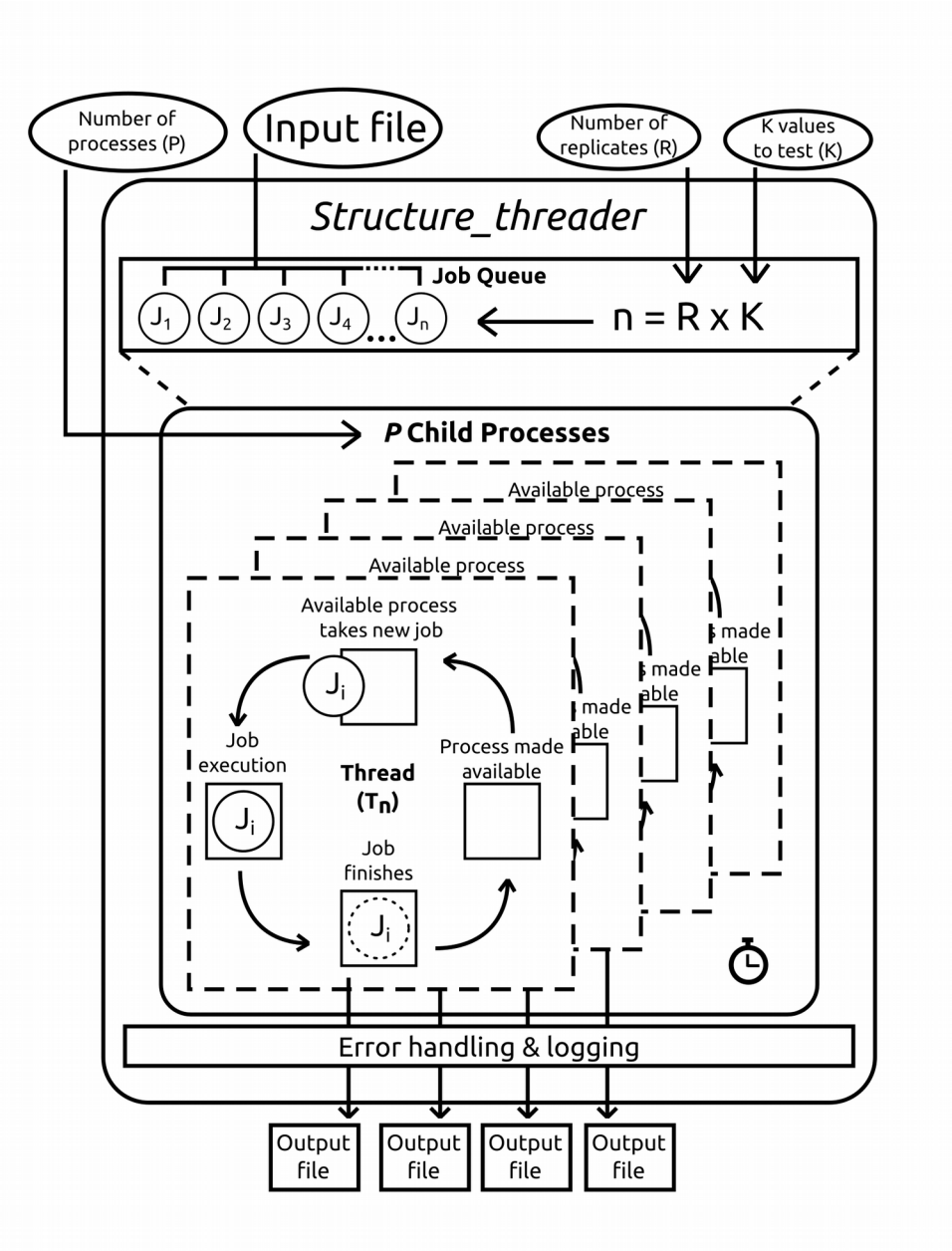

Figure 4.1: Threading strategy used in *Structure_threader*. Values in ellipses are read from the command line and passed to the main process, which generates a job queue. The jobs in the queue are then processed by the spawned child processes. The main process is responsible for handling and logging any errors that occur in the child processes.

3.4 Benchmarking STRUCTURE

3.4.1 Test dataset description

The test file used for the benchmarks consists of 100 individuals, represented by 80 SNP loci without missing data. This dataset was crafted based on data from the *1000 genomes project* (The 1000 Genomes Project Consortium, 2015) to perform the benchmarks and was constructed aiming for a size that would be neither too small, which could bias the benchmarking towards very quick runs, nor too large, to avoid the process taking too long to be practical.

This dataset was created from public data, and instructions on how to recreate it are available in the documentation and in the program's repository.

3.4.2 Benchmark details

The benchmarking process consisted of running the test dataset on STRUCTURE v2.3.4 for $1x10^6$ MCMC iterations, and a "burnin" length of $5x10^4$, under "admixture" model (for other parameters check the program's repository). These settings were performed for values of "K" varying from 1 to 4. Each value of "K" was run with 4 replicates, which means a total of 16 STRUCTURE runs were performed in each benchmark. All these runs were performed on the default, single threaded implementation and under the *Structure_threader*, *ParallelStructure* (using the "parallel_structure()" implementation, which initial testing found to be faster in all used machines) and *StrAuto* wrappers.

3.5 Benchmarking FASTSTRUCTURE

3.5.1 Test dataset description

The test dataset used for benchmarking FASTSTRUCTURE runs, is different from the one used for benchmarking STRUCTURE, since this program was designed to analyse larger datasets. The tested file consists of 1000 individuals, represented by 1000 SNP loci. Like the previous dataset, this one was also crafted from the same public data from the 1000 genomes project, and instructions for recreating it are available in the documentation. The used dataset itself is available in the program's repository.

3.5.2 Benchmark details

The benchmarking process consisted of running the above described dataset for values of "K" from 1 to 16 for each benchmark run.

The average run time of both replicates was used to plot and analyse the data. Since a FASTSTRUCTURE runs do not require replicates for downstream analyses, each value of "K" was run only once per benchmark, which means that a total of 16 FASTSTRUCTURE runs

were performed both in the default implementation and under the *Structure_threader* wrapper.

3.6 Benchmarking *MavericK*

3.6.1 Test dataset description

The test file used for the *MavericK* benchmarks is the same that was used to benchmark STRUCTURE, which is described above.

3.6.2 Benchmark details

The benchmarking process consisted of running the test dataset on *MavericK* v1.0.4 for $1x10^4$ MCMC iterations, plus a "burnin" length of 2500 iterations, with 5 replicates each (for other parameters check the program's repository). These settings were performed for values of "K" varying from 1 to 16.

4 Results & Discussion

Using *Structure_threader* as a wrapper for all programs has yielded increases in speed that scale almost linearly with the number of processes used, at least as long as physical cores are concerned, as can be seen in Figure 4.2 and Figure 4.3.

Considering the benchmark results in Figure 4.2, it is clear that both *Structure_threader* and *StrAuto* are more efficient methods to perform STRUCTURE runs on multiple core systems than *ParallelStructure* (on average 7% faster in the tested systems, varying from 1% to 25% faster). *Structure_threader* is also always faster than *StrAuto*, but by much smaller margins than when compared with *ParallelStructure* (on average 3% faster, varying from 0.3% to 7% faster). Regardless of the tested system and number of cores used, the differences in "speed up" are always in favour of *Structure_threader*. When compared to *ParallelStructure*, the difference increases with the requested scaling – the more physical cores are used, the better the relative performance of *Structure_threader*. Also worth noting is that the "speed up" values obtained here with *ParallelStructure* when using physical cores, are somewhat better than what is described in (Besnier & Glover, 2013), but this could be due to differences in benchmark workloads.

Speed up differences between *StrAuto* and *Structure_threader* are small, but can be compared in Figure 4.2. A more detailed comparison, can be made using the data tables available in Supplementary Material 4.1.

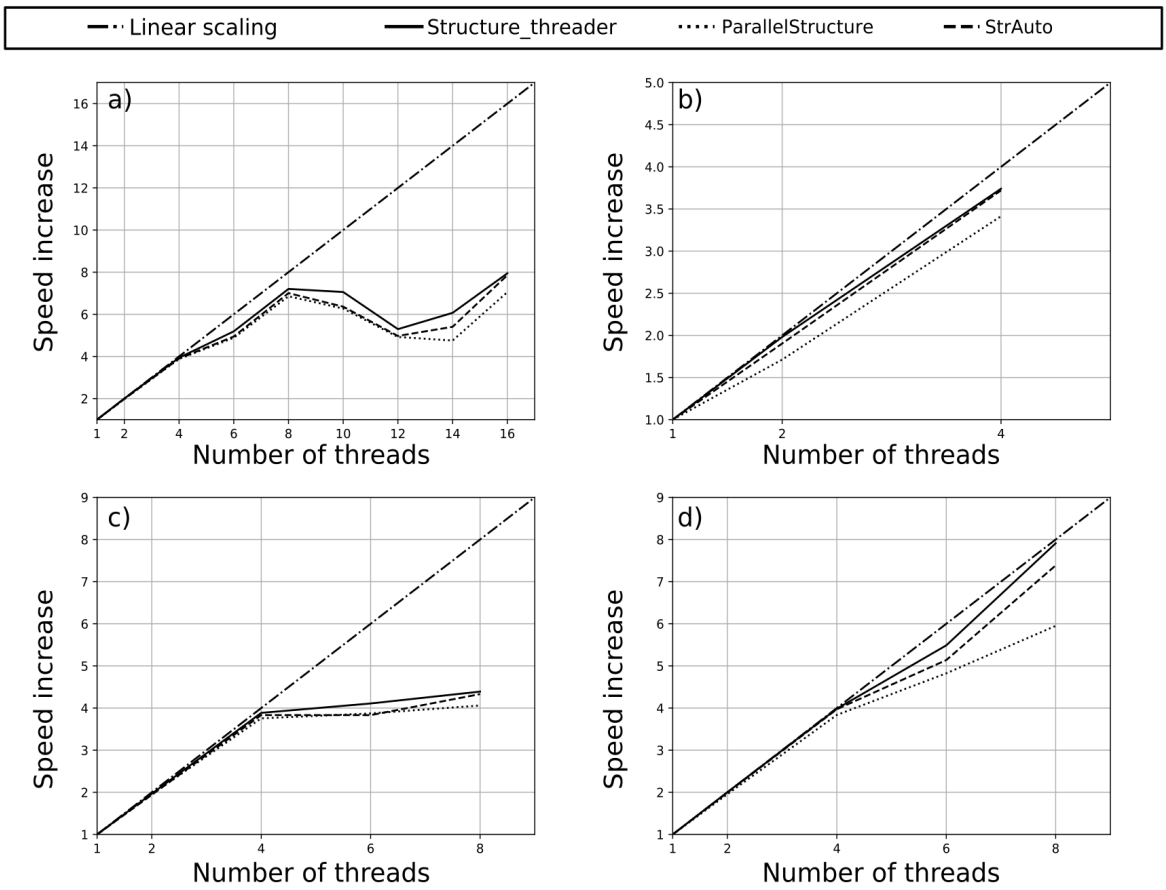

Figure 4.2: Plots of the STRUCTURE "speed up" as more threads are used in *Structure_threader*, *ParallelStructure* and *StrAuto*. Each plot represents a different system – a) is "Nehalem Rack", b) is "Ivy Bridge Desktop", c) is "Haswell Laptop" and d) is "Sandy Bridge Rack".

Unlike *ParallelStructure* and *StrAuto*, *Structure_threader* can also speed up the runs of other programs by running them in parallel. Similar to what is done for running STRUCTURE, wrapping FASTSTRUCTURE and *MavericK* in *Structure_threader*, provides considerable speed improvements, once again scaling almost linearly as long as *hyper-threading* is not in effect (Figure 4.3).

Although ideally the "speed up" factor should scale linearly with the number of used physical cores, this does not always happen in practice (Figure 4.2 and Figure 4.3). Of the three tested wrappers, *Structure_threader* scales the closest to linearly, even when using 8 physical cores, where the "speed up" factor varies between 6.24 and 7.91, depending on the system and the wrapped program. *ParallelStructure* shows the worst scaling of the tested wrappers, especially on 8 physical threads, where the "speed up" factor varies between 5.95 and 6.85.

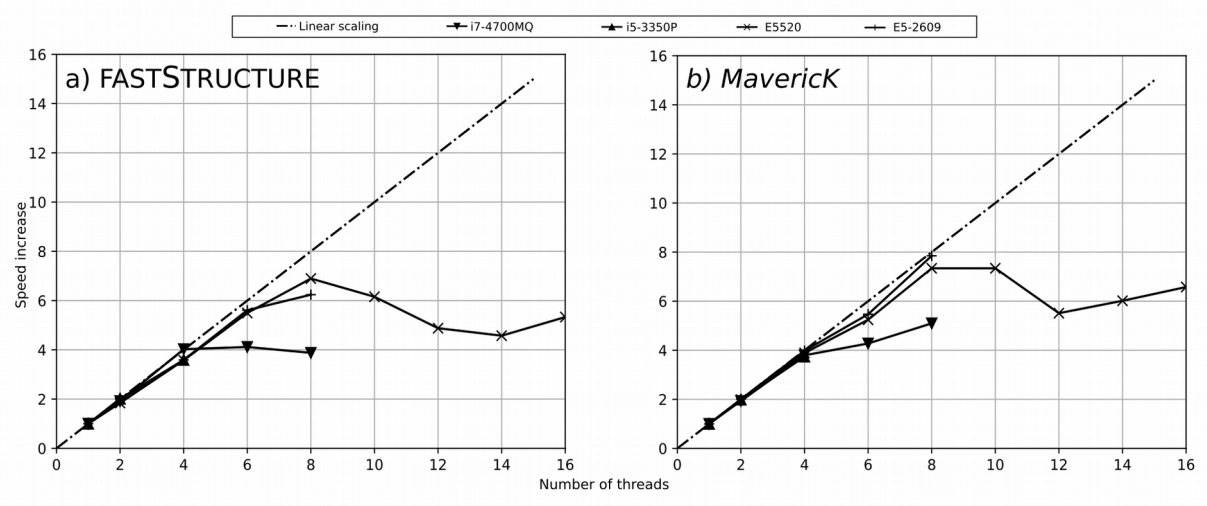

Figure 4.3: Plots of FASTSTRUCTURE and *MavericK* "speed up" as more threads are used when wrapped under *Structure_threader*.

The large drop in performance increase, regardless of the used wrapper program, happens when using hyper-threading (using more than eight cores in the *Nehalem Rack* system and more than four in the *Haswell Desktop* system – the CPUs of the other two systems do not have this feature), as is sometimes described under certain workloads (Leng, Ali, Hsieh, Mashayekhi, & Rooholamini, 2002; Marr et al., 2002; Saini et al., 2011). We are not sure why this happens on this particular workload, but the issue is not as evident when analysing smaller datasets as the one from (Besnier & Glover, 2013). It is therefore hypothesised that it may be related to "cache thrashing", a phenomenon that occurs when the CPU constantly refreshes the contents of L2 and L3 caches for quickly accessing different information. "Cache thrashing" is more likely to happen when working with larger datasets and when hyper-threading is active since both logical cores share L2 and L3 cache (Marr et al., 2002).

The automated plot drawing feature of *Structure_threader* is responsible for both a simplification of the analysis process (less steps per analysis), and also for the reduction in random error (consequence of less human intervention).

The mentioned plots produced by *Structure_threader* are provided in two formats. A static, vectorial image in "svg" format, especially suited for publication, and an interactive HTML version of the plot, suited for results exploration.

5 Conclusions

The observed difference in efficiency between *Structure_threader* and *ParallelStructure* can probably be explained by the programming languages utilized in the wrappers (Python vs. R) and the fact that *ParallelStructure* solves tasks in increasing order of complexity, whereas *Structure_threader* sorts them in decreasing order. This strategy

provides an increase in efficiency, since the sorting minimizes the time each CPU core is left idle.

Another important difference between *ParallelStructure* and *Structure_threader* is that the former is a framework to build scripts that perform the requested analyses, and the latter can either be used as a framework, or directly from the command line. This makes *Structure_threader* much easier to use, while simultaneously keeping the same type of flexibility *ParallelStructure* offers. Although both programs can be used to draw the clustering plots from the STRUCTURE results, the features offered by *Structure_threader* go far beyond the basic plotting that *ParallelStructure* is capable of.

The speed gains obtained with *Structure_threader* and *StrAuto* are very similar, with only a marginal difference favouring *Structure_threader*. This difference is likely due to the efficiency of *python*'s higher speed when compared to *bash*'s, and eventually due to a smaller overhead of *python*'s *multiprocess* module when compared to that of GNU parallel (Tange, 2011). Although both programs are run from the command line interface, *Structure_threader* is more user friendly than *StrAuto*, since it includes built-in help, handles user errors, and allows for a lot of parameters to be defined directly in the command line.

Structure_threader was designed to exploit the power of multi-core machines for speeding up multiple genetic clustering software runs, with emphasis on scalability. Our results demonstrate that in every tested scenario this goal is fulfilled in a more efficient way than previous solutions.

Furthermore, *Structure_threader* goes much farther than the two previous solutions in it's capabilities to perform tests for estimating the most biologically relevant "K" value, as well as plotting flexibility.

Although the automation process that *Structure_threader* provides does not decrease computation time, it should significantly speed up the analyses process, due to the human time that is saved. Furthermore, this automation is also one important step for reproducibility of the studies that use this software. That being said, it is also important that users interact with and explore the options and parametrization the wrapped programs offer. It is critical that these are well understood in order to obtain meaningful and statistically relevant results.

We find that the obtained decrease in run times, allied with the ease of use and automation, including that of follow up analysis, make *Structure_threader* a useful tool to any investigator working with population genetics/genomics data and the best current choice for performing genetic clustering analyses.

6 Acknowledgements

This study was financed by Portuguese National Funds, through FCT – Fundação para a Ciência e a Tecnologia, within the projects UID/BIA/00329/2013, SFRH/BD/51411/2011 and SFRH/BD/86736/2012.

We would further like to thank Bob Verity and an anonymous reviewer for their suggestions that thoroughly improved both the software and the manuscript.

7 Author Contributions

F. Pina-Martins has conceived the concept of the study, written most of the program code and written the manuscript. D. Silva has contributed to the ideas of the software, written code and revised the manuscript. J. Fino has contributed to the ideas of the software, written code and revised the manuscript. O. S. Paulo has mentored the work and revised the manuscript.

8 References

Besnier, F., & Glover, K. A. (2013). ParallelStructure: A R Package to Distribute Parallel Runs of the Population Genetics Program STRUCTURE on Multi-Core Computers. *PLoS ONE*, *8*(7). doi:10.1371/journal.pone.0070651

Chhatre, V. E., & Emerson, K. J. (2017). StrAuto: automation and parallelization of STRUCTURE analysis. *BMC Bioinformatics*, *18*, 192. doi:10.1186/s12859-017-1593-0

De Barro, P. J. (2005). Genetic structure of the whitefly Bemisia tabaci in the Asia–Pacific region revealed using microsatellite markers. *Molecular Ecology*, *14*(12), 3695–3718. doi:10.1111/j.1365-294X.2005.02700.x

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

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, *14*(8), 2611–2620. doi:10.1111/j.1365-294X.2005.02553.x

Lamaze, F. C., Sauvage, C., Marie, A., Garant, D., & Bernatchez, L. (2012). Dynamics of introgressive hybridization assessed by SNP population genomics of coding genes in stocked brook charr (Salvelinus fontinalis). *Molecular Ecology*, *21*(12), 2877–2895. doi:10.1111/j.1365-294X.2012.05579.x

Leng, T., Ali, R., Hsieh, J., Mashayekhi, V., & Rooholamini, R. (2002). An Empirical Study of Hyper-Threading in High Performance Computing Clusters. In *Linux Clusters: The HPC Revolution, 2002*. St. Petersburg.

Leprevost, F. da V., Barbosa, V. C., Francisco, E. L., Perez-Riverol, Y., & Carvalho, P. C. (2014). On best practices in the development of bioinformatics software. *Bioinformatics and Computational Biology*, *5*, 199. doi:10.3389/fgene.2014.00199

Marr, D., Binns, F., Hill, D. L., Hinton, G., Koufty, D. A., Miller, A. J., & Upton, M. (2002). Hyper-Threading Technology Architecture and Microarchitecture. *Intel Technology Journal*, *6*. Retrieved from https://www.researchgate.net/publication/237005389_Hyper-Threading_Technology_Architecture_and_Microarchitecture

Muchadeyi, F. C., Eding, H., Wollny, C. B. A., Groeneveld, E., Makuza, S. M., Shamseldin, R., … Weigend, S. (2007). Absence of population substructuring in Zimbabwe chicken ecotypes inferred using microsatellite analysis. *Animal Genetics*, *38*(4), 332–339. doi:10.1111/j.1365-2052.2007.01606.x

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

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, *197*(2), 573–589. doi:10.1534/genetics.114.164350

Renaut, S., Grassa, C. J., Moyers, B. T., Kane, N. C., & Rieseberg, L. H. (2012). The Population Genomics of Sunflowers and Genomic Determinants of Protein Evolution Revealed by RNAseq. *Biology*, *1*(3), 575–596. doi:10.3390/biology1030575

Rosenberg, N. A. (2004). distruct: a program for the graphical display of population structure. *Molecular Ecology Notes*, *4*(1), 137–138. doi:10.1046/j.1471-8286.2003.00566.x

Saini, S., Jin, H., Hood, R., Barker, D., Mehrotra, P., & Biswas, R. (2011). The impact of hyper-threading on processor resource utilization in production applications. In *2011 18th International Conference on High Performance Computing* (pp. 1–10). doi:10.1109/HiPC.2011.6152743

Tange, O. (2011). GNU Parallel—The Command-Line Power Tool. *Login: The USENIX Magazine*, *36*(1), 42–47.

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. doi:10.1038/nature15393

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

Chapter 5

New SNPs mined from Cork Oak (*Quercus suber* L.) EST data provide preliminary insights on the species' potential response to climatic change

1 Abstract

The cork oak (*Quercus suber* L.) is a forest tree species with a West Mediterranean distribution, spanning from Portugal and Morocco in the West, to Tunisia and Italy in the East, including a location in Bulgaria of human introduced individuals. In the present study, SNPs were mined from a 454 RNA-Seq dataset of pooled anonymous cork oak individuals originating from contrasting climatic regions. Based on this information, 375 individuals were then genotyped via MassArray™ technique, representing most of *Q. suber*'s distribution range. This data was then used to find associations between SNP alleles and a set of environmental variables, as well as to attempt to detect natural selection signatures, which should help predict the species' response to climatic change. An assessment of the species' genetic variability was also performed in order to accommodate for the effects of population structure in the capacity to detect selection. Environmental association analyses revealed 10 associations of environmental variables with five of the genotyped loci, three of which may provide an important base for downstream studies. Genetic variability indices reveal a similar situation to that of other European oaks where nuclear markers were analyzed: considerable diversity, but no clear structuring pattern. These results contrast with those of previous studies based on chloroplastidial DNA which suggested a genetically segregated distribution, but should be regarded with caution, due to the low number of markers used. Finally, evidence was found that genes such as *fructose-1,6-biphosphatase*, *glutaredoxin*, or *carboxylesterase* are likely to be involved in the local adaptation process, since their allele frequencies were found associated with variables likely to be related to the their putative function.

2 Introduction

Cork oak (*Quercus suber* L.) is an evergreen tree species whose distribution spans throughout most of the Western Mediterranean coastal regions, the Atlantic coast of the Iberian peninsula and slightly north of the Pyrenees. *Q. suber* also occurs in Bulgaria, a population of Iberian origin (Borelli & Varela, 2000), introduced by human action (Petrov & Genov, 2004) (Figure 5.1). The strictly defined spectrum of environmental conditions endured by *Q. suber* (Aronson, Pereira, & Pausas, 2012) make it an interesting subject for environmental associations studies in the Mediterranean area.

Studies on cork oak genetic variation encompassing most of the species' distribution range are focused on the species' evolutionary history (Lumaret et al., 2005; Magri et al., 2007; Simeone et al., 2009; Toumi & Lumaret, 1998) and for the most part, based on plastidial markers, which may impose a bias on conclusions drawn from them alone (Govindarajulu, Parks, Tennessen, Liston, & Ashman, 2015). Studies attempting to find genetic associations with environmental variables have been typically performed on "local" scopes (Ramírez-Valiente et al., 2010), resorting to microsatellite loci markers.

Despite its relatively small scope, the previously mentioned study suggested local adaptive pressures as the explanation for the associations found between some of the six tested loci with temperature, leaf growth and leaf size. More recently, a study analyzing six candidate genes on a broad cork oak sampling (Modesto et al., 2014) has also revealed allelic variation associations with precipitation and temperature related variables.

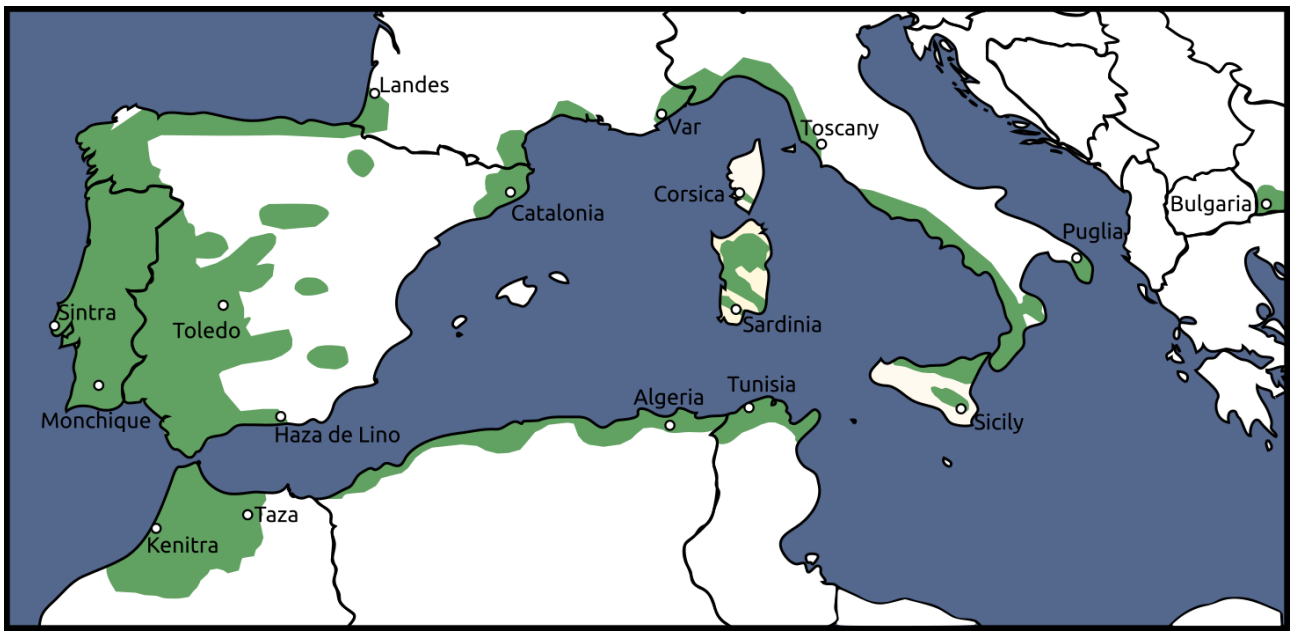

Figure 5.1: A map of *Q. suber* distribution area (green land areas) with the original sampling sites pinpointed. Adapted from EUFORGEN 2009 (www.euforgen.org).

Nowadays, an alternative marker type to those used in the above mentioned studies are Single Nucleotide Polymorphisms (SNPs) and one of the currently used strategies for their discovery is to mine them from Expressed Sequence Tags (ESTs) databases (Orsini, Jansen, Souche, Geldof, & De Meester, 2011). This is a direct consequence of the rapid decrease in production costs of these libraries which can currently be sequenced by Next Generation Sequencing (NGS) technologies (Kumar, Banks, & Cloutier, 2012), thus providing base information to develop new SNP markers for organisms with few available genomic resources.

Although SNP markers are usually less powerful than Simple Sequence Repeats (SSRs) for inferring population structuring (Glover et al., 2010), in some cases, such as those of species with high effective population size, the difference between both types of markers is minimal (Haasl & Payseur, 2011). On the other hand, SNP markers can be superior to SSRs regarding accuracy, reproducibility and robustness (Telfer et al., 2015).

Due to their nature, analyzing SNP markers brings challenges that did not affect previous markers, which were typically assumed to behave as strictly neutral. Non-neutral markers are known to provide better insights into genetic diversity, local

adaptation or evolutionary potential than their neutral counterparts (Kirk & Freeland, 2011), however, distinguishing both types of markers is not a simple task (Porcher, Giraud, & Lavigne, 2006). Methodologies for dealing with this issue are available (Foll & Gaggiotti, 2008), but when migration is high or the divergent selection pressure is weak (Thibert-Plante & Hendry, 2010), disentangling historical population structure from the adaptation process can be a daunting task.

In a global climatic change scenario (García-Ruiz, López-Moreno, Vicente-Serrano, Lasanta–Martínez, & Beguería, 2011), *Q. suber* becomes a particularly interesting study subject due to its relatively extended range. In order to predict its response to these changes (Kremer et al., 2012), it is important to understand the species' genetic architecture of adaptive traits (Alberto et al., 2013) and evolutionary history (Kremer, Potts, & Delzon, 2014). This issue has been previously studied in other genera such as *Pinus* (Alberto et al., 2013) or *Populus* (Olson et al., 2013), where the importance of these adaptive traits was highlighted to help predict the potential response of these tree species to global climatic alterations. On the other hand, *Q. suber* has also been studied regarding its potential response to climatic change, but from an ecological point of view (Correia, Bugalho, Franco, & Palmeirim, 2017; Vessella, López-Tirado, Simeone, Schirone, & Hidalgo, 2017) which did not consider the species' genetic adaptation potential.

Another factor that may play an important role regarding cork oak genetic background and evolutionary history is hybridization with other oaks, such as *Q. ilex*, *Q. coccifera* and *Q. cerris* (Belahbib et al., 2001; Burgarella et al., 2009; Costa et al., 2011; Petit et al., 2002) which is known to happen differentially across the species' range. This phenomenon may cause these species to maintain a higher genetic diversity than what would be expected if they did not exchange genetic material (Belahbib et al., 2001).

The first objective of this study is to perform environmental association and selection detection analyses in order to provide new insights into cork oak's adaptation process which may help predict its response to global climatic changes. This includes the identification and characterization of the genes where associated SNPs are found and relating their putative function with the environmental variable they are associated with. These analyses are based on SNP markers mined from ESTs and on climatic variables represented for individuals from across most of the species' range. This approach, however, required an initial assessment of the intraspecific genetic variability and population clustering to accommodate for the eventual role of genetic structure in the capacity to detect selection.

3 Methods

3.1 Field collection and environmental data

Most leaves were collected from cork oak trees planted on an international provenance trial (FAIR I CT 95 0202) established at “Monte Fava”, Alentejo, Portugal (38°00’ N; 8°7’ W) (Varela, 2000). The remaining individuals, from Portuguese (Sintra and Monchique) and Bulgarian locations, were collected directly from their original locations (Table 5.1, Figure 5.1). Twenty trees were sampled from each of the 17 sampling sites, spanning the full distribution range of the species. Sampling sites were selected considering both geographical distribution and environmental heterogeneity between the original locations, prioritizing populations that represent contrasting environments. Twenty *Quercus ilex* subsp. *rotundifolia* Lam. Individuals were also sampled from their native locations [Estremadura West region (Ranging from “Videla” 39º28'58''N; 8º37'54''W to “Pernes” 39º23'06''N; 8º39'02''W)] as well as twenty *Quercus coccifera* individuals [Estremadura West region (Ranging from “Foz do Arelho” 39º25'26''N; 9º12'26''W to “Alcabideche” 38º44'34''N; 9º27'11''W)]. This set of individuals from these two species are henceforth designated as *Putatively Introgressive Species (PIS)*. Plant material was stored at –80°C until DNA extraction.

Three spatial variables were recorded for each of the original sampling sites: altitude, latitude and longitude (Varela, 2000). Climatic data was gathered from the WorldClim database (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005) at 30 arc-seconds resolution (about 1 km) using a python script, *layer_data_extractor.py* (https://github.com/StuntsPT/Misc_GIS_scripts) as of commit bd36320. Nineteen Bioclimatic (BIO) variables were directly collected, BIO1 to BIO19. Correlations between variables were assessed using Pearson's correlation coefficient as implemented in the R script *eliminate_correlated_variables.R* (https://github.com/JulianBaur/R-scripts) as of commit 43e6553. This resulted in the exclusion of six variables due to high correlation (r>0.95). Each sampling location was thus characterized by three spatial variables and 13 environmental variables (Appendix II Table 1).

3.2 SNP mining from 454 EST data

The 454 dataset used in this work is described in (Pereira-Leal et al., 2014), designated as “L-19” and was obtained from RNA sequencing of leaf tissues from 8 *Q. suber* populations naturally growing on climatic divergent regions, in the scope of the FCT project SOBREIRO/0036/2009 (“Polymorphism detection and validation”).

EST data was analyzed using *4Pipe4* (Pina-Martins, Vieira, Seabra, Batista, & Paulo, 2016) as of commit 373c30e, with a minimum coverage of 15 reads, minimum less frequent allele frequency of 20% and minimum average base quality of 70 (configuration file with

all parameters available as Appendix II Data 1). BLAST queries of the largest Open Reading Frames (ORFs) were performed against NCBI's "nr" database (05-12-2013). Additionally, the mutation type of each ORF was characterized.

Table 5.1: Sampled sites, including original geographical location (coordinates in "Deg. Min. Sec" in the WGS84 system) , coupled with summary statistics per sampling location – expected heterozigozity (He), observed heterozigozity (Ho), Introgression coefficient (F_{IS}), Average Minimum Allele Frequency (MAF) and Hardy & Weinberg Equilibrium (HWE).

Sampling site	Num. Individuals	Latitude	Longitude	He	Ho	FIS	Average MAF	HWE Dhe q-value	HWE Ehe q-value
Algeria	20	36° 32' 24'' N	7° 9' 0'' E	0.372	0.337	0.0660	0.404	0.66	0.98
Bulgaria	20	41° 25' 59.88'' N	23° 10' 0.12'' E	0.361	0.363	-0.0324	0.391	0.89	0.71
Catalonia	20	41° 51' 0'' N	2° 31' 59.88'' E	0.392	0.427	-0.1184	0.418	0.99	0.59
Corsica	19	41° 37' 0.12'' N	8° 58' 0.12'' E	0.388	0.351	0.0400	0.371	0.68	0.98
Haza de Lino	20	36° 49' 59.88'' N	3° 17' 60'' W	0.379	0.427	-0.1696	0.412	0.99	0.21
Kenitra	18	34° 4' 59.88'' N	6° 34' 59.88'' W	0.312	0.322	-0.0633	0.347	0.89	0.71
Landes	20	43° 45' 0'' N	1° 19' 59.88'' W	0.290	0.337	-0.1832	0.371	0.99	0.21
Monchique	20	37° 19' 0.12'' N	8° 34' 0.12'' W	0.367	0.363	-0.0283	0.398	0.89	0.79
Puglia	19	40° 34' 0.12'' N	17° 40' 0.12'' E	0.353	0.340	-0.0073	0.347	0.81	0.98
Sardinia	20	39° 4' 59.88'' N	8° 50' 60'' E	0.347	0.353	-0.0582	0.365	0.89	0.79
Sicily	20	37° 7' 0.12'' N	14° 30' 0'' E	0.326	0.353	-0.1155	0.327	0.99	0.49
Sintra	19	38° 45' 0'' N	9° 25' 0.12'' W	0.394	0.404	-0.0493	0.416	0.89	0.96
Taza	20	34° 12' 0'' N	4° 15' 0'' W	0.318	0.323	-0.0538	0.386	0.89	0.71
Toledo	20	39° 22' 0.12'' N	5° 20' 60'' W	0.384	0.383	-0.0290	0.398	0.89	0.79
Tunisia	20	36° 57' 0'' N	8° 50' 60'' E	0.339	0.357	-0.0775	0.390	0.89	0.71
Tuscany	20	42° 25' 0.12'' N	11° 56' 60'' E	0.346	0.360	-0.0707	0.355	0.89	0.71
Var	20	43° 7' 59.9988'' N	6° 15' 0'' E	0.350	0.360	-0.0783	0.342	0.89	0.71
Q. ilex	20	39º 28' 58'' N	8º 37' 54'' W	0.079	0.080	-0.0638	0.361	0.086	0.996
Q. coccifera	20	39º 25' 26'' N	9º 12' 26'' W	0.045	0.040	0.0948	0.365	0.808	0.982
Mean	19.74	N/A	N/A	0.323	0.331	-0.0525	0.377	N/A	N/A
Mean *Q. suber*	19.71	N/A	N/A	0.354	0.362	-0.0605	0.379	N/A	N/A

The *4Pipe4* analysis pipeline was developed specifically to find SNPs in 454 datasets of pooled individuals when no reference information or strain information is available. The objective was to identify and mine SNPs unlikely to be false positives – located in high coverage zones and with high confidence in the base callings, in order to maximize the number of reliable markers for downstream genotyping assays.

Of the mined putative SNPs, those contained within contigs with ORFs whose best BLAST hit was against a mitochondrial or chloroplastidial sequence were discarded. Using the remaining SNPs putatively located in nuclear protein coding sequences, six potential assays were designed using Sequenom's "Assay Design Suite" based on

sequences from the flanking regions of each of the SNPs, which were produced using a python script (https://github.com/StuntsPT/4Pipe4_to_genotyping_array) as of commit 3644efd. From the generated assays, one of the two with most multiplexed SNPs (36 plex) was randomly selected.

In order to further explore the genotyped SNPs, the respective contig sequence was queried using BLAST against NCBI's nucleotide database “nt” as of 05-12-2013.

3.3 Sample preparation and genotyping

Total genomic DNA was extracted from liquid nitrogen-grounded leaves of all samples collected from *Q. suber*, *Q. ilex* and *Q. coccifera* individuals using the kit "innuPREP Plant DNA Kit" (Analytik Jena AG), according to the manufacturer's protocol.

The total amount of extracted DNA was quantified by spectrophotometry using a Nanodrop 1000 (Thermo Scientific) and integrity verified on an agarose gel (0.8%). DNA samples were then diluted to the same concentration and plated for genotyping.

DNA samples were then outsourced to “Sequenom GmbH, Hamburg” for genotyping using the “iPLEX Genotyping” technology on a “MassArray Analyser 4” with water randomly placed in 1-3 wells per plate to use as a negative controls. This technique consists in amplifying DNA samples in multiplex PCR reactions which are used for locus-specific single-base extension reactions. The resulting products are then discriminated by mass spectrometry.

3.4 Genetic diversity and differentiation

Deviations from Hardy & Weinberg Equilibrium (HWE) and expected and observed heterozygosity were calculated with *Genepop* 4.5.1 (Rousset, 2008), and filtered with a custom script (https://github.com/CoBiG2/RAD_Tools/blob/master/BioGenepop.py) that relies on the Biopython project (Cock et al., 2009) module “PopGen”. All *p*-values were corrected with a FDR test to account for multiple testing. The same software was also used to calculate F_{ST} values and to perform 10^6 permutations Mantel tests to correlate genetic with geographic distance (Rousset, 1997). F_{IS} values were obtained with the R package “diveRsity” v1.9.89 (https://cran.r-project.org/web/packages/diveRsity/) module “divBasic”. Plots of F_{ST} values were produced using the R script *Fst_heatmap.R* (https://github.com/StuntsPT/misc_plotters) as of commit 9911835.

Analyses involving geographic and environmental variables were performed excluding PIS and Bulgaria individuals.

All file format conversions for different software were performed with PGDSpider (Lischer & Excoffier, 2012). In order to increase the reproducibility value of this study, a “vcf” formatted file containing the raw SNP data is made available as Appendix II Data 2.

3.5 Outlier detection and environmental associations

In order to distinguish between neutral and non-neutral loci, the polymorphic genotyped SNPs were scanned for outliers using *Selestim* v1.1.4 (Vitalis, Gautier, Dawson, & Beaumont, 2014) with the following parameters: 200 pilot runs with a length of 5000, and a final run of $4e^{-6}$ MCMC iterations, a thinning interval of 50, and a burn-in length of 500000. The limit value for comparison with the generated pseudo observed distribution was 0.01.

The software Bayescan 2.1 (Foll & Gaggiotti, 2008) was also used for the same purpose, with the following parameters: 20 pilot runs with a length of 5000 iteration each, and a final run with a total of $4e^{-6}$ MCMC iterations, a thinning interval of 10 and a burnin length of 50000. The FDR test limit was set to 0.05.

Associations of SNP alleles to environmental and geospatial variables were performed with *Samβada* 0.4.2 (Stucki et al., 2014), with the model "BEST" and a significance value of 0.05. Due to concerns that eventual population structure could be causing an analysis bias, *Samβada* was run on three separate datasets – one containing all *Q. suber* individuals, one comprised only of individuals from Western sampling sites (*Sintra*, *Monchique*, *Kenitra*, *Toledo*, *Taza*, *Haza de Lino*, *Landes* and *Catalonia*) and one comprised of individuals from Eastern sampling sites (*Var*, *Algeria*, *Sardinia*, *Corsica*, *Tunisia*, *Tuscany*, *Sicilia* and *Puglia*), similar to the approach from (Modesto et al., 2014).

Two different approaches were used to determine wether any given locus is under the effects of selection. The first approach, called "Strict criteria" considered that any markers meeting at least three of the following criteria were considered "Non-neutral": (1) is detected as an outlier by *SelEstim*, (2) is detected as an outlier by *Bayescan*, (3) is associated to an environmental variable, and (4) is a non-synonymous mutation in the context of the largest ORF that can be drawn from the contig where the marker SNP is located. Combination of these criteria minimizes the flagging of false positive "non-neutral" loci, which is a known problem in the used methods (Vitalis et al., 2014). The second approach, called "Loose criteria" considered that only one of the above criteria is necessary for any given locus to be considered under selection (Table 5.2).

3.6 Genetic structure

Population genetic structure was inferred based on two datasets – one with all loci and one exclusively comprised of the loci previously defined as "Neutral" based on the "Strict criteria". Patterns of population structure were investigated using the software *fastSTRUCTURE* 1.0 (Raj, Stephens, & Pritchard, 2014) wrapped in *Structure_threader* v 1.1.0.post1 (Pina-Martins, Silva, Fino, & Paulo, 2016). *Structure_threader* was also used to infer the best value of "K" as well as to draw the individual cluster attribution plots, which were sorted by longitude. The Principal Components Analyses (PCA) were

performed with *vcf2PCA.R* (https://github.com/Nymeria8/R_little_scripts) as of commit "debec0e".

4 Results

4.1 SNP mining and genotyping

The 454 sequencing of L-19 dataset (Pereira-Leal et al., 2014) resulted in 1177063 reads. Of these, 201328 were discarded due to low quality, shortness or contaminant presence. From the remaining 975735 reads, 788232 assembled into 74950 contigs, with a maximum length of 3598 bp, N50, N90 and N95 of 602, 398 and 347, respectively. Of all the variation in the assembly, only 361 sites of 267 contigs matched the criteria to be considered putative SNPs by *4Pipe4* (on average, 1.35 putative SNPs per contig). Among these, six SNPs had a best BLAST hit against a mitochondrial or chloroplastidial protein, and were thus, discarded.

The *4Pipe4* report containing these results is available as Appendix II Data 3.

The remaining 355 putative SNPs, whose contigs' ORFs had a best BLAST hit against nuclear proteins, were used to design the MassArray™ genotyping assays.

Of the 36 SNPs comprising the performed assay, three (8%) failed to amplify, possibly due to undetected polymorphisms in the primer regions. The ratio between monomorphic and polymorphic SNPs was of 17/16, which translates to a false positive rate of approximately 51.5%. Only polymorphic SNPs were considered for subsequent analyses (Contig name to marker name translations can be found in Appendix II Table 2).

After querying the SNPs' contig sequences against the nucleotide database "nr", the SNP QSM001 was revealed to be located in a mitochondrial region, contrary to what the initial blast results against a protein database indicated, and was therefore discarded, resulting in a total of 15 SNP markers used for downstream analyses.

4.2 Genetic diversity and differentiation

Only two of the 15 polymorphic SNP loci were not found to be in *HWE*, QSN006, which displayed heterozygote excess, and QSN014, which displayed heterozygote deficiency (Table 5.2).

Regarding the *Q. suber* sampling sets (Table 5.1), observed heterozygosity (Ho) ranged from 0.322 (*Kenitra*) to 0.427 (*Haza de Lino*), while expected heterozygosity (He) ranged from 0.290 (*Landes*) to 0.394 (*Sintra*). F_{IS} values ranged from -0.183 (*Landes*) to 0.067 (*Algeria*), with only two positive values out of 17 measurements, which hints at a lot of gene flow between samples from different locations. Performing the *HWE* tests per set of samples revealed no significant deviations from equilibrium (Table 5.1).

When considering all 15 loci, pairwise F_{ST} values range from -0.018 (*Haza de Lino* vs. *Monchique*) to 0.530 (*Landes* vs. *Q. coccifera*), or 0.206 (*Var* vs. *Kenitra*) if considering only *Q. suber* individuals, with average values of 0.126 (all samples) and 0.060 (*Q. suber* only).

When considering only loci from the "Neutral" set, pairwise F_{ST} values range from -0.017 (*Taza* vs. *Algeria*) to 0.527 (*Landes* vs. *Q. coccifera*), or 0.181 (*Landes* vs. *Var*) if considering only *Q. suber* individuals, with average values of 0.084 (all samples) and 0.029 (*Q. suber* only).

4.3 Outlier detection and environmental associations

Selestim analyses identified seven SNPs that significantly deviate from a neutral distribution (Appendix II Figure 1): QSN002, QSN004, QSN005, QSN008, QSN011, QSN012 and QSN014 (Table 5.2).

Bayescan analyses identified four SNPs as outlier loci (Appendix II Figure 2): QSN001, QSN026, QSN009 and QSN010 (Table 5.2). Allele frequency graphs available as supplementary material (Appendix II Figure 3).

Using "Strict criteria" to assess the putative neutrality of each marker indicated five non-neutral SNPs: QSN002, QSN004, QSN008, QSN011 and QSN012 (Table 5.2). Using the "Loose criteria" indicated that all SNPs were under the effects of natural selection.

Environmental association analysis carried out with *Samβada* provided different results depending on the analyzed dataset. When all individuals are considered, five SNPs, QSN002, QSN004, QSN007, QSN008 and QSN011 are significantly associated with some of the considered variables (Table 5.2 and Appendix II Table 4). Specifically, QSN002 and QSN011 are associated with longitude, QSN004 and QSN007 with precipitation related variables (both to annual precipitation and the latter also to precipitation of the wettest and driest month) and QSN008 is associated with temperature related variables, namely "Isothermality" and mean diurnal range.

When analyzing only individuals from the Western sampling sites, two SNPs were found to be associated with environmental variables. Namely, QSN007 was associated with "Annual precipitation", "Precipitation of wettest month", "Precipitation of driest month" and "Altitude". These are the same associations that were found when all individuals are used, except for Altitude, which is exclusive to the Western individuals. The marker QSN012 was found to be associated with "Latitude", which was not found on the full dataset.

When performing the analysis with individuals exclusively from the Eastern sampling sites, only a single association is found, between QSN004 and "Annual precipitation", which is also present in the dataset with all individuals.

Table 5.2: Summary data of polymorphic SNPs. “Largest ORF silent” refers to whether the SNP alleles represent a silent mutation in the largest contig ORF. “HWE” refers to Hardy & Weinberg equilibrium. “Dhe” and “Ehe” refer to “deficit” and “excess” heterozygotes respectively. Associations written with a **bold** typeface occur on the dataset with individuals from all sampling sites, an *emphasis* typeface represent associations on the dataset with Eastern individuals, and underlined associations were fund in the dataset with only the Western individuals. Refer to section 3.4.1 for the definition of “Strict criteria” and “Loose criteria”.

SNP name	Putatively represented gene	HWE Dhe q-value	HWE Ehe q-value	Alleles	Num. ORFs	Largest ORF silent	SelEstim Outlier	Bayescan Outlier	Associations (**All**, West, *East*)	Genotype	Strict criteria	Loose criteria
QSN001	T-complex protein 1 subunit epsilon-like	0.74293	0.88788	AG	3	Yes	No	**Yes**	0	N/A	Neutral	**Non-Neutral**
QSN002	sphingoid long-chain bases kinase 1-like	0.74293	0.88788	CT	4	**No**	**Yes**	No	**Longitude**	CC	**Non-neutral**	**Non-neutral**
QSN003	alanyl-tRNA synthetase-like	0.92827	0.77813	AC	3	**No**	No	No	0	N/A	Neutral	**Non-neutral**
QSN004	fructose-1,6-bisphosphatase	0.85020	0.88788	GT	3	**No**	**Yes**	No	***Annual precipitation***	TT	**Non-neutral**	**Non-neutral**
QSN005	Galactosyltransferase family protein	0.85020	0.88788	AG	2	Yes	**Yes**	No	0	N/A	Neutral	**Non-neutral**
QSN006	Early light-induced protein	1.00000	**0.00000**	GT	2	**No**	No	**Yes**	0	N/A	Neutral	**Non-neutral**
QSN007	Glutaredoxin	0.74293	0.88788	CT	2	Yes	No	No	**Annual precipitation** **Precipitation of wettest month** **Precipitation of driest month** Altitude	CC	Neutral	**Non-neutral**
QSN008	Carboxylesterase	0.85020	0.88788	CT	1	**No**	**Yes**	No	**Latitude** **Mean diurnal range** **Isothermality**	CC	**Non-neutral**	**Non-neutral**
QSN009	NADH-ubiquinone oxidoreductase	0.92827	0.77813	CT	4	Yes	No	**Yes**	0	N/A	Neutral	**Non-neutral**
QSN010	fatty acid desaturase	0.98496	0.60525	CT	2	Yes	No	**Yes**	0	N/A	Neutral	**Non-neutral**
QSN011	Chlorophyll a/b binding protein	0.92827	0.78060	CG	1	**No**	**Yes**	No	**Longitude**	CC GG	**Non-neutral**	**Non-neutral**
QSN012	nuclear transcription factor Y subunit A-7-like	0.74293	0.97523	CT	2	**No**	**Yes**	No	Latitude	CC	**Non-neutral**	**Non-neutral**
QSN013	alcohol dehydrogenase class-3-like	0.60225	0.98539	AG	2	**No**	No	No	0	N/A	Neutral	**Non-neutral**
QSN014	extracellular calcium sensing receptor	**0.03750**	0.99750	CT	2	Yes	**Yes**	No	0	N/A	Neutral	**Non-neutral**
QSN015	uncharacterised	0.74293	0.88788	CG	1	**No**	No	No	0	N/A	Neutral	**Non-neutral**
QSM001	cytochrome oxidase subunit I (COI)	N/A	N/A	CT	1	**No**	N/A	N/A	N/A	N/A	N/A	N/A

A graphical representation of the pairwise F_{ST} values for *Q. suber* sampling sites can be seen in Figure 5.2. All pairwise F_{ST} values can be found in Appendix II Table 3.

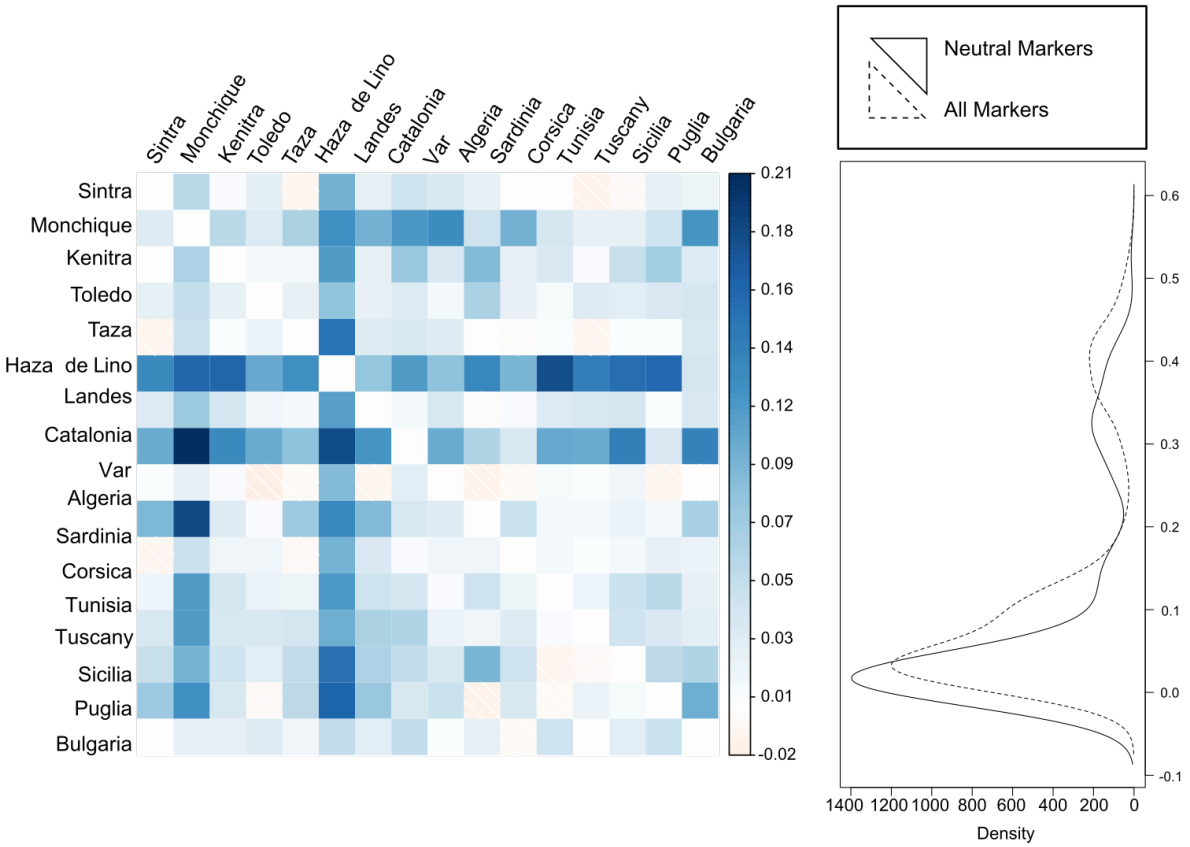

Figure 5.2: Pairwise F_{ST} plots between sampling sites of the neutral loci (upper matrix) and all loci (lower matrix) datasets of *Q. suber* only. The line plot on the right represents the F_{ST} value density of neutral loci (full line) and all loci (dashed line).1

Mantel tests revealed the probability of Isolation by distance (IBD) to be ~0 on all datasets.

4.4 Population genetic structure

Population genetic structure was separately inferred based on all 15 loci and on ten neutral loci. In both cases, the "bestk.py" test indicated 1 as the best fit value of K. This is strong evidence that no structuring can be found in the data. Considering this, K=2 was plotted (Figure 5.3), and although no clear structuring pattern is presented, individuals of *Q. ilex* and *Q. coccifera* are clustered together, in a somewhat segregated way from *Q. suber* individuals, but can not be distinguished from each other.

Since STRUCTURE is known to perform poorly when F_{ST} values are below 0.05, (Latch, Dharmarajan, Glaubitz, & Rhodes, 2006), a PCA analysis was also performed in order to confirm the results (Appendix II Figure 4). Similarly to what happened with the previous

analysis, the PCA is also able to distinguish *PIS* from *Q. suber*, but no evident segregation pattern can be found regarding cork oak individuals.

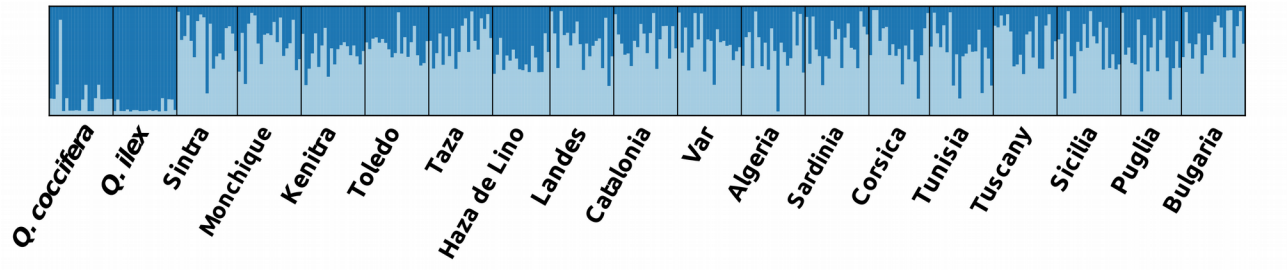

Figure 5.3: Plot of the clustering obtained with the software FASTSTRUCTURE. *Q. coccifera* and *Q. ilex* represent the *PIS* individuals, and all other labels represent the original sampling site of *Q. suber* individuals.

5 Discussion

5.1 SNP mining and genotyping

This study demonstrates that assemblies of anonymous individuals based on 454 technology can provide useful variation data which can be further explored by genotyping. Although the number of validated SNPs obtained by this approach is relatively low (49.51%), this value is in line with similar studies which revealed a 46% rate of false positives (Oliver et al., 2011). However, in the mentioned study, reference sequences were used, which should in principle, reduce the number of false positive SNPs comparatively to when this information is missing. This value is also below what is expected according to *4Pipe4*'s benchmarks (Pina-Martins, Vieira, et al., 2016), however the difference might be explained by the complexity differences between cork oak, which is a diploid eukaryotic organism, and the bacterial data that was used to benchmark *4Pipe4*.

As such, the amount of SNP markers used in this study is relatively low. However, it has been shown that as lineage divergence time increases, the number of markers required to correctly infer population structuring decreases (Haasl & Payseur, 2011), and fossil records indicate this to be the case of *Q. suber* (Carrión, Parra, Navarro, & Munuera, 2000), which suggests that any marked lineage divergence should be detectable, even using few markers, assuming any eventual structuring signal was not eroded due to high gene flow. That being stated, caution is generally advised when interpreting genetic structure results based on the 15 markers developed in this work. Despite this shortcoming, the genotyped SNP markers are adequate to obtain a clear signal in the detection of both outliers and environmental associations. The way they were mined also allows for further exploration than the typical markers obtained from Reduced Representation Libraries, since all of the markers have a large identified flanking region.

5.2 Genetic diversity and differentiation

Although values of genetic diversity will vary according to the type of markers from which they are inferred, measures found in this work can be generally compared with those from other publications focused on other European *Quercus* species, such as *Quercus petraea* and *Quercus robur*. In this regard, the data herein presented is in line with that from (Ballian, Belletti, Ferrazzini, Bogunić, & Kajba, 2010; Guichoux et al., 2013; Neophytou, Aravanopoulos, Fink, & Dounavi, 2010), which characterize these species (using nuclear markers) as low inbred, but essentially unstructured panmitic populations (Table 5.3).

Table 5.3: Comparison of genetic variability measures between the data from the present study and data from similar studies with other European oak species.

Author	This study	Ballian et al. 2010	Neophytou et al. 2010		Guichoux et al. 2013	
Scope	West Mediterranean	Bosnia and Herzegovina	Central – East Europe		Across France	
Species	*Q. suber*	*Q. robur*	*Q. robur*	*Q. petraea*	*Q. robur*	*Q. petraea*
Marker	SNPs	SSRs	SSRs	SSRs	SNPs	SNPs
F_{ST}	0.06	0.05	0.04	0.05	0.01	0.13
F_{IS}	-0.06	0.23	0.11	0.08	0	0
He	0.35	0.86	0.82	0.8	0.22	0.22
Ho	0.36	0.68	0.74	0.75	N/A	N/A

Overall, intraspecific pairwise F_{ST} values are relatively low, suggesting little differentiation among samples from different locations.

Moreover, regardless of the considered marker set, the probability of isolation by distance is always ~0, according to the Mantel test. This suggests that gene flow is indeed an important factor regarding the species' population genetics.

The rate of false positives is a known problem when attempting to determine weather or not a marker is under the effects of selection (Gautier, 2015; Stucki et al., 2014; Vitalis et al., 2014). All the SNP markers developed for this work can be considered as "non-neutral" under at least one of the used methods, designated as "Loose criteria". As such, more restrictive criteria have also been used to determine weather or not any given locus can be considered "non-neutral". This method, deemed "Strict criteria" selected five SNP markers as being under the effects of selection. On any account, this is a high ratio of SNPs under selection, however, it may be due to the fact that the markers were mined from ESTs, and from individuals from across the species' range, which makes them a lot more likely to be under selection than average markers taken from random non-coding genome regions.

5.3 Population genetic structure

The 15 markers developed for this work do not reveal much as far as population structure is concerned. Regardless of the used set of markers, the most likely value of K is one, despite the datasets containing multiple species. This can either be due to a true lack of population structure, or to the fact that the used markers do not have enough power to discern it. It is, nevertheless, possible to segregate between *PIS* and cork oak individuals. Although (Guichoux et al., 2013) states that this is to be expected, even based on few SNP markers, the fact that *PIS* cannot be distinguished from each other is somewhat unexpected. This might be due to the fact that the used markers were identified and developed exclusively from *Q. suber* genetic data, and as such, may not accurately represent the genetic diversity specific to the two *PIS*, or, once again, due to lack of power to make a distinction. It is, however, worth noting that *Q. ilex* and *Q. coccifera* have been found to frequently share haplotypes (Jiménez, de Heredia, Collada, Lorenzo, & Gil, 2004) and the sampled individuals of both *PIS* are from relatively close locations. It is in fact, quite possible that the lack of differentiation is due to a combination of these factors.

The *PIS* were included in the analyses in an attempt to gain an insight on the amount of introgression these species might incur along *Q. suber*'s distribution range. Although the three *Quercus* species included in this study are close enough for the developed markers to be common to all of them, and yet intraspecifically variable, the found variation was insufficient to provide information about any eventual differential introgression.

Regarding cork oak individuals, independently of dataset or value of K plotted, no evident structuring pattern could be found. This lack of structuring is supported by previous studies (Ramírez-Valiente, Valladares, & Aranda, 2014; Soto, Lorenzo, & Gil, 2007) regarding nuclear genetic diversity of *Q. suber*, despite the differences in geographic scope. Once again it should be highlighted that this can either be due to (1) a true lack of structure, (2) a lack of marker power to detect it, or (3) a combination of both these factors – weak structure and low marker power.

5.4 Environmental associations and outlier detection

Despite the markers used in this work providing no hint of structure, in order to perform environmental associations, three datasets were considered. This was similar to what was done in (Modesto et al., 2014), in an attempt to overcome an eventual analysis bias due to undetected population structure.

Of the 11 associations found between six markers and ecologically relevant variables across the species' natural distribution range, five of them are with geospatial variables. These variables may be functioning as a proxy for co-varying environmental data which was not considered *per se* on this study, such as photoperiod. Alternatively, these may

simply be a spurious correlation, not necessarily due to causation. As such, a functional overview of the genes containing the SNPs associated with environmental variables (Table 5.2) should help understand how likely causation is to be involved.

The gene containing QSN004 encodes a fructose-1,6-bisphosphatase, which is a regulatory enzyme critical for gluconeogenesis (Marcus & Harrsch, 1990). Water stress is reported to have an effect on the activity of this enzyme (Botha & Small, 1985), which might explain the genotypic association with "Annual precipitation". It is interesting to note that this association is present on the dataset with all individuals and on the dataset comprised of individuals exclusively from Eastern sampling sites. This marker is also considered "non-neutral" based on the "Strict criteria" for selection detection, which makes it particularly interesting for downstream functional studies.

The gene where QSN007 is located, codes for a glutaredoxin, a protein belonging to a family of enzymes involved in redox reactions (Fernandes & Holmgren, 2004). In other plants, such as *Arabidopsis thaliana* and *Solanum lycopersicum*, glutaredoxin plays a role in drought stress (Guo, Huang, Xie, Song, & Zhou, 2010), which could explain this SNP's association with precipitation related variables. This association can be found in both the dataset with all individuals and the one with only western individuals. Furthermore, in the smaller dataset, another association can be found, between QSN007 and the "Altitude" variable. In this case, "Altitude" can either be a proxy for another variable, or simply a spurious association, since the function of the protein where QSN007 is found can hardly be associated with the variable.

QSN008 is located in a gene coding for a carboxylesterase, which has catalytic activity (Krejci, Duval, A Chatonnet, P Vincens, & J Massoulié, 1991), and is known to play a role in ecological situations of heat stress (Lan, Li, Wang, & Ma, 2010). The putative function of this gene is in accordance with the associations found for the genotype with temperature related variables and even, to some extent, latitude. This association, however is exclusive to the full dataset. This can either mean that it only happens on a "global" scope, or that the reason this association is detected is due to a difference in allelic frequencies between the defined Eastern and Western groups.

The marker QSN012 is located in a gene that encodes a DNA binding protein, nuclear transcription factor Y subunit A-7 (Theologis et al., 2000). It is not clear why such marker can be associated with the variable "Latitude". This makes this association a likely candidate to being one of those cases whee causation is not involved.

Finally, the gene where QSN002 is located encodes a sphingoid long-chain bases kinase 1-like enzyme, which is involved in stomatal opening and closing (Nakagawa et al., 2012), and in disease resistance and apoptosis (Zhang et al., 2013). The marker QSN011 is located in a chlorophyll a-b binding protein encoding gene, which plays a crucial role in the Light Harvesting Complex (LHC), specifically in balancing the excitation energy

between the two photosystems (Liu & Shen, 2004) and also in thermal dissipation of excess absorbed light energy in the photosystem (Li et al., 2004). This protein is also known to play a role in drought stress (Xu et al., 2011). Associating any kind of function to a geospatial variable like "Longitude" does not make much biological sense, so either the variable is working as a proxy for another unrepresented environmental variable, or this correlation is non causative (Gautier, 2015; Stucki et al., 2014). Both these markers' associations are only present in the full dataset. This makes these markers less interesting for downstream analyses than QSN004, QSN007 or QSN008.

It is interesting to note that many of the genes containing SNPs associated with environmental variables are involved in heat and water stress, which coincides with what was reported in (Modesto et al., 2014). This could be due to the same environmental pressures affecting the allele frequencies of different genes involved in similar adaptive traits.

Due to their associations to environmental variables of identified ecological relevance, QSN004, QSN007 and QSN008 are interesting as candidate markers for downstream analyses. Further studying these markers and their respective sequences with case-control trials and on a functional level may be of paramount importance in understanding *Q. suber*'s response to the looming threat of climatic change.

5.5 Final remarks

This study shows what can be achieved based on SNP mining from EST data. Despite the method not having provided a great number of markers to explore, it allowed a greater ratio of markers to be further explored than what is typically obtained using Reduced Representation Libraries on organisms without reference genomes. This is due to the large flanking regions EST SNP mining typically provides, which allows a detailed exploration of putative gene functions. In this regard, the results provided by *4Pipe4* were a success.

The overall aim of this research was to make a contribution to better understanding the adaptation of cork oak and provide genomic tools that can help to forecast the consequences of climatic change for this species.

Genetic variability levels of the SNP markers developed for this work seem to be in line with those found for other European oaks. The possible comparisons, however, are of limited scope due to marker and scope differences. Doing a full re-analysis of the data from (Guichoux et al., 2013) to compare the results under the same models is a possibility to enable a better comparison, but falls out of the scope of this work. Nevertheless, the simple performed comparisons suggest that cork oak's standing genetic variability should not be too different from other European oaks, namely *Q. petraea* or *Q. robur*.

The novel markers developed for this study provide a different perspective on *Q. suber*'s evolutionary history than what was previously described with plastidial markers (Magri et al., 2007). Although the population structuring results are not conclusive, the few markers developed in this first step work show that more research needs to be performed with nuclear markers on cork oak in order to confirm whether there is more to *Q. suber*'s evolutionary history than what cpDNA can reveal.

This study also uncovered three markers (QSN004, QSN007, QSN008) that correlate to environmental variables ("Annual precipitation", "Precipitation of wettest month", Precipitation of driest month", "Mean diurnal range" and "Isothermality") which are likely to change in the near-medium term due to climatic changes. The corresponding genes could provide specific answers regarding cork oak's adaptation potential in downstream projects, especially if coupled with other markers from previous studies, such as those from (Modesto et al., 2014). The association of several climatic variables with identified genes related with temperature and water stress, suggests differential local adaptation pressures across the species range, but once again, further research is required to validate these findings.

Studies attempting to forecast the cork oak's response to climatic change from an ecological modeling perspective are non concordant (Correia et al., 2017; Vessella et al., 2017). Integrating this information with genetic/genomic based predictions should considerably improve the accuracy of such predictions. Results from this work highlight the presence of markers under natural selection, thus suggesting that with a larger number of markers it should be possible to make the aforementioned forecasts.

The combination of results presented here, regarding population genetic structure and candidate genes for further exploration, should provide a solid starting point to attempt to predict the species' response to climatic change, including the consequences for the "Montado" ecosystem and the economy based on the cork exploration. This study serves as a first step in this direction, and as a starting point for future research regarding the cork oak's adaptation potential.

6 Acknowledgements

We would like to thank R. Nunes, A. S. Rodrigues, C. Ribeiro and I. Modesto, for help with sample collection and to A. Kremer for the insightful revision of the first manuscript draft. Funding was provided by projects SOBREIRO/0036/2009 (under the framework of the Cork Oak ESTs Consortium), PTDC/AGR-GPL/104966/2008 and UID/BIA/00329/2013 from Fundação para a Ciência e Tecnologia (FCT) – Portugal. F. Pina-Martins was funded by FCT grant SFRH/BD/51411/2011, under the PhD program "Biology and Ecology of Global Changes", Univ. Aveiro & Univ. Lisbon, Portugal. D. Batista was funded b FCT grant SFRH/BPD/104629/2014.

7 References

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

Aronson, J., Pereira, J. S., & Pausas, J. G. (2012). *Cork Oak Woodlands on the Edge: Ecology, Adaptive Management, and Restoration*. Island Press.

Ballian, D., Belletti, P., Ferrazzini, D., Bogunić, F., & Kajba, D. (2010). Genetic variability of Pedunculate Oak (Quercus robur L.) in Bosnia and Herzegovina. *Periodicum Biologorum*, *112*(3), 353–362.

Belahbib, N., Pemonge, M. H., Ouassou, A., Sbay, H., Kremer, A., & Petit, R. J. (2001). Frequent cytoplasmic exchanges between oak species that are not closely related: Quercus suber and Q. ilex in Morocco. *Molecular Ecology*, *10*(8), 2003–2012.

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

Botha, F. C., & Small, J. G. (1985). Effect of Water Stress on the Carbohydrate Metabolism of Citrullus lanatus Seeds during Germination. *Plant Physiology*, *77*(1), 79–82.

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

Carrión, J. S., Parra, I., Navarro, C., & Munuera, M. (2000). Past distribution and ecology of the cork oak (Quercus suber) in the Iberian Peninsula: a pollen-analytical approach. *Diversity and Distributions*, *6*(1), 29–44. doi:10.1046/j.1472-4642.2000.00070.x

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, *25*(11), 1422–3.

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

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

Fernandes, A. P., & Holmgren, A. (2004). Glutaredoxins: Glutathione-Dependent Redox Enzymes with Functions Far Beyond a Simple Thioredoxin Backup System. *Antioxidants & Redox Signaling*, *6*(1), 63–74. doi:10.1089/152308604771978354

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

García-Ruiz, J. M., López-Moreno, J. I., Vicente-Serrano, S. M., Lasanta–Martínez, T., & Beguería, S. (2011). Mediterranean water resources in a global change scenario. *Earth-Science Reviews*, *105*(3–4), 121–139. doi:10.1016/j.earscirev.2011.01.006

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

Glover, K. A., Hansen, M. M., Lien, S., Als, T. D., Høyheim, B., & Skaala, Ø. (2010). A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genetics*, *11*(1), 2. doi:10.1186/1471-2156-11-2

Govindarajulu, R., Parks, M., Tennessen, J. A., Liston, A., & Ashman, T.-L. (2015). Comparison of nuclear, plastid, and mitochondrial phylogenies and the origin of wild octoploid strawberry species. *American Journal of Botany*, *102*(4), 544–554. doi:10.3732/ajb.1500026

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

Guo, Y., Huang, C., Xie, Y., Song, F., & Zhou, X. (2010). A tomato glutaredoxin gene SlGRX1 regulates plant responses to oxidative, drought and salt stresses. *Planta*, *232*(6), 1499–1509. doi:10.1007/s00425-010-1271-1

Haasl, R. J., & Payseur, B. A. (2011). Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, *106*(1), 158–171. doi:10.1038/hdy.2010.21

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*(15), 1965–1978. doi:10.1002/joc.1276

Jiménez, P., de Heredia, U. L., Collada, C., Lorenzo, Z., & Gil, L. (2004). High variability of chloroplast DNA in three Mediterranean evergreen oaks indicates complex evolutionary history. *Heredity*, *93*(5), 510–515. doi:10.1038/sj.hdy.6800551

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

Krejci, E., Duval, N., A Chatonnet, P Vincens, & J Massoulié. (1991). Cholinesterase-like domains in enzymes and structural proteins: functional and evolutionary relationships and identification of a catalytically essential aspartic acid. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(15), 6647–6651. doi:10.1073/pnas.88.15.6647

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

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

Kumar, S., Banks, T. W., & Cloutier, S. (2012). SNP Discovery through Next-Generation Sequencing and Its Applications. *International Journal of Plant Genomics*, *2012*. doi:10.1155/2012/831460

Lan, P., Li, W., Wang, H., & Ma, W. (2010). Characterization, sub-cellular localization and expression profiling of the isoprenylcysteine methylesterase gene family in Arabidopsis thaliana. *BMC Plant Biology*, *10*(1), 212. doi:10.1186/1471-2229-10-212

Latch, E. K., Dharmarajan, G., Glaubitz, J. C., & Rhodes, O. E. (2006). Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, *7*(2), 295–302. doi:10.1007/s10592-005-9098-1

Li, X.-P., Gilmore, A. M., Caffarri, S., Bassi, R., Golan, T., Kramer, D., & Niyogi, K. K. (2004). Regulation of Photosynthetic Light Harvesting Involves Intrathylakoid Lumen pH Sensing by the PsbS Protein. *Journal of Biological Chemistry*, *279*(22), 22866–22874. doi:10.1074/jbc.M402461200

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

Liu, X.-D., & Shen, Y.-G. (2004). NaCl-induced phosphorylation of light harvesting chlorophyll a/b proteins in thylakoid membranes from the halotolerant green alga, Dunaliella salina. *FEBS Letters*, *569*(1–3), 337–340. doi:10.1016/j.febslet.2004.05.065

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*(5), 853–861. doi:10.1093/aob/mci237

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

Marcus, F., & Harrsch, P. B. (1990). Amino acid sequence of spinach chloroplast fructose-1,6-bisphosphatase. *Archives of Biochemistry and Biophysics*, *279*(1), 151–157. doi:10.1016/0003-9861(90)90475-E

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

Nakagawa, N., Kato, M., Takahashi, Y., Shimazaki, K., Tamura, K., Tokuji, Y., … Imai, H. (2012). Degradation of long-chain base 1-phosphate (LCBP) in Arabidopsis: functional characterization of LCBP phosphatase involved in the dehydration stress response. *Journal of Plant Research*, *125*(3), 439–449. doi:10.1007/s10265-011-0451-9

Neophytou, C., Aravanopoulos, F. A., Fink, S., & Dounavi, A. (2010). Detecting interspecific and geographic differentiation patterns in two interfertile oak species (Quercus petraea (Matt.) Liebl. and Q. robur L.)

using small sets of microsatellite markers. *Forest Ecology and Management*, *259*(10), 2026–2035. doi:10.1016/j.foreco.2010.02.013

Oliver, R. E., Lazo, G. R., Lutz, J. D., Rubenfield, M. J., Tinker, N. A., Anderson, J. M., … Jackson, E. W. (2011). Model SNP development for complex genomes based on hexaploid oat using high-throughput 454 sequencing technology. *BMC Genomics*, *12*(1), 77. doi:10.1186/1471-2164-12-77

Olson, M. S., Levsen, N., Soolanayakanahally, R. Y., Guy, R. D., Schroeder, W. R., Keller, S. R., & Tiffin, P. (2013). The adaptive potential of Populus balsamifera L. to phenology requirements in a warmer global climate. *Molecular Ecology*, *22*(5), 1214–1230. doi:10.1111/mec.12067

Orsini, L., Jansen, M., Souche, E. L., Geldof, S., & De Meester, L. (2011). Single nucleotide polymorphism discovery from expressed sequence tags in the waterflea Daphnia magna. *BMC Genomics*, *12*, 309.

Pereira-Leal, J. B., Abreu, I. A., Alabaça, C. S., Almeida, M. H., Almeida, P., Almeida, T., … Ricardo, C. P. (2014). A comprehensive assessment of the transcriptome of cork oak (Quercus suber) through EST sequencing. *BMC Genomics*, *15*(1), 371. doi:10.1186/1471-2164-15-371

Petit, R. J., Csaikl, U. M., Bordács, S., Burg, K., Coart, E., Cottrell, J., … Kremer, A. (2002). Chloroplast DNA variation in European white oaks: Phylogeography and patterns of diversity based on data from over 2600 populations. *Forest Ecology and Management*, *156*(1–3), 5–26. doi:10.1016/S0378-1127(01)00645-4

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

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

Pina-Martins, F., Vieira, B. M., Seabra, S. G., Batista, D., & Paulo, O. S. (2016). 4Pipe4 – A 454 data analysis pipeline for SNP detection in datasets with no reference sequence or strain information. *BMC Bioinformatics*, *17*, 41. doi:10.1186/s12859-016-0892-1

Porcher, E., Giraud, T., & Lavigne, C. (2006). Genetic differentiation of neutral markers and quantitative traits in predominantly selfing metapopulations: confronting theory and experiments with Arabidopsis thaliana. *Genetical Research*, *87*(1), 1–12. doi:10.1017/S0016672306007920

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, *197*(2), 573–589. doi:10.1534/genetics.114.164350

Ramírez-Valiente, J. A., 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*(5), 1031–1044. doi:10.1007/s10682-010-9365-6

Ramírez-Valiente, J. A., Valladares, F., & Aranda, I. (2014). Exploring the impact of neutral evolution on intrapopulation genetic differentiation in functional traits in a long-lived plant. *Tree Genetics & Genomes*, 1–10. doi:10.1007/s11295-014-0752-y

Rousset, F. (1997). Genetic Differentiation and Estimation of Gene Flow from F-Statistics Under Isolation by Distance. *Genetics*, *145*(4), 1219–1228.

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

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

Soto, A., Lorenzo, Z., & Gil, L. (2007). Differences in fine-scale genetic structure and dispersal in Quercus ilex L. and Q. suber L.: consequences for regeneration of mediterranean open woods. *Heredity*, *99*(6), 601–607. doi:10.1038/sj.hdy.6801007

Stucki, S., Orozco-terWengel, P., Bruford, M. W., Colli, L., Masembe, C., Negrini, R., … Consortium, the N. (2014). High performance computation of landscape genomic models integrating local indices of spatial association. *ArXiv:1405.7658 [q-Bio]*. Retrieved from http://arxiv.org/abs/1405.7658

Telfer, E. J., Stovold, G. T., Li, Y., Silva-Junior, O. B., Grattapaglia, D. G., & Dungey, H. S. (2015). Parentage Reconstruction in Eucalyptus nitens Using SNPs and Microsatellite Markers: A Comparative Analysis of Marker Data Power and Robustness. *PLOS ONE*, *10*(7), e0130601. doi:10.1371/journal.pone.0130601

Theologis, A., Ecker, J. R., Palm, C. J., Federspiel, N. A., Kaul, S., White, O., … Davis, R. W. (2000). Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana. *Nature*, *408*(6814), 816–820. doi:10.1038/35048500

Thibert-Plante, X., & Hendry, A. P. (2010). When can ecological speciation be detected with neutral loci? *Molecular Ecology*, *19*(11), 2301–2314. doi:10.1111/j.1365-294X.2010.04641.x

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*(4), 647–656. doi:10.1007/s001220050941

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

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

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

Xu, Y.-H., Liu, R., Yan, L., Liu, Z.-Q., Jiang, S.-C., Shen, Y.-Y., … Zhang, D.-P. (2011). Light-harvesting chlorophyll a/b-binding proteins are required for stomatal response to abscisic acid in Arabidopsis. *Journal of Experimental Botany*, err315. doi:10.1093/jxb/err315

Zhang, H., Li, L., Yu, Y., Mo, J., Sun, L., Liu, B., … Song, F. (2013). Cloning and characterization of two rice long-chain base kinase genes and their function in disease resistance and cell death. *Molecular Biology Reports*, *40*(1), 117–127. doi:10.1007/s11033-012-2040-y

Chapter 6

New insights on adaptation and population structure of Cork Oak using genotyping by sequencing

1 Abstract

Species respond to global climatic changes in a local context. Understanding this process is paramount due to the pace at which these changes are occurring. Tree species are particularly interesting to study in this regard due to their long generation times, sedentarism, and because of their ecological and economic importance. *Quercus suber* L. is an evergreen forest tree species of the Fagaceae family with an essentially Western Mediterranean distribution. Even though this species' evolutionary history has been frequently studied, large-scale genetic studies have essentially relied on plastidial markers, whereas those performed with nuclear markers have been done on locally focused sampling strategies. The potential response of *Q. suber* to global climatic changes has also been studied, under ecological modelling. In this work, the "Genotyping by Sequencing" (GBS) technique is used to derive 2,547 SNP markers in order to assess the species' evolutionary history from a nuclear DNA perspective, to gain insights on how local adaptation may be shaping the species' genetic background, and to attempt to forecast how the cork oak may respond to global climatic changes from a genetic perspective. Results reveal an essentially unstructured species, where a balance between gene flow and local adaptation keeps the species' gene pool somewhat homogeneous across its distribution, but at the same time allows variation clines for the individuals to cope with local conditions. "Risk of Non-Adaptedness" (RONA) analyses, based on environmental association results, suggest that for the considered variables and most sampled locations, the cork oak does not require large shifts in allele frequencies to survive the predicted climatic changes. However, more research is required to integrate these results with the ecological modelling ones.

2 Introduction

2.1 Adaptation

Global climatic changes have been shown to cause alterations in species' traits (Benito Garzón, Alía, Robson, & Zavala, 2011; Walther et al., 2002). Understanding how species respond to such alterations in their environmental context is becoming an increasingly important question due to the pace at which they are taking place (Kremer et al., 2012; Primack et al., 2009). Too avoid obliteration, species may respond to climatic changes by either altering their distribution range, effectively going extinct in the original location but persisting somewhere else, or by adapting to the new conditions. The latter can occur "instantly", due to phenotypic plasticity, or across several generations, by local adaptation (Aitken, Yeaman, Holliday, Wang, & Curtis-McLane, 2008). The kind of response species can provide is known to depend on factors like location, distribution range, and/or genetic background (Gienapp, Teplitsky, Alho, Mills, & Merilä, 2008;

Ohlemuller, Gritti, Sykes, & Thomas, 2006). Climatic change is modelled for the Mediterranean region (Giorgi & Lionello, 2008), and some of the species in this region have been studied regarding the impact of these changes (Lindner et al., 2010), but much is still unknown.

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

In this work, we address the case of the cork oak (*Quercus suber* L.). With a distribution ranging most of the West Mediterranean region (Figure 6.1), this oak species is the most selective evergreen oak of the Mediterranean basin in terms of precipitation and temperature conditions (Vessella, López-Tirado, Simeone, Schirone, & Hidalgo, 2017). European oaks in particular, are known to have endured past climatic alterations, but how they can cope with the current, rapidly occurring changes is not yet fully understood (Kremer et al., 2012; Kremer, Potts, & Delzon, 2014). Despite this tree's ecological and economic importance, little is known regarding the consequences of global climatic change on its future (Benito Garzón, Sánchez de Dios, & Sainz Ollero, 2008). Some recent works have been performed to attempt to answer this very question, but focusing on range expansion and contraction with the assumption of a genetically homogeneous species (Correia, Bugalho, Franco, & Palmeirim, 2017; Vessella et al., 2017). Both these studies also highlight the need for a genetic study regarding the adaptation potential of *Q. suber*.

In this regard, studies integrating genetic information and response to climatic alterations of *Q. suber* are rare and of small scale (Modesto et al., 2014) when compared with other oak species (Rellstab et al., 2016). Studies such as Jose Alberto Ramírez-Valiente, Valladares, Huertas, Granados, & Aranda (2011) have revealed that some traits can be associated to genetic variants, however, these were performed on a local scope and using a relatively low number of markers, which limits their utility in a larger scope. Knowing gene flow and local adaptation dynamics of *Q. suber* is paramount to understanding the species' potential to endure rapid climatic changes through adaptation (Savolainen, Lascoux, & Merilä, 2013).

Genomic resources represent a new way to study the genetic mechanisms responsible for local adaptation (Rellstab, Gugerli, Eckert, Hancock, & Holderegger, 2015), through the use of environmental association analyses, which correlate environmental data with genetic markers, thus highlighting loci putatively involved in the adaptation process (Rellstab et al., 2016). The same methods, can thus, in principle, be used to assess the degree of maladaptation to predicted future local conditions (Rellstab et al., 2016). Applying this kind of methodology on *Q. suber* would fill the gap mentioned in (Correia

et al., 2017; Vessella et al., 2017). The Risk of Non-adaptedness (RONA) method was developed in (Rellstab et al., 2016) with this very goal, however, no public implementation is provided in the mentioned work.

2.2 Population structure

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

The evolutionary history of *Q. suber* has been studied in the past using multiple methodologies and in different geographic ranges. The most recent large-scale studies on the subject suggest that cork oak is divided into four strictly defined lineages (Magri et al., 2007; Simeone et al., 2009). Two of these lineages range from the south-east of France, to Morocco, including the Iberian peninsula and the Balearic Islands, a third lineage ranges from the Monaco region to Algeria and Tunisia, including the islands of Corsica and Sardinia. The fourth lineage spans the entire Italic peninsula, including Sicilia. Based only on plastidial markers, these lineages have been shown to hardly share any haplotypes. Notwithstanding, later works based on nuclear DNA have hinted at a different scenario, where the species is not as categorically divided(Costa et al., 2011; J. A. Ramírez-Valiente, Valladares, & Aranda, 2014) (see also Chapter 5). These works are, however, limited in either geographic scope or number of markers to confidently conclude that such segregation is only present in plastidial markers.

2.3 Objectives

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

3 Material & Methods

3.1 Sample and environmental data collection

In order to provide a comprehensive view of the species genetic background, samples were collected from 17 locations spanning most of *Q. suber*'s distribution. Fresh leaves were collected from six individuals from, *Bulgaria*, *Corsica*, *Kenitra*, *Monchique*, *Puglia*, *Sardinia*, *Sicilia*, *Tuscany*, *Tunisia* and *Var*, and from five individuals from *Algeria*, *Catalonia*, *Haza de Lino*, *Landes*, *Sintra*, *Taza* and *Toledo* for a total of 95 individuals (Table 6.1, Figure 6.1).

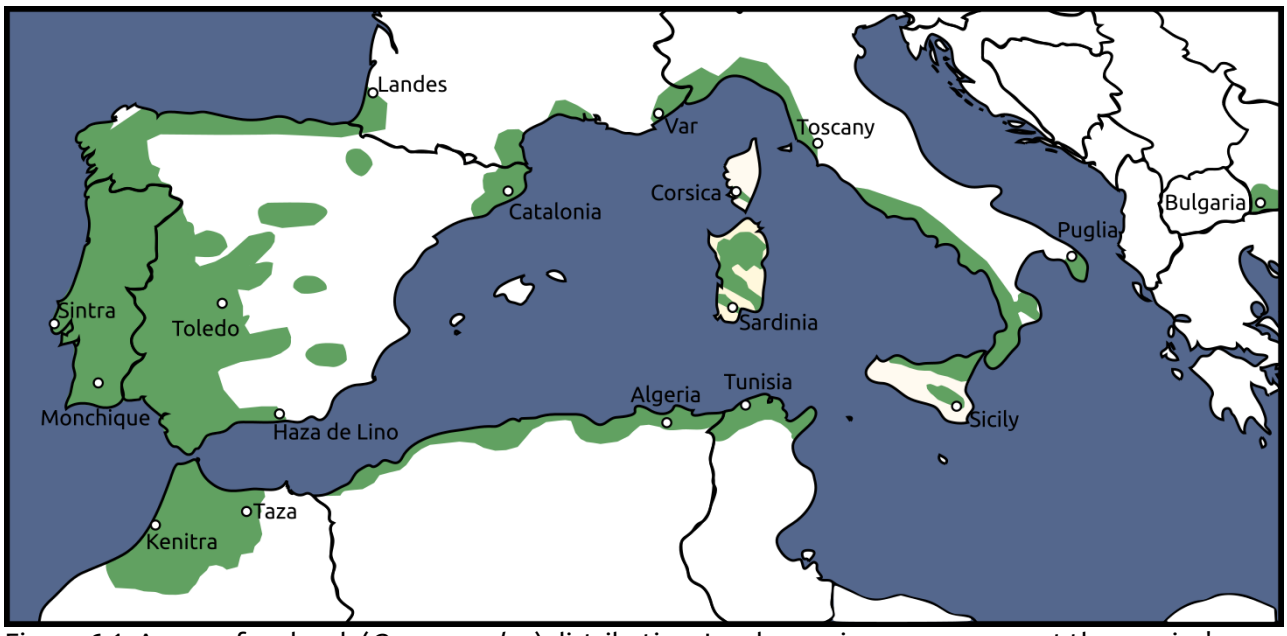

Figure 6.1: A map of cork oak (*Quercus suber*) distribution. Land areas in green represent the species' range. White dots represent the sampling locations. Adapted from EUFORGEN 2009 (www.euforgen.org).

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

Altitude, latitude and longitude spatial variables (Varela, 2000) were recorded for each of the native sampling sites. Nineteen Bioclimatic (BIO) variables, BIO1 to BIO19 were collected from the WorldClim database (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005) at 30 arc-seconds (~ 1 km) resolution for "Current conditions ~1960-1990" and "Future" predictions for 2070 (*rcp26* and *rcp85* conditions for the following "Global Climate Models" (GCMs): BCC-CSM1-1, CCSM4, GFDL-CM3, GISS-E2-R, HadGEM2-ES, IPSL-CM5A-LR, MRI-CGCM3, MPI-ESM-LR and NorESM1-M as these are available under permissive licenses and calculated for both *rcp26* and *rcp85*). An average of the mentioned datasets was obtained for each coordinate and variable used in the analyses (Appendix III Table 1

and 2 respectively). Data was extracted from the GeoTiff files using a python script, *layer_data_extractor.py* (https://github.com/StuntsPT/Misc_GIS_scripts) as of commit "bd36320".

Table 6.1: Coordinates and number of sampled individuals for every sampling site.

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

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

3.2 Library preparation and sequencing

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

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

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

3.3 Genomic data analyses

The raw GBS data received from "Genomic Diversity Facility" was analysed using the program *ipyrad* v0.5.15, which is based on *pyrad* (Eaton, 2014), using the provided "*conda*" environment - *MUSCLE* v3.8.31 (Edgar, 2004) and *VSEARCH* v2.0.3 (Rognes, Flouri, Nichols, Quince, & Mahé, 2016). Sequence assembly was performed for the GBS *datatype*, with a *clustering threshold* of 0.95, a *mindepth* of 8 and maximum *barcode* mismatch of 0. Each sampling site had to be represented by at least three individuals for a SNP to be called, except the locations of *Kenitra* and *Taza*, where only one individual was required, due to the lower representation of these sampling sites. Full parameters can be found in Appendix III Data 1. The demultiplexed "fastq" files were submitted to NCBI's Sequence Read Archive SRA) as "Bioproject" PRJNA413625.

Downstream analyses were automated using "*GNU Make*". This file, containing every detail of every step of the analyses for easier reproducibility is presented as Appendix III Data 2. For improved reproducibility, a docker image with all the software, configuration files, parameters and the *Makefile*, ready to use is also provided (https://hub.docker.com/r/stunts/q.suber_gbs_data_analyses/). The intent is not to allow the analyses process to be treated as a "black box", but rather to provide a full environment that can be reproduced, studied and modified by the scientific community.

Processed data from *ipyrad* was then filtered using *VCFtools* v0.1.14 (Danecek et al., 2011) with the following criteria: each sample has to be represented in at least 55% of the SNPs, and after this each SNP has to be represented in at least 80% of the individuals. Furthermore, due to the relatively small sample size, the minimum allele frequency (MAF) of each SNP has to be at least 0.05 for it to be retained.

In order to minimize the effects of linkage disequilibrium, analyses downstream from this point were performed using only the SNP closest to the centre of the sequence where each SNP was found. This sub dataset was obtained using the python script

vcf_parser.py (https://github.com/CoBiG2/RAD_Tools/blob/master/vcf_parser.py) as of commit "0893296".

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

Descriptive statistics, such as Hardy-Weinberg Equilibrium (HWE), F_{ST} and F_{IS} were calculated using *Genepop* v4.6 (Rousset, 2008). The same software was further used to perform Mantel tests to determine an eventual effect of Isolation by Distance (IBD) by correlating "'F/(1-F)'-like with common denominator" with "Ln(distance)" following on 1,000,000 permutations. This test was performed excluding individuals sampled from *Bulgaria* due to their introduced origin (Chapter 5, section 2).

3.4 Population Structure

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

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

The STRUCTURE method was performed with STRUCTURE v2.3.4, (Pritchard et al., 2000) using the admixture model with an inferred *alpha.* To achieve the best results using STRUCTURE, 20 replicates of each "K" were run at 200000 iterations (10% burnin), and the three best values of delta K were then run for a single replicate at 2000000 iterations (10% *burnin*). The same method was also performed as implemented under *MavericK*. In this case, two runs were performed: an initial single "pilot" run of 5000 iterations, with a *burnin* of 500 using an admixture model, a free *alpha* parameter of "1" and "thermodynamic integration" (TI) turned off. Tuned *alpha* and *alphaPropSD* values were extracted from the pilot run and used in the "tuned" run as parameters for the admixture model. This run was comprised of five runs of 10000 iterations (10% burnin), with TI turned on and set to 20 rungs of 10000 samples with 20% burnin. Both programs were wrapped under *Structure_threader* v 1.2.2 (Pina-Martins, Silva, Fino, & Paulo, 2016) for values of "K" between 1 and 8. The most suitable value of "K" was calculated using the *evanno* (Earl & vonHoldt, 2012) and TI methods for and STRUCTURE and *MavericK* respectively. Full parameter files are available as Appendix III Data 2.

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

3.5 Outlier detection and environmental associations

Outlier detection was performed using two programs: *SelEstim* v1.1.4 (Vitalis, Gautier, Dawson, & Beaumont, 2014) (50 pilot runs of length 1000 followed by a main run of length 10^6, with a burnin of 1000, a thinning interval of 20, and a detection threshold of 0.01) and *BayeScan* v2.1 (Foll & Gaggiotti, 2008) (20 pilot runs of length 5000 followed by a main run of 500000 iterations, a burnin of 50000, a thinning interval of 10, and a detection threshold of 0.05) (full commands and parameters available in Appendix III Data 2), since these methods show the lowest rate of false positives (Narum & Hess, 2011; Vitalis et al., 2014). Only SNPs indicated as outliers by both programs were considered outliers for the purpose of this work. This was done to reduce the chance of false positives, which is a known issue in this type of analyses (Gautier, 2015; Vitalis et al., 2014).

The software *BayPass* v2.1 (Gautier, 2015) wrapped under the script *Baypass_workflow.R* (https://github.com/StuntsPT/pyRona/blob/master/Baypass_workflow.R) as of commit "5b406fb" was used to assess associations of SNPs to environmental variables using the "AUX" model (20 pilot runs of length 1000, followed by a main run of length 500000 with a burnin of 5000 and a thinning interval of 25). Any association with a Bayes Factor (BF) above 15 was considered significant. Similar to what was done for the Mantel tests, association analyses were performed excluding individuals from *Bulgaria* sampling site.

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

3.6 Risk of non-adaptedness

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

In short, for every significant association between a SNP and an environmental variable, the RONA method plots each location's individuals' allele frequencies (corrected by

Baypass to eliminate any possible effects of population structure) vs. the respective environmental variable. This is done for both the current value and the future prediction. A correlation between allele frequencies and the current variable values is then calculated and the corresponding best fit line is inferred. The distance between the fitted line and the two coordinates is then compared per location and its normalized difference is considered the RONA value for each association and location (which can vary between 0 and 1). In theory, the higher the difference in conditions between the current values and the prediction, the more *Q. suber* should have to shift its allele frequencies to survive in the location under the new conditions.

Two alternative climate prediction models were used to calculate a RONA value for each location, a low emission scenario (RPC26) and a high emission scenario (RPC85) in order to account for uncertainties in the models' assumptions.

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

4 Results

Genotyping by sequencing (Elshire et al., 2011), a technique based on restriction enzyme genomic complexity reduction followed by short-read sequencing, was employed to discover SNP markers from a total of 95 *Q. suber* individuals sampled from 17 geographical locations (Table 6.1).

A total of 225,214,094 reads (100 bp) generated by the GBS assay was processed by *ipyrad* (Eaton, 2014) computational pipeline. The first step of the analysis process consisted in the assembly of raw reads into 7,456 loci, containing 12,330 SNPs. Twelve *Q. suber* samples were discarded due to low sequence representation, resulting in the retention of 83 individuals. After filtering according to the criteria presented in 3.3, 2,547 SNPs remained, which were used for all further analyses. The filtering process also further removed two samples due to too much missing data (>55%), and therefore, of the 83 remaining samples, only 81 were used in the analyses (Table 6.2).

The calculated F_{IS} values for each sampling site are available in Table 6.2. These range from -0.0234 (*Landes*) to 0.0987 (*Puglia*) with an average value of 0.0531. Pairwise F_{ST} values are available in Figure 6.2 and Appendix III Table 4. These range from 0.0038 between *Sintra* and *Monchique* to 0.1225 between *Kenitra* and *Var* (average F_{ST} of 0.0553).

Table 6.2: Number of individuals used in analysis, F_{IS} values, and Hardy-Weinberg Equilibrium (HWE) *p*-values for each sampling site.

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

Hardy-Weinberg Equilibrium tests revealed that a heterozygote deficit exists in most sampling sites (Table 6.2), in fact, only *Bulgaria* and *Landes* sampling sites seem not to have an excess of homozygote individuals. When looking at HWE results per marker, of the 2,547 SNPs, only 109 reveal a heterozygote deficit, whereas 23 reveal a deficit of homozygotes. Performing the same test on all individuals as a single large population also revealed a deficit of heterozygotes. The performed Mantel test revealed no evidence of IBD between the *Q. suber* samples.

4.1 Population structure

In order to perform clustering analyses, it is important to estimate the value of "K", which represents how many *demes* the data can be clustered into. The software *MavericK* is especially interesting for cluster estimation due to its innovative method for estimating "K", called "Thermodynamic Integration" (TI), which has shown superior performance in this task relative to other methods (Verity & Nichols, 2016). In this case, the "TI" method determined the best "K" value to be "1" on both the full dataset and the "neutral" dataset and "2" in the "non-neutral" dataset (Appendix III Figure 1). The classic method for the STRUCTURE software, the *evanno* method revealed that K=2 had the best

ΔK, followed by K=3 and K=4 on all datasets. It is, however, important to note that the *evanno* method is not able to evaluate the ΔK value for K=1.

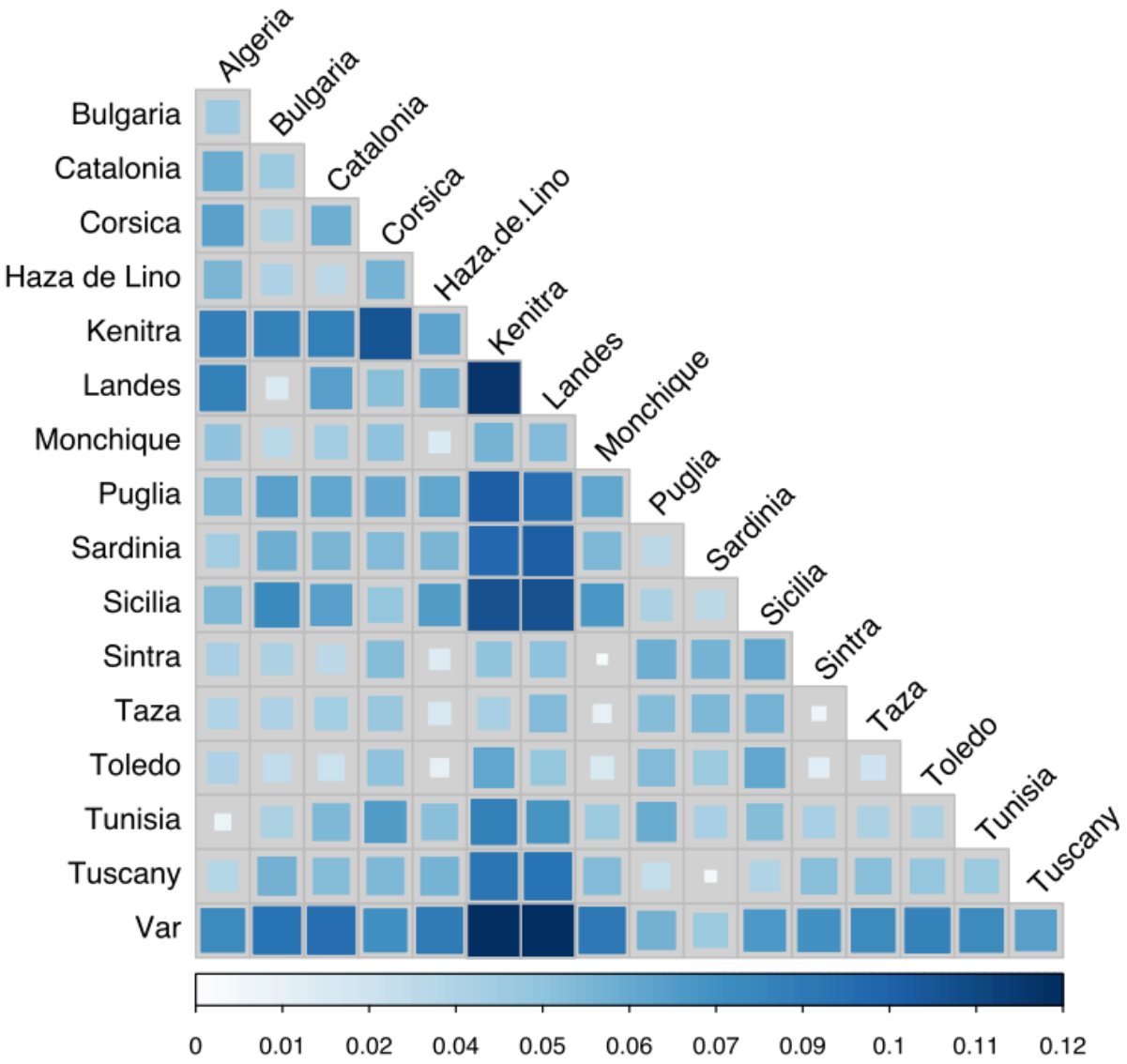

Figure 6.2: Pairwise F_{ST} plot between sampling sites. Darker blue represents a higher pairwise F_{ST} value, and lighter blue represents a lower value.

Q-matrix plots provide the relatedness of each genotype to each considered deme (as many demes are represented as the value of "K"). The Q-matrix plot of *MavericK*'s results produced using all loci (Figure 6.3A) can be interpreted as a rough split between Western individuals (from locations *Sintra*, *Monchique*, *Kenitra*, *Toledo*, *Landes*, *Taza*, *Haza de lino* and *Catalonia*), which are mostly assigned to cluster "1" and Eastern ones (from locations *Var*, *Algeria*, *Sardinia*, *Corsica*, *Tunisia*, *Tuscany*, *Sicilia*, *Puglia* and *Bulgaria*), which are mostly assigned to cluster "2". Individuals from *Bulgaria* are a notable exception, since individual genotypes are mostly assigned to cluster "1" similar to those of individuals from Western locations (due to the species' introduced origin (Varela, 2000)).

This West – East split is, however, not completely clear, as individuals' genomes are never completely attributed to a single cluster. In fact, most individuals have a considerable part of their genome attributed to both cluster "1" and "2". Furthermore, individuals from some eastern locations have their genomes mostly attributed to cluster "1" (*Var 21*, *Corsica 3, Corsica 11, Corsica 14* and *Puglia 5*), and individuals from *Tunisia* are split between both clusters on a close to 50-50 proportion.

The Q-plot obtained using the "neutral" loci subset (Figure 6.3B) is nearly identical to the one with all the loci, and can be interpreted in the same way.

The Q-plot produced using only the 13.4% "non-neutral" loci (Figure 6.3C), however does bear a different clustering pattern from the previous ones. In this case, the East – West split is more evident, as the individual genomes' attribution to each cluster is not as evenly split, but rather a much more pronounced attribution to either cluster.

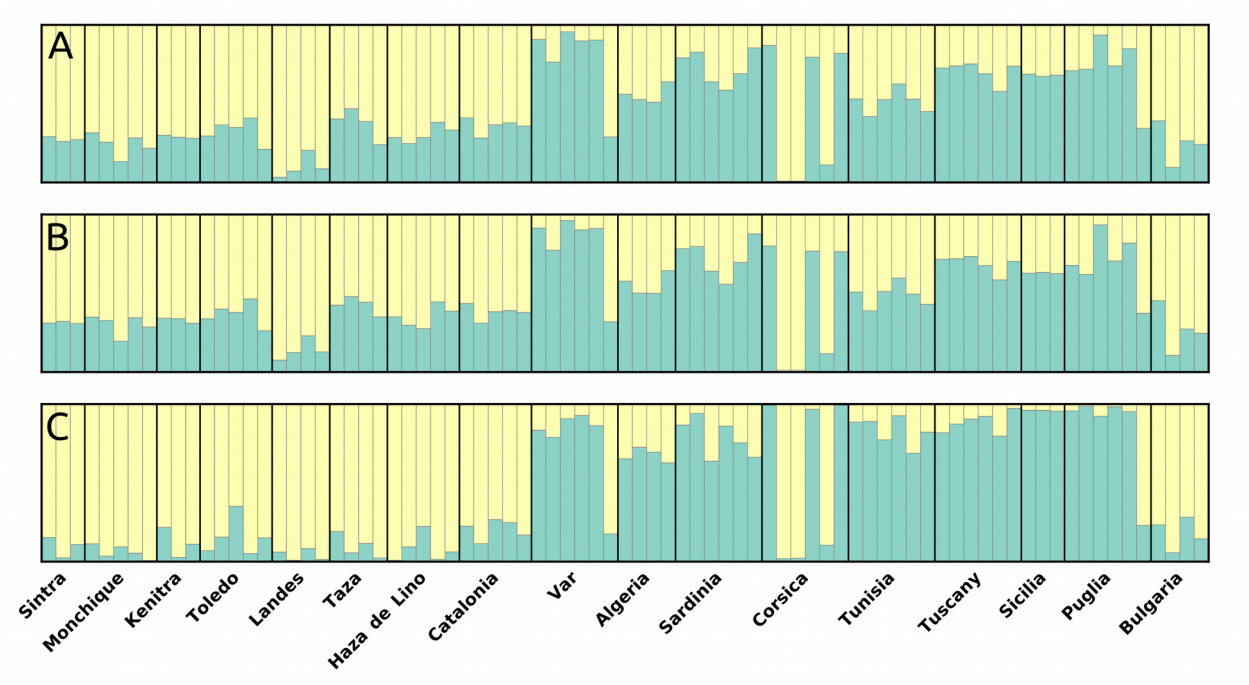

Figure 6.3: *MavericK* clustering plots for K=2. Sampling sites are presented from West to East. "A" is the Q-value plot for the dataset with all loci, "B" is for the dataset with only "neutral" loci, and "C" if for the dataset with only "non-neutral" loci.

The Q-plot obtained from STRUCTURE (Appendix III Figure 2) reveals a generally similar pattern to that of *MavericK* on all datasets.

The PCA clustering method (largest eigenvector values of 0.0431 and 0.0241) is essentially concordant with the previous methods, revealing two loosely defined groupings (Appendix III Figure 3). The first group containing individuals from *Algeria*, *Corsica*, *Puglia*, *Sardinia*, *Sicilia*, *Tuscany*, *Tunisia* and *Var* and the second group containing individuals from *Bulgaria*, *Catalonia*, *Corsica*, *Haza de Lino*, *Kenitra*, *Landes*, *Monchique*,

Puglia, *Sintra*, *Taza*, *Toledo* and *Var*. The groups are loosely defined, because they somewhat resemble an East – West split, but individuals from *Corsica*, *Puglia* and *Var* are present in both groups. Just as in the Q-plots, Bulgarian individuals group with Western ones, despite existing on the edge of the species' Eastern range. Finally, a less pronounced sub-grouping is discernible: one comprising three individuals from *Corsica*; a second comprising all *Landes* individuals, plus three individuals from *Bulgaria*; and a third sub-group consisting of two individuals from *Puglia* and three from *Var* .

4.2 Outlier detection and environmental association

The software *BayeScan* and *SelEstim* detected 32 and 48 outlier SNPs respectively (Appendix III Table 5). The 31 markers common to both methods were considered as being putatively under the effect of natural selection.

Ten of the 31 sequences were matched to *Q. lobata* genome scaffolds. Of these, seven were not annotated, and four could be matched to an annotated region (Table 6.3).

BayPass detected 374 associations between 329 SNPs and 14 of the 16 tested environmental variables (no associations were found with neither "Temperature Annual Range" nor "Precipitation Seasonality"). These associations can be found in Appendix III Table 6. Despite this relatively high number of associations, it is important to note that 72 of these associations were between a SNP and a geospatial variable – 9 associations with "Latitude", 55 with "Longitude" and 8 with "Altitude". Of all environmental variables, the one with most markers associated is "Precipitation of Driest Month" with 79 associations, followed by "Mean Temperature of Driest Quarter" with 51 associations, and "Temperature Seasonality" with 33 associations.

Sequences containing 144 of the 329 markers associated with environmental variables were matched to entries in the *Q. lobata* genome, however, of these only 47 were annotated (Table 6.4).

Since 19 SNP markers are simultaneously associated with an environmental variable and considered outliers, the union of these two SNP sets, resulted in a sub-dataset of 341 SNP markers deemed "non-neutral". The remaining 2206 SNPs were grouped in another sub-dataset, deemed "neutral".

Table 6.3: Summary of best BLAST hit results for loci with SNPs considered outliers against the genome of *Q. lobata*.

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

Table 6.4: Summary of BLAST hits for loci with SNPs associated to one or more environmental variables. “MTDQ” and “MTWQ” stand for “Mean Temperature of Driest Quarter” and “Mean Temperature of Wettest Quarter” respectively.

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

4.3 Risk of non-adaptedness (RONA)

A summary of the RONA analyses for both a low emission scenario (RPC26) and a high emission scenario (RPC85) predictions can be found in Figure 6.4 and Appendix III Table 7. The most represented environmental variables are “Precipitation of Driest Month” (79 SNPs, mean R^2=0.1597), “Mean Temperature of Driest Quarter” (51 SNPs, mean R^2=0.1466) and “Temperature Seasonality” (33 SNPs, mean R^2=0.1545). The values of RONA per sampling site are always higher for RPC85 than for RPC26, except for “Precipitation of Driest Month” in *Tunisia* where RPC85 has a lower RONA than RPC26, and in *Kenitra* where they are the same (the “Precipitation of Driest Month” variable in *Kenitra* is not predicted to change from current conditions (0 mm^2), regardless of the model).

Under the RPC26 predictions, the highest RONA values for “Mean Temperature of Driest Quarter” is *Landes* (0.1482), for “Temperature Seasonality” is *Toledo* (0.0690) and for “Precipitation of Driest Month” is *Landes* (0.0356). Under the RPC85 predictions, *Catalonia* presents the highest values of RONA for “Mean Temperature of Driest Quarter” (0.3921), *Landes* presents the highest RONA for “Precipitation of Driest Month” (0.1157), whereas *Toledo* has the highest value (0.1478) for “Temperature Seasonality”. It is important to note that the high RONA values of *Catalonia* are twice as high as the second highest RONA value on the RPC26 prediction and more than three times as high for RPC85.

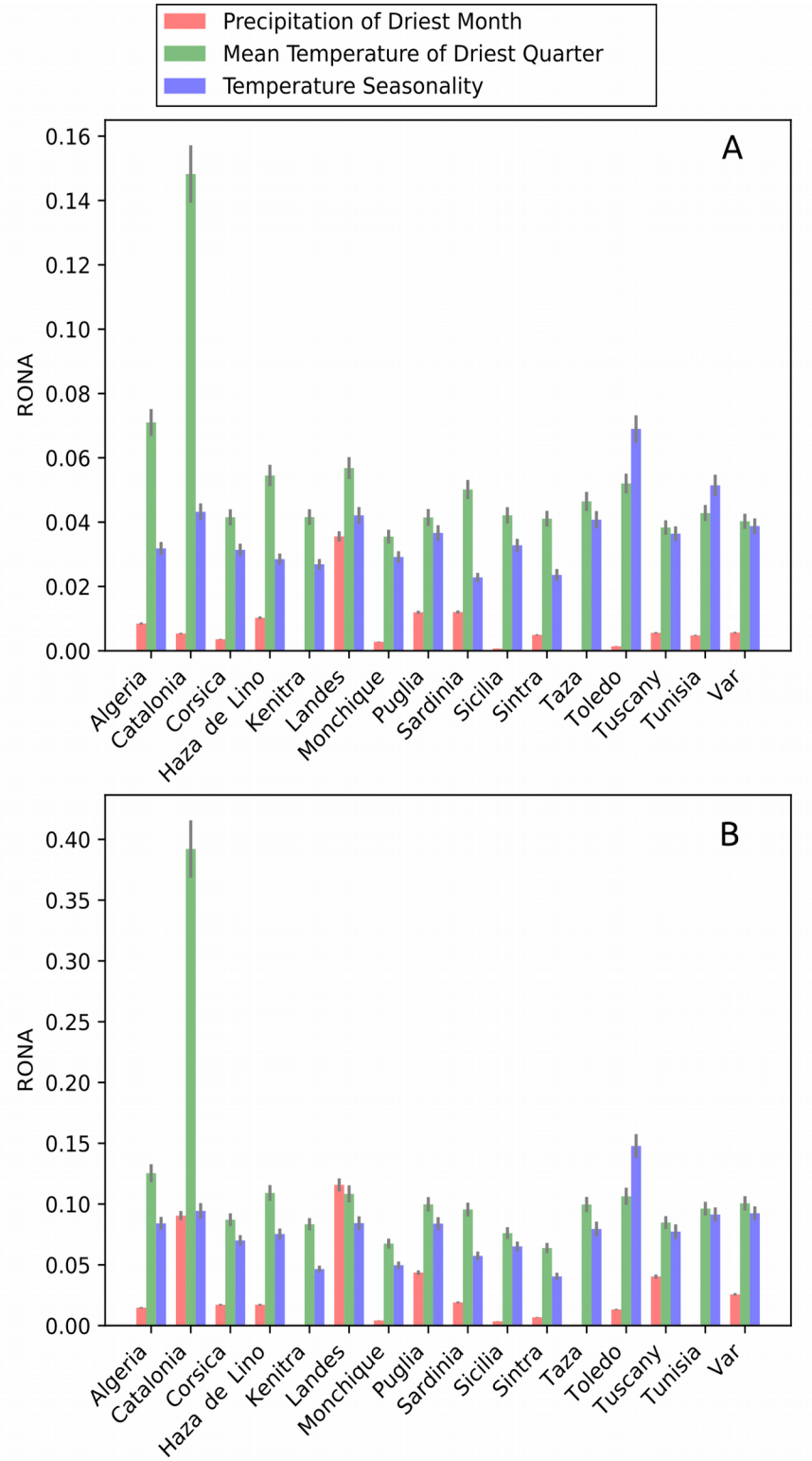

Figure 6.4: Risk of Non-Adaptedness plot for the three SNPs with most associations. Bars represent weighted means (by R^2 value) and lines represent standard error. (A) is the plot for RPC26 and (B) is for RPC85 prediction models.

5 Discussion

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

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

After stringent quality filtering, a set of 2,547 SNPs was used in this study. This number is lower than that of some studies with similar data (Berthouly-Salazar et al., 2016), which obtained ~22k SNPs (albeit using a more frequent cutting enzyme), but still more than (De Kort et al., 2014), which obtained 1630 SNPs, very close to that of (Escudero, Eaton, Hahn, & Hipp, 2014) and (Pais, Whetten, & Xiang, 2017). Even though this number may seem small, in the universe of *Q. suber*'s genome of ~750 Mbp, this is to date the largest number of molecular markers available for this species and represents a step forward to increase the power of population genetics studies.

5.1 Population genetic structure

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

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

Only two sampling sites did not reveal significant deviations from HWE (*Bulgaria* and *Landes*) regarding heterozygote deficiency. No sampling site exhibited heterozygote excess. Although this pattern is not usual, few individual markers deviate from HWE (4.28% reveal excess heterozygotes and 0.90% deficit heterozygotes). This may be due

to the fact that each sampling site does not represent a real biological population, or to non random mating across the species distribution range.

Similar to what is observed with F_{IS}, F_{ST} values are on average (0.0553) higher than on unstructured trees species (0.0125) (Guichoux et al., 2013), but lower than other well structured trees such as eucalypts (0.095) (Cappa et al., 2013). This data supports what the clustering analyses reveal: an incomplete segregation in two clusters, as seen on Figure 6.3. Although clustering analyses using all loci do not provide a clear structuring signal (and the "TI" method clearly favours a scenario of a single large panmitic population), the produced *Q. suber* Q-plots do show some degree of segregation between Western and Eastern individuals.

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

In Figure 6.3C, where the Q-plot was produced using only loci putatively under selection, the division between Western and Eastern individuals is much clearer than in Figure 6.3A and Figure 6.3B. The respective "TI" test also supports this segregation by indicating K=2 as the most likely clustering pattern. Conversely, Figure 6.3B, which was drawn based on loci deemed "neutral", a pattern very similar to the Q-plots of all loci emerges, which supports a scenario of an incomplete segregation between individuals from Eastern and Western locations. This evidence, combined with the relatively low pairwise F_{ST} and F_{IS} values, suggests that there is a balance between local adaptation and gene flow. Whereas the former is responsible for maintaining the species' standing genetic variation and the latter for the species's response to local environmental differences. Intense gene flow would also explain the relatively low proportion of outlier SNPs, which may be counteracting reactions to weak selective pressures. At the same time, this balance may provide the species a relatively large genetic variability to respond to strong selection (De Kort et al., 2014; Kremer et al., 2012).

Data from this work do not seem to support the four glacial refugia hypothesis proposed in Magri et al., (2007). It could be argued that the mentioned refugia had in fact existed, but gene flow would have erased any evidence of their existence, which is thought to have occurred in other tree species (Eidesen et al., 2007), however it seems just as likely to assume a scenario where there were no refugia at all, and the cpDNA segregation is due to the different dispersal capacities of pollen and acorns (Sork, 1984).

Two hypotheses can thus be proposed to explain the observed genetic structure. (1) The observed genetic structure can be explained by the balance between gene flow and local adaptation. In this scenario, these two processes are responsible for both creating and maintaining this level of divergence. This hypothesis seems to be fully supported by the SNP data. (2) The observed pattern can be explained by differential hybridization of *Q. suber* with *Q. cerris* in the East (Bagnoli et al., 2016) and with *Q. ilex* in the West

(Burgarella et al., 2009). In this scenario, the balance between gene flow and adaptation is responsible for maintaining the current divergence levels, but not for their origin. The SNP data in not sufficient to corroborate this hypothesis, and as such, the issue will remain open for investigation.

5.2 Outlier detection and environmental association analyses

The method used to detect outlier loci flagged ~1.2% of the total SNPs, which is in line with what was found on other similar studies (Berdan, Mazzoni, Waurick, Roehr, & Mayer, 2015; Chen et al., 2012). Of the 31 outlier markers found, only four had a match to an annotated location in *Q. lobata*'s genome. This low proportion is likely due to a combination of factors, such as the distance between *Q. suber* and *Q. lobata*, and the incomplete annotation of *Q. lobata*'s genome. On the other hand, it emphasizes the need for more genomic resources in this area, which can potentially provide important functional information of these SNPs in *Q. suber*'s genome, that will at least for now remain unknown. Of particular note is SNP 493, whose sequence is a match to a region of the *Q. lobata* genome, annotated as "Similar to NCS2: S-norcoclaurine synthase 2 (*Papaver somniferum*)", a protein family member usually expressed upon infections and stressful conditions (van Loon, Rep, & Pieterse, 2006). This can be a particularly interesting marker for downstream studies regarding adaptation to infection response.

The environmental association analyses (EAA) served two purposes in this work. On one hand, the reported associations work as a proxy for detecting local adaptation, and on the other hand, allow the attribution of a RONA score to each sampling site. *Q. suber* is known to be very sensitive to precipitation and temperature conditions (Vessella et al., 2017), and as such, it was expected beforehand that some of the markers obtained in this study were to be associated with some of these conditions (Rellstab et al., 2016). In order to understand how important the found associations are for the local adaptation process, it is necessary to understand the putative function of the genomic region where each SNP was found. Querying the available sequences against *Q. lobata*'s genome annotations, has provided insights regarding some of the markers' sequences putative function. The proportion of sequences that were a match to an annotated region, however, is rather small – only ~14.3% of the queried sequences were matched to such regions. Reasons for the low ratio of annotations are likely the same as for the outlier loci.

Of the 47 SNPs associated with an environmental variable that returned hits to annotated regions of *Q. lobata*'s genome, four are likely located in a mitochondrial region, seven in chloroplastidial regions, and 36 in nuclear regions. While all these associations are potentially interesting to explore, doing so falls outside the grander scope of this work. Nevertheless, 6 SNPs are particularly interesting to take a closer look

at, mostly due to how much information is available regarding the identified genomic region function.

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

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

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

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

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

5.3 Risk of non-adaptedness

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

implementation developed for this work, named *pyRONA* suffers from most of the same limitations as the original application, even though it is based on an arguably superior association detection method (Gautier, 2015), but introduces a correction to the average values based on the R^2 of each marker association (by using weighted means). The automation brought by this new implementation, easily allows two different emission scenarios (RPC26 and RPC85) to be tested and compared.

With the exception of *Catalonia*, which seems to have an exceptionally high highest RONA value under both prediction models, the other locations present relatively low RONA values for the tested variables. The variable "Mean Temperature of Driest Quarter" appears to be the tested variable that requires the greatest changes in allele frequencies to ensure adaptation of the species to the local projected changes, although "Temperature Seasonality" is not far behind. These RONA values, are nevertheless smaller than those presented in Rellstab et al. (2016). This might be due to various factors, such as the different variables tested, the geographic scope of the study, the species' respective tolerance to environmental ranges, the differences between species' standing genetic variation, the association detection method, or likely a combination of several of these factors.

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

This study, however, is limited to the considered environmental variables. Other factors that were not included in this work may have a larger effect on *Q. suber*'s RONA.

6 Conclusions

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

Despite the relatively large geographic distances involved, the nuclear markers used in this work indicate lesser genetic structuring than previously thought from cpDNA markers, that clearly segregated the species in several well defined demes (Magri et al., 2007). The SNP data from this work can thus be used to propose two new hypotheses to replace the current view of a genetic structure carved by population recessions and expansions from glacial refugia. The observed genetic structure origin and maintenance

can be explained either by balance between gene flow and local adaptation, or alternatively, differential hybridization of *Q. suber* with *Q. ilex* in the West and *Q. cerris* in the East is responsible for the geographic differences, which are then maintained by the mentioned balance between gene flow and local adaptation (albeit more research is required to confirm this second hypothesis).

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

This study starts by providing a new perspective into the population genetics of *Q. suber*, and, based on this data, suggests an initial conjecture on the species' future, despite the used technique's limitations. Even though studies regarding *Q. suber*'s response to climatic change are not new (Correia et al., 2017; Vessella et al., 2017), this is the first work where this response is investigated from an adaptive perspective. One aspect that could thoroughly improve its reliability would be the availability of more genomic resources, especially a thoroughly annotated genome of the species. Such resource would allow the identification of more markers, and assess the reliability of more associations, which would also allow a more refined method for assessing which loci are more likely to be under the effects of selection. Fortunately, such efforts are underway at the time of writing, and further work in this area should benefit from it in the near future.

7 Acknowledgements

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

8 References

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

Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T., & Curtis-McLane, S. (2008). Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications*, *1*(1), 95–111. doi:10.1111/j.1752-4571.2007.00013.x

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Giorgi, F., & Lionello, P. (2008). Climate change projections for the Mediterranean region. *Global and Planetary Change*, *63*(2–3), 90–104. doi:10.1016/j.gloplacha.2007.09.005

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

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

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*(15), 1965–1978. doi:10.1002/joc.1276

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

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

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

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

Lindner, M., Maroschek, M., Netherer, S., Kremer, A., Barbati, A., Garcia-Gonzalo, J., … Marchetti, M. (2010). Climate change impacts, adaptive capacity, and vulnerability of European forest ecosystems. *Forest Ecology and Management*, *259*(4), 698–709. doi:10.1016/j.foreco.2009.09.023

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Varela, M. C. (2000). *Evaluation of genetic resources of cork oak for appropriate use in breeding and gene conservation strategies*. EC FAIR Programme.
Verity, R., & Nichols, R. A. (2016). Estimating the Number of Subpopulations (K) in Structured Populations. *Genetics*, *203*(4), 1827–1839. doi:10.1534/genetics.115.180992
Vessella, F., López-Tirado, J., Simeone, M. C., Schirone, B., & Hidalgo, P. J. (2017). A tree species range in the face of climate change: cork oak as a study case for the Mediterranean biome. *European Journal of Forest Research*, 1–15. doi:10.1007/s10342-017-1055-2
Vitalis, R., Gautier, M., Dawson, K. J., & Beaumont, M. A. (2014). Detecting and Measuring Selection from Gene Frequency Data. *Genetics*, *196*(3), 799–817. doi:10.1534/genetics.113.152991
Walther, G.-R., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T. J. C., … Bairlein, F. (2002). Ecological responses to recent climate change. *Nature*, *416*(6879), 389–395. doi:10.1038/416389a
Xia, Y., Suzuki, H., Borevitz, J., Blount, J., Guo, Z., Patel, K., … Lamb, C. (2004). An extracellular aspartic protease functions in Arabidopsis disease resistance signaling. *The EMBO Journal*, *23*(4), 980–988. doi:10.1038/sj.emboj.7600086
Yang, K.-Z., Xia, C., Liu, X.-L., Dou, X.-Y., Wang, W., Chen, L.-Q., … Ye, D. (2009). A mutation in THERMOSENSITIVE MALE STERILE 1, encoding a heat shock protein with DnaJ and PDI domains, leads to thermosensitive gametophytic male sterility in Arabidopsis. *The Plant Journal*, *57*(5), 870–882. doi:10.1111/j.1365-313X.2008.03732.x
Zoldos, V., Papes, D., Brown, S. C., Panaud, O., & Siljak-Yakovlev, S. (1998). Genome size and base composition of seven *Quercus* species: inter- and intra-population variation. *Genome*, *41*(2), 162–168. doi:10.1139/g98-006

Chapter 7

Final Remarks

1 General overview

According to the most current scentific literature, global climatic change will cause a general increase in temperature which may vary from 1 to 4°C in the next 70 years, depending on the prediction model and location (IPCC, 2014; Oreskes, 2004; Walther et al., 2002). These changes, which are also bound to affect precipitation patterns (Beniston et al., 2007), are likely to have a negative impact on current biodiversity (IPCC, 2014). Mediterranean forest tree species are expected to be more affected than those of temperate and boreal areas, due to their sensitivity to precipitation seasonality (Briceño-Elizondo, Garcia-Gonzalo, Peltola, Matala, & Kellomäki, 2006; Loustau et al., 2005; Saxe, Cannell, Johnsen, Ryan, & Vourlitis, 2002).

The scientific questions addressed in this thesis arise from the need to understand how species may respond to global climatic change. In particular, it focuses on the need to better understand the potential response of *Quercus suber* to such rapid alterations. For that, it was necessary to gain a better understanding on the evolutionary history of this Mediterranean species, as well as identifying the effects of natural selection across the species' genome (Kremer, Potts, & Delzon, 2014).

This large scale assessment was possible due to the "mainstreaming" of *Second generation Sequencing* technologies, often called *Next Generation Sequencing*, or NGS. This technology has changed the way biological problems are thought of and approached (Kumar, Banks, & Cloutier, 2012). Current low cost per sequenced base pair of these methods allows for non-model organisms like *Q. suber* to be sequenced in a larger scope than ever (Schuster, 2008). Even for organisms whose whole genome sequencing remains a challenge, alternative approaches like *Reduced Representation Libraries*, in which Genotyping by Sequencing (GBS) is included, allow the scientific community to study their genome at a whole new level (Van Tassell et al., 2008), for a fraction of previous costs.

New types of data require new methodologies, and in the case of NGS, researchers now have to deal both with new data types, and also with unprecedented amounts of it (Markowetz, 2017). Software that was used without issue on typical Sanger sequencing data, or microsatellite loci did not scale to what NGS demands. Workflows and protocols that were considered good practice for "small" data became either impractical, or could not even be applied to the new reality.

Bioinformatics, which was already a rising discipline, took centre stage in dealing with the terabytes of FASTQ and similar files generated by the new sequencing machines (Markowetz, 2017). Bioinformatics software now had to scale, perform and run stable on datasets that were unimaginable just a few years prior. This led to an increase in the quality patterns researchers now hold bioinformatics software (and the respective

documentation) to (Leprevost, Barbosa, Francisco, Perez-Riverol, & Carvalho, 2014). Automation became a necessity and reproducibility became a particularly important issue.

This is the context under which the works for this thesis were performed, and it is under these conditions that it was set out to make a contribution for science.

2 Field contributions

The works on this thesis were performed in pursue of three main objectives. In order to optimize the accomplishment of the two evolutionary biology objectives, the technical, bioinformatics focused objective had to be fulfilled first.

2.1 Technical issues

In place of writing scripts and programs tailored specifically for each specific analysis, t generic, reusable software, that can be used by the larger scientific community was developed and made available.

Software in scientific research is now a crucial part of the scientific method, and it can be argued that anything less than the full release of source code is indefensible for any scientific results (Ince, Hatton, & Graham-Cumming, 2012). The programs conceived and built for this thesis are released under open source licenses with the code being made available, which is crucial from a reproducibility standpoint. These programs have been built with the specific purpose of improving the automation, performance and reproducibility value of workflows, and respect the guidelines recommended in Leprevost et al., 2014.

2.1.1 4Pipe4

Chapter 2 describes the *4Pipe4* program which was developed for this thesis. It was a fundamental piece in mining the SNP markers from a 454 dataset with no reference sequence nor strain information available. It outperformed every other method available in the false positives metric at the time of publishing. It was used to generate the EST data that Chapter 5 builds upon.

The 454 technology has been surpassed by other technologies, until finally having been declared “end of life” by Roche in 2016. *4Pipe4*, however, still stands as an example for automation and reproducibility – a considerable improvement over the too frequent alternative approach of using “custom scripts” (Barbazuk, Emrich, Chen, Li, & Schnable, 2007; Tollenaere et al., 2012).

2.1.2 NCBI Mass Sequence Downloader

Chapter 3 describes a program that is used to download large datasets from the NCBI databases. It can run either in the command line (CLI) or with the use of a Graphical User Interface (GUI), and is a far more reliable way to obtain large sequences from NCBI in large scale than performing downloads of several gigabytes from a web browser. It was especially useful during the exploratory phase of Chapter 5 and 6, where the size of the sequence databases to query was particularly important.

NCBI_Mass_sequence_downloader, however, is catered to a niche audience. It is, nonetheless, a good example of software that improves automation and, consequently, the research's reproducibility value (Gentleman et al., 2004) as well as the performance of downstream analyses (due to dataset reduction).

2.1.3 Structure_threader

Chapter 4 regards a program designed to automate and accelerate the runs of population genetics clustering software. It was fundamental for this work, because it reduced human intervention in an involved and mistake prone process (thus reducing random error and increasing reproducibility value), sped up the analysis process by a factor of approximately 8x, and produced the Q-value plots that can be seen in Chapter 5 and 6.

Structure_threader is putatively the greatest contribution of this thesis to the field of population genetics/genomics. By running the analyses wrapped in *Structure_threader*, any user will instantly find improvements in terms of automation, performance and reproducibility, when compared to the alternatives. *Structure_threader*, however is not designed for the absolute novice. It requires some degree of familiarity with the wrapped programs, and minimal knowledge of how to use a CLI interface. In this respect, the program could be improved with the addition of an optional GUI.

2.1.4 Further automation

In order to improve the reproducibility value of Chapter 6, most of the analysis process has been automated. Every step of the analysis performed after the SNP calling stage has been performed in an automated manner recurring to a "Makefile". This method ensures that the performed analyses are easy to repeat and that the entire procedure is logged. Furthermore, every change to the process is recorded via *git* version control system. But in order to maximize the reproducibility value of the study, a docker container including all the software required to perform the analysis is also provided. This automated approach ensures that at least for the foreseeable future, the analyses performed here will remain completely reproducible. Although the container and

"Makefile" are tuned for this specific analysis, they are built in a way that is flexible enough to allow them to be broadly used as a base for other, similar procedures.

2.2 Evolutionary biology questions

Evolutionary biology is the main focus of this thesis. As such, some of the most important "breakthroughs" of this thesis fall under this category. Below are some of the arguably most relevant ones.

2.2.1 New *Q. suber* evolutionary history hypotheses

In Chapter 5 and 6, the evolutionary history of cork oak was revisited, and approached from a genomic perspective, which revealed a different reality from what had been known, based essentially on plastidial molecular markers. Whereas previous studies had classified *Q. suber* as a species clearly split in four different lineages (Magri et al., 2007; Simeone et al., 2009), evidence from high density SNP data obtained and analysed in this thesis revealed a much lesser level of segregation (albeit inconclusively in Chapter 5). The data also suggest that gene flow in *Q. suber* is much more prevalent, at least from a nuclear genome standpoint, than what was previously considered.

In these two chapters, the role of local adaptation was assessed across most of *Quercus suber*'s distribution. Association analyses revealed several SNP markers whose allele frequencies across sampling sites suggest that they are either under the effects of natural selection or, at the very least, correlated with a certain environmental variable (functional analyses of the gene where these markers lie frequently suggested causation). Likewise, outlier analyses were used to identify loci putatively under selection. These loci were subsequently used to generate sub-datasets of "neutral" and "non-neutral" markers. Although data from Chapter 5 did not allow for any substantiated hypothesising regarding the species' evolutionary history, combining information from selection detection analyses with population structure in Chapter 6 did. In this chapter, clustering analyses performed exclusively with markers putatively under selection hinted at a two cluster scenario, roughly segregating Eastern and Western individuals. This pattern is not evident when all loci are analysed together, in which case a single cluster is indicated as the most likely scenario. These results cannot be explained by previously proposed hypotheses, which ponted glacial refugia as the origin of the four *Q. suber* lineages (Magri et al., 2007). Therefore two new hypotheses were proposed in this chapter as explanations for the cork oak's evolutionary history:

1. The observed genetic structure can be explained by a balance between gene flow and local adaptation. In this scenario, these two processes are responsible for both generating and maintaining the observed level of divergence.

2. The observed genetic structure can be explained by differential hybridization of *Q. suber* with *Q. cerris* in the East and with *Q. ilex* in the West. In this scenario, the balance between gene flow and adaptation is responsible for maintaining the current divergence levels, but not for their origin.

It should be reinforced that hypothesis 1 is fully supported by the SNP data, and although the data does not reject hypothesis 2, further research regarding the mentioned species' introgression levels is required to confirm it.

2.2.2 New identified markers

In both Chapter 5 and 6, several SNP markers were associated with environmental variables. These markers and the respective flanking regions were compared to sequences from publicly available resources – Genbank (Sayers et al., 2010) and the genome annotation of *Quercus lobata* (Sork et al., 2016). Although far from a complete assessment, this resulted in the creation of a small database of markers that are not only associated with an environmental variable, but whose putative molecular function is known. This data can be used as a basis for future "case – control" studies in *Q. suber*.

2.2.3 Risk of Non-adaptedness

During the analysis of *Genotyping by Sequencing* (GBS) data (Chapter 6), it was possible to assess a *Risk of Non-Adaptedness* (RONA) (Rellstab et al., 2016) value for each sampled site. Although this was not one of the main objectives of this thesis, it is likely one of its most interesting results, despite its known limitations. These RONA results are particularly interesting, not only because it is the first time such metric is assessed for the species, as it is also the first time the adaptation potential of the species is considered from a genetic point of view. RONA analyses suggest that the species is likely genetically equipped to survive in most of its current distribution. Recent works have tried to answer this very question, but from an exclusively ecological point of view (Correia, Bugalho, Franco, & Palmeirim, 2017; Vessella, López-Tirado, Simeone, Schirone, & Hidalgo, 2017), with both providing different predictions (although not antagonistic) regarding the future of cork oak, none of them incompatible with the genetic perspective revealed in Chapter 6.

3 Future perspectives

Science does not often provide definitive answers, but frequently provides further questions on the studied subject. The results of this thesis are no exception, but not only do they provide a way to answer a new set of questions, they also hint at an answer to those that were placed. Results presented in this thesis provide many avenues of future exploration regarding the questions it proposed to provide answers for.

As is the case for most code ever written, improvements for the programs produced in the context of this thesis are always possible – this is software, after all. *NCBI Mass Sequence Downloader* can be given more options to simplify the querying of the NCBI databases, and more formats other than FASTA can be supported. This will, of course, depend on user interest. *Structure_threader* can also in the future integrate with more clustering programs, and eventually be given a GUI for easier use by novice users. Most importantly, the use of these programs by the community is already a gain for automation and reproducibility, since they are improvements on manual and often error, and consequently bias, prone procedures.

Regarding the automated analysis process: in this thesis, the choice was to use a version controlled "Makefile" and a docker container. In an ideal world, approaches such as this start being adopted in large scale by the scientific community, as it visibly increases the reproducibility value of the analyses. But this process too, can be further improved – for instance, the generation of environmental data can still be further automated, and although using "GNU Make" was a good choice, other, more recent "build systems" exist.

As for evolutionary biology perspectives – the species' evolutionary history just became a controversial theme. Further exploration in this front could be performed via Whole Genome Sequencing (WGS), using a smaller sampling than what was used in the works presented here, since what is now known can help make a more informed sampling strategy. A reference genome of *Q. suber* would be of paramount importance to make such an approach a successful project.

It would also be interesting to further understand how much local adaptation influences *Q. suber*'s evolutionary history and current standing genetic variation. A study could be devised using the markers mined in Chapters 5 and 6, applied in transects of natural populations, case-control studies, or even provenance trials as "candidate genes" in order to gain further insights on the role of this type of natural selection from a functional point of view.

Regarding the RONA values of cork oak, this is a method that could certainly be improved upon. This could be done by expanding beyond the limits of linear regressions and correlations, and by attempting to associate multiple loci of "small effect" with environmental variables, instead of just looking for "large effect" markers as is currently implemented in the method. Furthermore, the genetic perspective presented here, should be combined with the ecological approaches performed elsewhere to maximize forecast accuracy.

As an ending statement – this thesis shows some of what can be done when combining bioinformatics with evolutionary biology, a combination that is only likely to grow tighter. Moving evolutionary biology from genes to whole genomes is certain to bring

new answers to many old (and new) questions, and the challenges that will be coupled with these advances, are bound to be, at least, just as interesting.

4 References

Barbazuk, W. B., Emrich, S. J., Chen, H. D., Li, L., & Schnable, P. S. (2007). SNP discovery via 454 transcriptome sequencing. *The Plant Journal*, *51*(5), 910–918. doi:10.1111/j.1365-313X.2007.03193.x

Beniston, M., Stephenson, D. B., Christensen, O. B., Ferro, C. A. T., Frei, C., Goyette, S., … Woth, K. (2007). Future extreme events in European climate: an exploration of regional climate model projections. *Climatic Change*, *81*(S1), 71–95. doi:10.1007/s10584-006-9226-z

Briceño-Elizondo, E., Garcia-Gonzalo, J., Peltola, H., Matala, J., & Kellomäki, S. (2006). Sensitivity of growth of Scots pine, Norway spruce and silver birch to climate change and forest management in boreal conditions. *Forest Ecology and Management*, *232*(1–3), 152–167. doi:10.1016/j.foreco.2006.05.062

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

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., … Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, *5*, R80. doi:10.1186/gb-2004-5-10-r80

Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, *482*(7386), 485–488. doi:10.1038/nature10836

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

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

Kumar, S., Banks, T. W., & Cloutier, S. (2012). SNP Discovery through Next-Generation Sequencing and Its Applications. *International Journal of Plant Genomics*, *2012*. doi:10.1155/2012/831460

Leprevost, F. da V., Barbosa, V. C., Francisco, E. L., Perez-Riverol, Y., & Carvalho, P. C. (2014). On best practices in the development of bioinformatics software. *Bioinformatics and Computational Biology*, *5*, 199. doi:10.3389/fgene.2014.00199

Loustau, D., Bosc, A., Colin, A., Ogee, J., Davi, H., Francois, C., … Delage, F. (2005). Modeling climate change effects on the potential production of French plains forests at the sub-regional level. *Tree Physiology*, *25*(7), 813–823. doi:10.1093/treephys/25.7.813

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

Markowetz, F. (2017). All biology is computational biology. *PLOS Biology*, *15*(3), e2002050. doi:10.1371/journal.pbio.2002050

Oreskes, N. (2004). BEYOND THE IVORY TOWER: The Scientific Consensus on Climate Change. *Science*, *306*(5702), 1686–1686. doi:10.1126/science.1103618

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

Saxe, H., Cannell, M. G. R., Johnsen, Ø., Ryan, M. G., & Vourlitis, G. (2002). Tree and forest functioning in response to global warming: Tansley review no. 123. *New Phytologist*, *149*(3), 369–399. doi:10.1046/j.1469-8137.2001.00057.x

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., … Ye, J. (2010). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *38*(Database issue), D5-16.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, *5*(1), 16–18.

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

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

Tollenaere, C., Susi, H., Nokso-Koivisto, J., Koskinen, P., Tack, A., Auvinen, P., … Laine, A.-L. (2012). SNP Design from 454 Sequencing of Podosphaera plantaginis Transcriptome Reveals a Genetically Diverse Pathogen Metapopulation with High Levels of Mixed-Genotype Infection. *PLoS ONE*, *7*(12), e52492. doi:10.1371/journal.pone.0052492

Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., … Sonstegard, T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, *5*(3), 247–252. doi:10.1038/nmeth.1185

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

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

Appendix I

Supplementary Material for Chapter 4

1 Tables

Table 1: Software run times. “Speed up” value for single thread shows the run time in seconds.

STRUCTURE (Structure_threader)				
Replica 1				
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	02:41:45	04:23:35	04:25:47	03:02:44
2	01:21:51	02:11:21	02:13:16	01:31:53
4	00:41:14	01:06:01	01:07:45	00:48:20
6	00:39:18	00:48:02	00:51:29	N/A
8	00:36:41	00:33:38	00:36:13	N/A
10	N/A	N/A	00:38:14	N/A
12	N/A	N/A	00:50:36	N/A
14	N/A	N/A	00:41:21	N/A
16	N/A	N/A	00:33:20	N/A
Replica 2				
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	02:40:30	04:23:15	04:27:33	03:01:27
2	01:21:52	02:11:33	02:13:01	01:31:59
4	00:41:42	01:06:12	01:07:36	00:49:01
6	00:39:07	00:47:58	00:51:15	N/A
8	00:36:43	00:32:58	00:37:48	N/A
10	N/A	N/A	00:37:20	N/A
12	N/A	N/A	00:50:05	N/A
14	N/A	N/A	00:46:30	N/A
16	N/A	N/A	00:33:48	N/A
Average				
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	02:41:08	04:23:25	04:26:40	03:02:06
2	01:21:52	02:11:27	02:13:09	01:31:56
4	00:41:28	01:06:07	01:07:41	00:48:41
6	00:39:13	00:48:00	00:51:22	N/A
8	00:36:42	00:33:18	00:37:01	N/A
10	N/A	N/A	00:37:47	N/A
12	N/A	N/A	00:50:21	N/A
14	N/A	N/A	00:43:56	N/A
16	N/A	N/A	00:33:34	N/A
“Speed up”				
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	9668	15805	16000	10926
2	1.97	2.00	2.00	1.98
4	3.89	3.98	3.94	3.74
6	4.11	5.49	5.19	N/A
8	4.39	7.91	7.20	N/A
10	N/A	N/A	7.06	N/A
12	N/A	N/A	5.30	N/A
14	N/A	N/A	6.07	N/A
16	N/A	N/A	7.94	N/A

STRUCTURE (ParallelStructure)				
Replica 1				
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	02:41:45	04:23:35	04:25:47	03:02:44
2	01:22:27	02:38:56	02:14:36	01:45:06
4	00:43:06	01:37:33	01:08:39	00:52:08
6	00:41:59	01:21:55	00:54:29	N/A
8	00:39:40	00:44:41	00:38:37	N/A
10	N/A	N/A	00:41:33	N/A
12	N/A	N/A	00:53:42	N/A
14	N/A	N/A	00:55:52	N/A
16	N/A	N/A	00:37:41	N/A
Replica 2				
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	02:40:30	04:23:15	04:27:33	03:01:27
2	01:22:38	02:38:02	02:14:45	01:47:35
4	00:42:42	01:33:42	01:08:41	00:54:32
6	00:41:14	01:19:23	00:54:42	N/A
8	00:39:42	00:43:51	00:39:11	N/A
10	N/A	N/A	00:43:28	N/A
12	N/A	N/A	00:54:36	N/A
14	N/A	N/A	00:56:18	N/A
16	N/A	N/A	00:37:56	N/A
Average				
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	02:41:08	04:23:25	04:26:40	03:02:06
2	01:22:33	02:14:41	02:14:41	01:46:21
4	00:42:54	01:08:40	01:08:40	00:53:20
6	00:41:37	00:54:36	00:54:36	N/A
8	00:39:41	00:44:16	00:38:54	N/A
10	N/A	N/A	00:42:31	N/A
12	N/A	N/A	00:54:09	N/A
14	N/A	N/A	00:56:05	N/A
16	N/A	N/A	00:37:49	N/A
“Speed up”				
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	9668	15805	16000	10926
2	1.95	1.96	1.98	1.71
4	3.76	3.84	3.88	3.41
6	3.87	4.82	4.88	N/A
8	4.06	5.95	6.86	N/A
10	N/A	N/A	6.27	N/A
12	N/A	N/A	4.92	N/A
14	N/A	N/A	4.75	N/A
16	N/A	N/A	7.05	N/A

Table 1(cont.): Software run times. “Speed up” value for single thread shows the run time in seconds.

STRUCTURE (StrAuto)				fastStructure (Structure_threader)			
Replica 1				Replica 1			
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	02:41:45	04:23:35	04:25:47	03:02:44	00:55:11	01:20:13	01:40:29	00:46:06
2	01:23:17	02:12:14	02:14:18	01:39:20	00:28:08	00:43:39	00:52:09	00:23:47
4	00:41:56	01:06:09	01:08:30	00:48:49	00:12:41	00:24:36	00:28:01	00:13:45
6	00:42:01	00:51:12	00:53:59	N/A	00:12:43	00:15:17	00:17:51	N/A
8	00:37:11	00:35:39	00:37:53	N/A	00:13:43	00:13:36	00:15:32	N/A
10	N/A	N/A	00:41:44	N/A	10 N/A	N/A	00:15:42	N/A
12	N/A	N/A	00:52:49	N/A	12 N/A	N/A	00:21:42	N/A
14	N/A	N/A	00:48:45	N/A	14 N/A	N/A	00:24:39	N/A
16	N/A	N/A	00:33:42	N/A	16 N/A	N/A	00:18:31	N/A
Replica 2					Replica 2			
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	02:40:30	04:23:15	04:27:33	03:01:27	00:49:29	01:28:15	01:40:08	00:49:01
2	01:22:49	02:11:53	02:14:43	01:32:00	00:26:26	00:41:59	00:56:44	00:23:24
4	00:42:07	01:06:29	01:08:16	00:49:04	00:13:18	00:22:19	00:28:07	00:12:44
6	00:42:05	00:51:20	00:53:51	N/A	00:12:42	00:14:44	00:18:38	N/A
8	00:37:13	00:35:41	00:38:15	N/A	00:13:17	00:13:23	00:13:34	N/A
10	N/A	N/A	00:42:05	N/A	10 N/A	N/A	00:16:53	N/A
12	N/A	N/A	00:54:26	N/A	12 N/A	N/A	00:19:25	N/A
14	N/A	N/A	00:49:52	N/A	14 N/A	N/A	00:19:12	N/A
16	N/A	N/A	00:34:12	N/A	16 N/A	N/A	00:19:06	N/A
Average					Average			
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	02:41:08	04:23:25	04:26:40	03:02:06	00:52:20	01:24:14	01:40:19	00:47:34
2	01:23:03	02:12:04	02:14:31	01:35:40	00:27:17	00:42:49	00:54:27	00:23:36
4	00:42:02	01:06:19	01:08:23	00:48:57	00:13:00	00:23:28	00:28:04	00:13:15
6	00:42:03	00:51:16	00:53:55	N/A	00:12:43	00:15:01	00:18:15	N/A
8	00:37:12	00:35:40	00:38:04	N/A	00:13:30	00:13:30	00:14:33	N/A
10	N/A	N/A	00:41:55	N/A	10 N/A	N/A	00:16:18	N/A
12	N/A	N/A	00:53:38	N/A	12 N/A	N/A	00:20:34	N/A
14	N/A	N/A	00:49:19	N/A	14 N/A	N/A	00:21:56	N/A
16	N/A	N/A	00:33:57	N/A	16 N/A	N/A	00:18:49	N/A
“Speed up”					“Speed up”			
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	9668	15805	16000	10926	1 3140	5054	6019	2854
2	1.94	1.99	1.98	1.90	2 1.92	1.97	1.84	2.02
4	3.83	3.97	3.90	3.72	4 4.03	3.59	3.57	3.59
6	3.83	5.14	4.95	N/A	6 4.12	5.61	5.50	N/A
8	4.33	7.39	7.01	N/A	8 3.88	6.24	6.89	N/A
10	N/A	N/A	6.36	N/A	10 N/A	N/A	6.15	N/A
12	N/A	N/A	4.97	N/A	12 N/A	N/A	4.88	N/A
14	N/A	N/A	5.41	N/A	14 N/A	N/A	4.57	N/A
16	N/A	N/A	7.85	N/A	16 N/A	N/A	5.33	N/A

Table 1(cont.): Software run times. “Speed up” value for single thread shows the run time in seconds.

MavericK (Structure_threader)				
Replica 1				
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	00:16:47	00:28:34	00:30:39	00:19:01
2	00:08:41	00:14:14	00:15:25	00:09:34
4	00:04:27	00:07:10	00:07:50	00:05:05
6	00:03:56	00:05:12	00:05:50	N/A
8	00:03:18	00:03:39	00:04:10	N/A
10	N/A	N/A	00:04:09	N/A
12	N/A	N/A	00:05:34	N/A
14	N/A	N/A	00:05:05	N/A
16	N/A	N/A	00:04:39	N/A
Replica 2				
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	00:16:51	00:28:28	00:30:30	00:18:59
2	00:08:42	00:14:14	00:15:25	00:09:33
4	00:04:25	00:07:09	00:07:54	00:05:05
6	00:03:56	00:05:13	00:05:50	N/A
8	00:03:17	00:03:37	00:04:09	N/A
10	N/A	N/A	00:04:10	N/A
12	N/A	N/A	00:05:32	N/A
14	N/A	N/A	00:05:04	N/A
16	N/A	N/A	00:04:39	N/A
Average				
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	00:16:49	00:28:31	00:30:35	00:19:00
2	00:08:42	00:14:14	00:15:25	00:09:34
4	00:04:26	00:07:10	00:07:52	00:05:05
6	00:03:56	00:05:13	00:05:50	N/A
8	00:03:18	00:03:38	00:04:10	N/A
10	N/A	N/A	00:04:10	N/A
12	N/A	N/A	00:05:33	N/A
14	N/A	N/A	00:05:05	N/A
16	N/A	N/A	00:04:39	N/A
“Speed up”				
Threads	i7 4700MQ	Intel Xeon E5-2609	Intel Xeon E5520	i5 3350P
1	1009	1711	1835	1140
2	1.93	2.00	1.98	1.99
4	3.79	3.98	3.89	3.74
6	4.28	5.47	5.24	N/A
8	5.10	7.85	7.34	N/A
10	N/A	N/A	7.34	N/A
12	N/A	N/A	5.51	N/A
14	N/A	N/A	6.02	N/A
16	N/A	N/A	6.58	N/A

Appendix II

Supplementary Material for Chapter 5

1 Tables

Table 1: Environmental variable values for each sampling site (Coordinates in decimal degrees).

Vars	Lat.	Long.	Alt	BIO1 Annual Mean Temp.	BIO2 Mean Diurnal Range	BIO3 Isotherm.	BIO4 Temp. Seasonality	BIO6 Min Temp. of Coldest Month	BIO7 Temp. Annual Range	BIO8 Mean Temp. of Wettest Quarter	BIO9 Mean Temp. of Driest Quarter	BIO10 Mean Temp. of Warmest Quarter	BIO12 Annual Precipitation	BIO13 Precipitation of Wettest Month	BIO14 Precipitation of Driest Month	BIO15 Precipitation Seasonality
Algeria	36,54	7,15	742	143.0	114.0	38.0	6097	16.0	293	69	219	226	744	123.0	6.0	59.0
Catalonia	41,85	2,533	663	127.0	71.0	31.0	5435	25.0	226	175	62	199	887	99.0	46.0	21.0
Cosrica	41,616	8,966	337	141.0	59.0	29.0	5019	55.0	198	121	206	209	639	92.0	9.0	48.0
Haza de Lino	36,833	-3,3	1316	123.0	117.0	38.0	6224	-3.0	304	60	207	208	573	71.0	9.0	49.0
Kenitra	34,083	-6,583	138	181.0	107.0	45.0	4509	69.0	237	134	237	240	553	107.0	0.0	75.0
Landes	43,75	-1,333	9	136.0	86.0	40.0	4613	36.0	212	110	176	195	1286	158.0	64.0	23.0
Monchique	37,316	-8,566	723	133.0	93.0	43.0	4183	44.0	214	93	186	191	731	111.0	4.0	63.0
Puglia	40,566	17,666	128	159.0	87.0	35.0	5720	49.0	247	133	234	234	575	75.0	18.0	40.0
Sardinia	39,083	8,85	899	123.0	80.0	33.0	5651	23.0	241	71	197	200	825	125.0	9.0	55.0
Sicily	37,116	14,5	273	161.0	76.0	34.0	5187	63.0	219	144	227	230	432	82.0	3.0	65.0
Sintra	38,75	-9,416	161	149.0	71.0	42.0	3472	72.0	168	112	190	194	819	127.0	5.0	64.0
Taza	34,2	-4,25	318	187.0	122.0	39.0	6284	49.0	311	111	271	271	521	86.0	1.0	66.0
Toledo	39,366	-5,35	545	151.0	118.0	36.0	6733	20.0	319	77	241	241	469	59.0	5.0	45.0
Tuscany	42,416	11,95	173	151.0	88.0	33.0	6151	35.0	266	121	231	231	709	97.0	20.0	34.0
Tunisia	36,95	8,85	20	182.0	98.0	40.0	5430	73.0	245	116	249	254	825	152.0	3.0	70.0
Var	43,133	6,25	12	151.0	83.0	36.0	5085	46.0	225	126	216	217	726	99.0	10.0	43.0

Table 2: Contig names of the assembly used to mine the data, followed by the gene annotation and SNP name.

Contig name	Represented gene	SNP name
c2092	T-complex protein 1 subunit epsilon-like	QSN001
c886	sphingoid long-chain bases kinase 1-like	QSN002
c9224	alanyl-tRNA synthetase-like	QSN003
c924	fructose-1,6-bisphosphatase	QSN004
rep_c13347	Galactosyltransferase family protein	QSN005
rep_c13353	Early light-induced protein	QSN006
rep_c13589	Glutaredoxin	QSN007
rep_c13804	Carboxylesterase	QSN008
rep_c14161	NADH-ubiquinone oxidoreductase	QSN009
rep_c17000	fatty acid desaturase	QSN010
rep_c22049	Chlorophyll a/b binding protein	QSN011
rep_c22615	nuclear transcription factor Y subunit A-7-like	QSN012
rep_c29429	alcohol dehydrogenase class-3-like	QSN013
rep_c29438	extracellular calcium sensing receptor	QSN014
rep_c31383	Uncharacterized	QSN015
rep_c32576	cytochrome oxidase subunit I (COI)	QSM001

Table 3: Pairwise F_{ST} values for all loci (lower triangular) and neutral loci (upper triangular) only datasets.

	Neutrals																		
Pops	Sintra	Monchique	Kenitra	Toledo	Taza	Haza de Lino	Landes	Catalonia	Var	Algeria	Sardinia	Corsica	Tunisia	Tuscany	Sicilia	Puglia	Bulgaria	*Q. ilex*	*Q. coccifera*
Sintra	0	0.02	0.13	0	0.04	0.03	0.1	0.02	0.08	0.04	0.07	0.03	0.05	0.03	0.11	0.1	0.01	0.25	0.33
Monchique	0.02	0	0.05	0	0.02	-0.01	0.1	0.02	0.04	0.03	0.02	0	0	-0.01	-0.01	0.02	0.02	0.34	0.41
Kenitra	0.07	0.03	0	0.05	0.03	0.06	0.12	0.1	0.12	0.13	0.04	0.1	0.04	0.02	0.02	0.04	0.12	0.32	0.39
Toledo	0.01	0	0.06	0	0.01	0.01	0.11	0.02	0.08	0.03	0.09	0.02	0.03	0.01	0.05	0.07	0.03	0.31	0.39
Taza	0.05	0.02	0.05	0.02	0	0.02	0.08	0.02	0.03	0.01	0.06	0.02	0.01	0.03	0.03	0.03	0.04	0.31	0.38
Haza de Lino	0.01	-0.01	0.04	0	0.02	0	0.15	0.03	0.03	0.03	0	0	0	-0.01	0	0	0.03	0.27	0.34
Landes	0.16	0.13	0.16	0.16	0.1	0.13	0	0.08	0.12	0.08	0.13	0.09	0.18	0.14	0.16	0.16	0.04	0.43	0.53
Catalonia	0.03	0.03	0.07	0.04	0.01	0.01	0.11	0	0.01	0.03	0	0	0.03	0.03	0.04	0	0.03	0.3	0.37
Var	0.13	0.1	0.21	0.13	0.1	0.08	0.18	0.12	0	0.1	0.06	0.03	0.1	0.1	0.14	0.03	0.14	0.25	0.33
Algeria	0.01	0	0.02	0	-0.02	0	0.09	-0.01	0.03	0	-0.01	0	0.01	0	0.01	-0.01	0	0.29	0.38
Sardinia	0.03	0.09	0.19	0.03	0	0.07	0.13	0.09	0.03	0.03	0	0.04	0.01	0.01	0.02	0.01	0.07	0.3	0.36
Corsica	0.01	-0.01	0.04	0.01	0.01	-0.01	0.1	0.03	0.01	0.01	0.01	0	0.01	0	0.01	0.02	0.02	0.34	0.41
Tunisia	0.04	0.01	0.11	0.04	0.02	0.02	0.12	0.04	0.04	0.01	0.04	0.01	0	0.01	0.04	0.05	0.02	0.18	0.27
Tuscany	0.04	0.03	0.11	0.03	0.03	0.04	0.1	0.06	0.06	0.02	0.01	0.03	0.01	0	0.04	0.03	0.02	0.27	0.34
Sicilia	0.07	0.05	0.1	0.04	0.02	0.05	0.15	0.06	0.05	0.03	0.09	0.04	-0.01	-0.01	0	0.05	0.06	0.32	0.39
Puglia	0.05	0.07	0.13	0.04	-0.01	0.05	0.16	0.08	0.03	0.04	-0.01	0.03	0	0.02	0.01	0	0.1	0.25	0.33
Bulgaria	0.02	0	0.02	0.02	0.03	0.01	0.05	0.03	0.05	0	0.02	-0.01	0.04	0	0.03	0.04	0	0.33	0.42
Q. Ilex	0.38	0.4	0.33	0.39	0.37	0.3	0.42	0.35	0.36	0.43	0.35	0.38	0.25	0.3	0.38	0.36	0.47	0	0.03
Q. coccifera	0.43	0.46	0.39	0.45	0.42	0.35	0.5	0.4	0.42	0.48	0.4	0.43	0.32	0.35	0.43	0.42	0.53	0.03	0
	All loci																		

Table 4: Significant associations between genotypes and environmental variables identified by *Samβada*.

All locations

Marker	Environmental variable	Loglikelihood	Gscore	WaldScore	Efron	McFadden	McFaddenAdj	CoxSnell	Nagelkerke	AIC	BIC	Beta_0	Beta_1
QSN002_CC	Longitude	-202.8947	17.3839	16.5228	0.0549	0.0411	0.0316	0.0537	0.0200	409.79	428.80	-0.6211	0.0590
QSN004_TT	Annual precipitation	-184.2542	23.6584	18.9342	0.0817	0.0603	0.0501	0.0724	0.0315	372.51	391.52	1.6112	-0.0035
QSN007_CC	Annual precipitation	-151.9607	27.3310	24.0417	0.1346	0.0825	0.0704	0.0879	0.0470	307.92	326.70	-3.7095	0.0035
QSN007_CC	Precipitation of wettest month	-154.3532	22.5459	21.8855	0.1134	0.0681	0.0560	0.0731	0.0385	312.71	331.48	-1.6576	0.0348
QSN007_CC	Precipitation of driest month	-155.5226	20.2073	19.0313	0.0886	0.0610	0.0489	0.0658	0.0343	315.05	333.82	-3.5799	0.0228
QSN008_CC	Latitude	-197.4191	22.8535	21.2239	0.0712	0.0547	0.0451	0.0702	0.0271	398.84	417.84	8.1973	-0.1969
QSN008_CC	Mean diurnal range	-199.0764	19.5390	18.1127	0.0657	0.0468	0.0372	0.0603	0.0231	402.15	421.15	-2.1562	0.0291
QSN008_CC	Isothermality	-199.3411	19.0095	17.6672	0.0620	0.0455	0.0359	0.0587	0.0224	402.68	421.68	-4.0192	0.1225
QSN011_CC	Longitude	-155.9218	37.4050	31.6006	0.1125	0.1071	0.0956	0.1126	0.0619	315.84	334.83	-0.9895	-0.1053
QSN011_GG	Longitude	-175.2156	35.4616	31.0857	0.1102	0.0919	0.0815	0.1071	0.0493	354.43	373.42	-1.2114	0.0933

West

Marker	Environmental variable	Loglikelihood	Gscore	WaldScore	Efron	McFadden	McFaddenAdj	CoxSnell	Nagelkerke	AIC	BIC	Beta_0	Beta_1
QSN007_CC	Precipitation of wettest month	-69.2538	35.3398	25.4537	0.2787	0.2033	0.1803	0.2176	0.1186	142.51	158.39	-5.4845	0.0423
QSN007_CC	Annual precipitation	-72.8401	28.1673	22.6345	0.2503	0.1620	0.1390	0.1777	0.0922	149.68	165.56	-3.9927	0.0040
QSN007_CC	Precipitation of driest month	-76.1961	21.4553	19.8940	0.1998	0.1234	0.1004	0.1384	0.0685	156.39	172.27	-1.6271	0.0363
QSN007_CC	Altitude	-72.9770	27.8936	18.8700	0.2531	0.1604	0.1374	0.1761	0.0912	149.95	165.83	0.3729	-0.0032
QSN012_CC	Latitude	-84.4641	25.2431	20.0323	0.1615	0.1300	0.1094	0.1494	0.0711	172.93	189.13	10.9117	-0.3097

East

Marker	Environmental variable	Loglikelihood	Gscore	WaldScore	Efron	McFadden	McFaddenAdj	CoxSnell	Nagelkerke	AIC	BIC	Beta_0	Beta_1
c924_TT	Bio12	-81.0333	31.9724	25.9382	0.2232	0.1648	0.1442	0.1832	0.0929	166.07	182.32	4.7175	-0.0083

2 Data

Data 1: 4Pipe4rc configuration file used for mining the SNPs from the EST dataset

https://gist.github.com/StuntsPT/52c64503388710f7c4f38f193c81cbc2

Data 2: VCF file containing the raw SNP data.

https://gist.github.com/StuntsPT/5582de33110c60cf95e9b123f12647ac

Data 3: 4Pipe4 SNP mining report.

https://stuntspt.github.io/EST_data_mining_reports/index.html

3 Figures

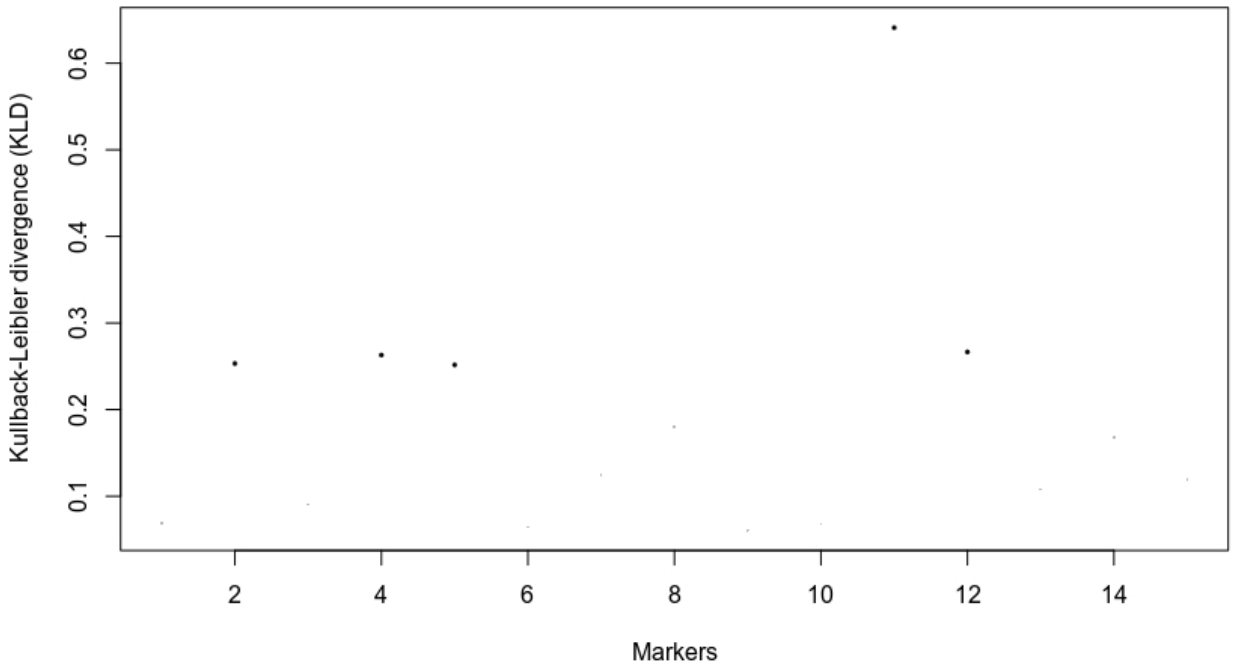

Figure 1: *SelEstim* outlier plot

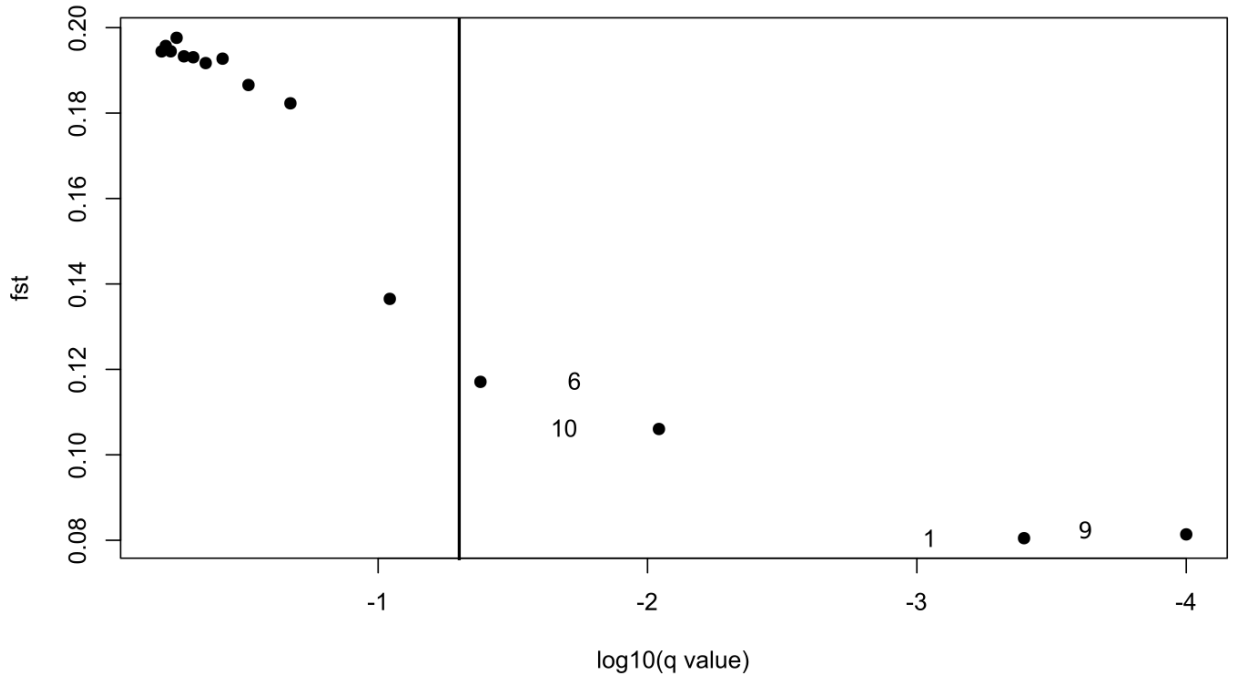

Figure 2: *Bayescan* outlier detection plot

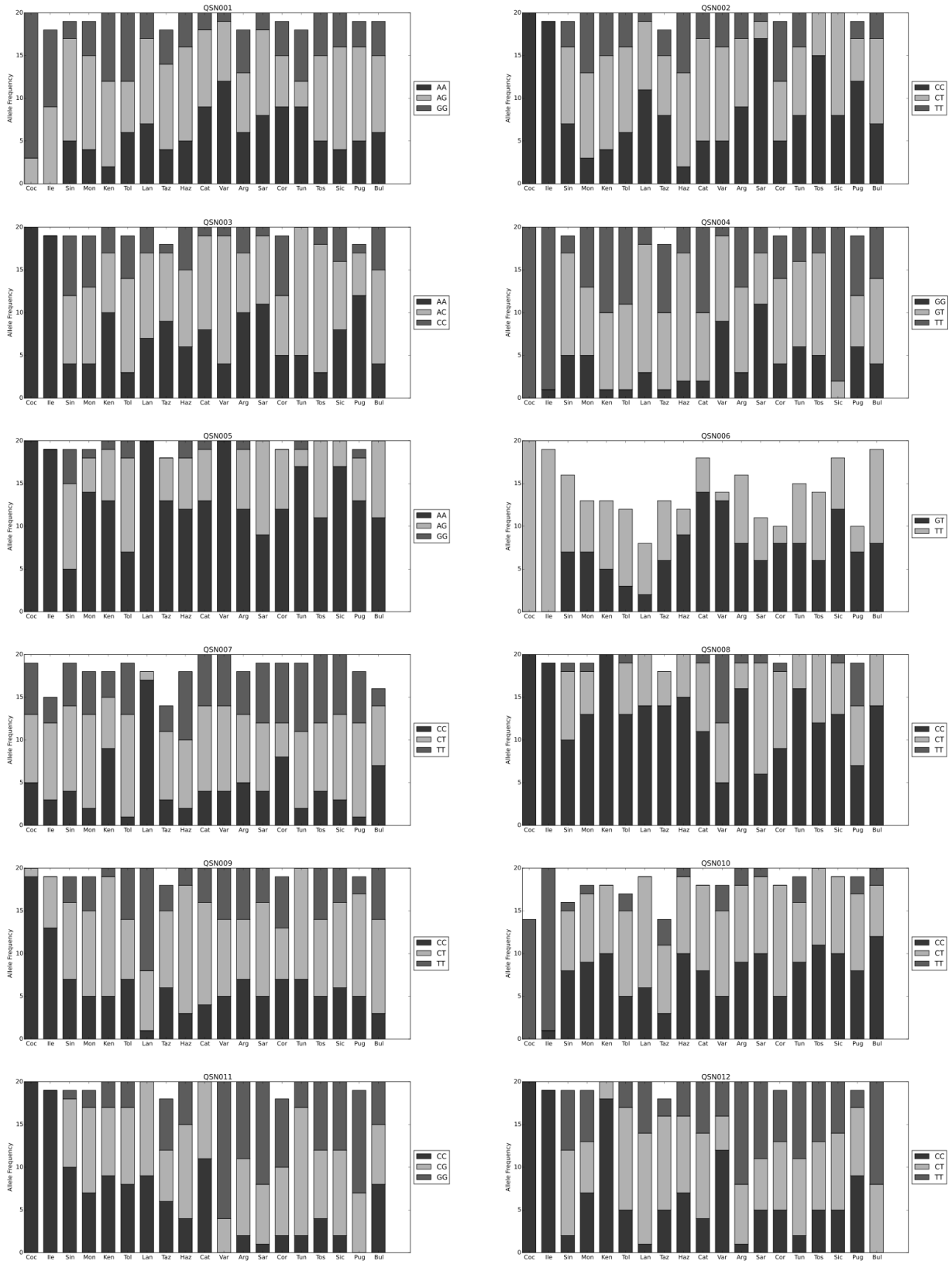

Figure 3: Allele frequency plots

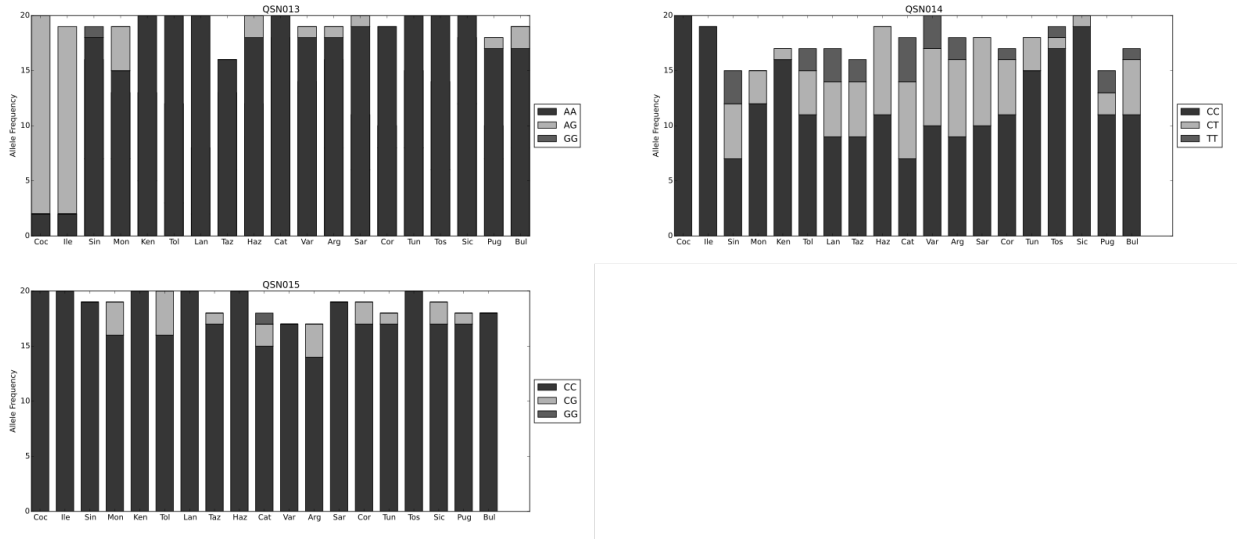

Figure 3 (continued): Allele frequency plots

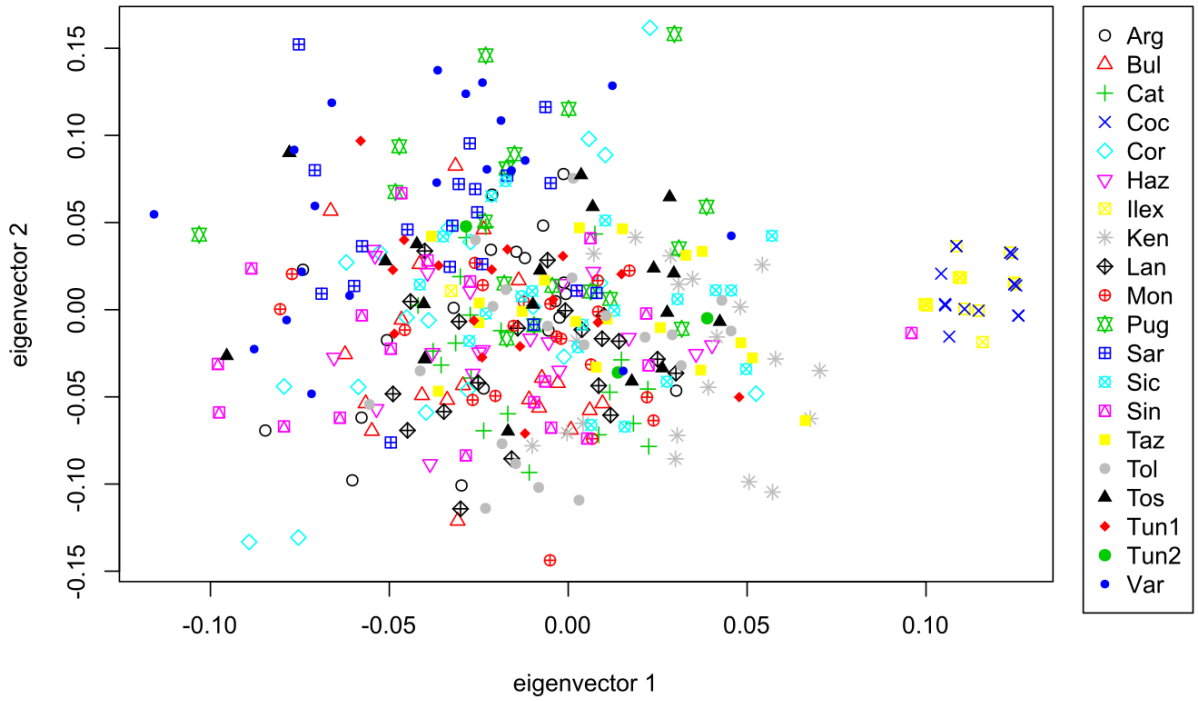

Figure 4: PCA plot of the SNP matrix

Appendix III

Supplementary Material for Chapter 6

1 Tables

Table 1: Bioclimatic variables for the RPC26 prediction model.

Sample	Latitude (Dec. Deg)	Longitude (Dec. Deg)	Altitude (m)	BIO1 Annual Mean Temp.	BIO2 Mean Diurnal Range	BIO3 Isotherm.	BIO4 Temp. Seasonality	BIO6 Min Temp. of Coldest Month	BIO7 Temp. Annual Range	BIO8 Mean Temp. of Wettest Quarter	BIO9 Mean Temp. of Driest Quarter	BIO10 Mean Temp. of Warmest Quarter	BIO12 Annual Precip.	BIO13 Precip. of Wettest Month	BIO14 Precip. of Driest Month	BIO15 Precip. Seasonality
Algeria	36.54	7.15	742	170.67	116	37.78	6352.56	40.33	302.78	99.22	251.11	257.44	635.78	102.22	4.67	58.44
Catalonia	41.85	2.53	663	148.78	72.11	30.22	5774.78	44.89	236.56	168.89	132.33	226.89	856.33	98.33	45.11	22.33
Corsica	41.62	8.97	337	158.11	59	28.33	5261.11	69.67	205.56	137.89	225.56	229.56	650.67	94.11	8.44	50
Haza de Lino	36.83	-3.30	1316	145.44	118	37.56	6479.67	16.22	311.56	87.22	232.78	234.67	499.78	72.78	7.33	52
Kenitra	34.08	-6.58	138	197.11	109.33	44.56	4714.89	82.56	243	149.67	255.67	259.11	483.56	89.67	0	75.44
Landes	43.75	-1.33	9	151.56	88.33	39.11	4935.44	48.78	223.56	105.44	202.11	216.22	1239.56	152.44	57.56	24.67
Monchique	37.32	-8.57	723	146.89	93.44	42	4406.33	55.89	219.11	99.56	203	207.22	694.78	112.56	3.56	65
Puglia	40.57	17.67	128	176.67	87.22	33.33	6006.11	64.89	257.33	145	253.33	256.11	564	74.89	16.11	42.11
Sardinia	39.08	8.85	899	144.78	80.11	32.11	5825.78	43.11	246.67	91.67	220.11	224.56	775.33	117	7.11	57.44
Sicilia	37.12	14.50	273	177.33	76.22	33.11	5438.11	76.11	227.56	158.67	246.56	250.56	424.89	82	2.89	68.11
Sintra	38.75	-9.42	161	165	71.78	41.11	3683.56	85	173.56	127.11	208.67	212.56	781.67	130.56	4.22	68.89
Taza	34.20	-4.25	318	206	124.89	38.22	6619.33	63.56	321.89	135.89	294.67	295.89	449.78	79.78	1	68.11
Toledo	39.37	-5.35	545	166.22	123.22	36.22	7261.22	30.44	336.33	104.22	264.56	264.67	466.67	68.89	4.78	49.89
Tuscany	42.42	11.95	173	168.78	90.78	32.33	6430	51.11	276.67	142.78	248.22	254	711.89	94.67	19.11	34.44
Tunisia	36.95	8.85	20	195.56	102	38.56	5831.11	81.33	262.22	132	268.56	274.33	796.33	137.44	3.78	67.56
Var	43.13	6.25	12	167.11	82.78	35	5381.89	61.44	234	142.67	234.33	239	731.11	100.44	9.11	45.78

Table 2: Bioclimatic variables for the RPC85 prediction model.

Sample	Latitude (Dec. Deg)	Longitude (Dec. Deg)	Altitude (m)	BIO1 Annual Mean Temp.	BIO2 Mean Diurnal Range	BIO3 Isotherm.	BIO4 Temp. Seasonality	BIO6 Min Temp. of Coldest Month	BIO7 Temp. Annual Range	BIO8 Mean Temp. of Wettest Quarter	BIO9 Mean Temp. of Driest Quarter	BIO10 Mean Temp. of Warmest Quarter	BIO12 Annual Precip.	BIO13 Precip. of Wettest Month	BIO14 Precip. of Driest Month	BIO15 Precip. Seasonality
Algeria	36.54	7.15	742	192.22	119.67	37.11	6754.11	55.11	317	118.11	278	284.33	504.11	86.44	3.67	61.22
Catalonia	41.85	2.53	663	169	72.78	28.56	6219.67	59.44	249.89	156.22	248.44	253.22	711.67	86.78	29.78	26.78
Corsica	41.62	8.97	337	176.33	59.22	27.22	5557.67	85.67	212.78	146	247.78	252.44	573.56	92.56	6.22	57.67
Haza de Lino	36.83	-3.30	1316	165.89	120.11	36.67	6839.89	30	323.67	92.78	257.33	259.11	378.89	57	6.11	53
Kenitra	34.08	-6.58	138	217.33	111.89	44.89	4865.22	97.22	247.33	166.89	275.89	279.78	352.22	65.22	0	74.89
Landes	43.75	-1.33	9	170.11	91.22	38.11	5307	63.22	237.33	117.56	230.78	240.89	1118.33	147.89	42.22	34.22
Monchique	37.32	-8.57	723	162.44	94.78	41.78	4563.78	67.22	224.56	111.56	219.56	223.78	547.56	110.44	3.33	73.11
Puglia	40.57	17.67	128	197.22	88	32.56	6381	82.33	267.78	159.89	279.56	282.67	495.89	72.78	11.11	51.22
Sardinia	39.08	8.85	899	161.89	80.67	31.22	6090.44	58	254.22	101.33	240.89	246.22	636.67	99.89	6	60.56
Sicilia	37.12	14.50	273	195.67	76.89	32.44	5691.67	91.89	234.22	172.89	265.78	272.89	354.89	71.56	2.44	69.78
Sintra	38.75	-9.42	161	179.33	72.44	40.56	3837	97	176.67	135.11	223.67	228	652.56	119.67	3.89	75
Taza	34.20	-4.25	318	230.44	128.67	37.67	6976.67	79.11	336.33	156.78	321.56	323.67	332.56	58.44	0.89	67.56
Toledo	39.37	-5.35	545	190.67	127.67	34.89	7945.78	43.33	360.22	110.22	296.67	297.56	377.22	56.89	2.89	53.56
Tuscany	42.42	11.95	173	190.56	91.11	31.33	6870.11	69.22	288.33	152.22	273.22	282.22	633.78	94.56	13.33	45.11
Tunisia	36.95	8.85	20	215	103.89	37.67	6138.33	96.44	272.33	146.67	292.56	298	648.56	117	3.11	70
Var	43.13	6.25	12	186.33	83.78	33.78	5792.67	77.22	245.33	145.78	261.44	264.56	665.33	101.11	5.89	54.67

Table 3: Bioclimatic variables for the “current conditions” as extracted from the worldclim.org database.

Sample	Latitude (Dec. Deg)	Longitude (Dec. Deg)	Altitude (m)	BIO1 Annual Mean Temp.	BIO2 Mean Diurnal Range	BIO3 Isotherm.	BIO4 Temp. Seasonality	BIO6 Min Temp. of Coldest Month	BIO7 Temp. Annual Range	BIO8 Mean Temp. of Wettest Quarter	BIO9 Mean Temp. of Driest Quarter	BIO10 Mean Temp. of Warmest Quarter	BIO12 Annual Precip.	BIO13 Precip. of Wettest Month	BIO14 Precip. of Driest Month	BIO15 Precip. Seasonality
Algeria	36.54	7.15	742	143	114	38	6097	16	293	69	219	226	744	123	6	59
Catalonia	41.85	2.53	663	127	71	31	5435	25	226	175	62	199	887	99	46	21
Corsica	41.62	8.97	337	141	59	29	5019	55	198	121	206	209	639	92	9	48
Haza de Lino	36.83	-3.30	1316	123	117	38	6224	-3	304	60	207	208	573	71	9	49
Kenitra	34.08	-6.58	138	181	107	45	4509	69	237	134	237	240	553	107	0	75
Landes	43.75	-1.33	9	136	86	40	4613	36	212	110	176	195	1286	158	64	23
Monchique	37.32	-8.57	723	133	93	43	4183	44	214	93	186	191	731	111	4	63
Puglia	40.57	17.67	128	159	87	35	5720	49	247	133	234	234	575	75	18	40
Sardinia	39.08	8.85	899	123	80	33	5651	23	241	71	197	200	825	125	9	55
Sicilia	37.12	14.50	273	161	76	34	5187	63	219	144	227	230	432	82	3	65
Sintra	38.75	-9.42	161	149	71	42	3472	72	168	112	190	194	819	127	5	64
Taza	34.20	-4.25	318	187	122	39	6284	49	311	111	271	271	521	86	1	66
Toledo	39.37	-5.35	545	151	118	36	6733	20	319	77	241	241	469	59	5	45
Tuscany	42.42	11.95	173	151	88	33	6151	35	266	121	231	231	709	97	20	34
Tunisia	36.95	8.85	20	182	98	40	5430	73	245	116	249	254	825	152	3	70
Var	43.13	6.25	12	151	83	36	5085	46	225	126	216	217	726	99	10	43

Table 4: Pairwise F_{ST} values between all individuals, grouped by sampling sites.

	Algeria	**Bulgaria**	**Catalonia**	**Corsica**	**Haza de Lino**	**Kenitra**	**Landes**	**Monchique**	**Puglia**	**Sardinia**	**Sicilia**	**Sintra**	**Taza**	**Toledo**	**Tunisia**	**Tuscany**
Bulgaria	0.0429															
Catalonia	0.0607	0.0446														
Corsica	0.0661	0.0386	0.0599													
Haza de Lino	0.0556	0.0368	0.0317	0.0565												
Kenitra	0.0835	0.0812	0.0825	0.1051	0.0637											
Landes	0.0832	0.0182	0.0667	0.0509	0.0588	0.1186										
Monchique	0.0494	0.0326	0.0422	0.0502	0.0174	0.0574	0.0531									
Puglia	0.0544	0.0655	0.0634	0.0618	0.0625	0.0995	0.0920	0.0636								
Sardinia	0.0424	0.0593	0.0557	0.0527	0.0553	0.0947	0.0998	0.0551	0.0312							
Sicilia	0.0544	0.0780	0.0669	0.0477	0.0681	0.1059	0.1057	0.0710	0.0384	0.0313						
Sintra	0.0402	0.0380	0.0316	0.0526	0.0163	0.0478	0.0501	0.0038	0.0591	0.0572	0.0630					
Taza	0.0357	0.0377	0.0412	0.0462	0.0197	0.0403	0.0531	0.0126	0.0515	0.0550	0.0566	0.0074				
Toledo	0.0377	0.0296	0.0262	0.0499	0.0128	0.0632	0.0474	0.0200	0.0528	0.0451	0.0629	0.0152	0.0233			
Tunisia	0.0096	0.0383	0.0539	0.0677	0.0513	0.0822	0.0721	0.0450	0.0607	0.0402	0.0517	0.0394	0.0381	0.0385		
Tuscany	0.0346	0.0583	0.0522	0.0545	0.0568	0.0891	0.0898	0.0536	0.0289	0.0050	0.0360	0.0508	0.0503	0.0467	0.0449	
Var	0.0768	0.0906	0.0920	0.0743	0.0855	0.1225	0.1220	0.0865	0.0580	0.0453	0.0690	0.0725	0.0767	0.0814	0.0767	0.0665

Table 5: List of SNPs detected by *Bayescan* and *SelEstim* as outlier loci. Common markers are represented in **bold** typeface.

Bayescan	Selestim
37	**37**
	70
	79
	99
145	**145**
180	**180**
249	**249**
	381
	387
490	**490**
497	**497**
619	**619**
671	**671**
	673
749	**749**
768	**768**
818	**818**
	825
	1011
1132	**1132**
1139	**1139**
1208	**1208**
1258	**1258**
	1282
1293	**1293**
1297	**1297**
1338	**1338**
	1345
1353	**1353**
1513	**1513**
1528	**1528**
1589	**1589**
	1601
	1618
	1646
1658	**1658**
1793	**1793**
1874	**1874**
1896	**1896**
	1920
	2083
2102	**2102**
2126	
	2140
	2195
2270	**2270**
2419	**2419**
2427	**2427**
2514	**2514**

Table 6: List of SNPs with associations to environmental variables

Environmental Variable	SNP name	BF(dB)	Environmental Variable	SNP name	BF(dB)	Environmental Variable	SNP name	BF(dB)
Latitude	762	16.78	Longitude	1881	32.31	Isothermality	254	16.72
Latitude	1271	19.5	Longitude	1934	18.23	Isothermality	626	19.78
Latitude	1503	15.07	Longitude	1982	17.72	Isothermality	714	18.61
Latitude	1568	15.41	Longitude	1984	20.62	Isothermality	787	18.3
Latitude	1589	16.9	Longitude	2085	24.02	Isothermality	860	15.5
Latitude	1601	23.35	Longitude	2120	28.9	Isothermality	896	16.24
Latitude	1645	15.07	Longitude	2142	17.42	Isothermality	960	19.61
Latitude	1714	20.53	Longitude	2185	18.14	Isothermality	1071	18.41
Latitude	1957	24.8	Longitude	2374	21.16	Isothermality	1084	18.86
Longitude	17	20.13	Longitude	2413	23.45	Isothermality	1175	15.5
Longitude	37	40.37	Longitude	2422	24.96	Isothermality	1245	15.9
Longitude	42	30.39	Longitude	2458	16.58	Isothermality	1276	15.5
Longitude	70	23.55	Altitude	408	15.46	Isothermality	1419	21.07
Longitude	99	25.98	Altitude	524	17.76	Isothermality	1458	15.88
Longitude	141	17.81	Altitude	778	19	Isothermality	1609	15.96
Longitude	174	19.99	Altitude	872	19.21	Isothermality	1680	27.95
Longitude	180	17.76	Altitude	1663	15.05	Isothermality	1706	15.9
Longitude	199	15.61	Altitude	1906	16.78	Isothermality	1774	16.96
Longitude	226	16.07	Altitude	2244	18.57	Isothermality	1920	15.77
Longitude	234	16.42	Altitude	2419	15.46	Isothermality	1922	15.98
Longitude	249	16.94	Annual Mean Temp.	235	19.99	Isothermality	1937	27.26
Longitude	258	15.28	Annual Mean Temp.	401	20.83	Isothermality	1995	15.44
Longitude	346	19.64	Annual Mean Temp.	499	15.28	Isothermality	2046	17.64
Longitude	361	22.84	Annual Mean Temp.	619	27.03	Isothermality	2113	16.52
Longitude	406	16.17	Annual Mean Temp.	638	20.78	Isothermality	2525	18.94
Longitude	409	23.17	Annual Mean Temp.	649	15.3	Temp. Seasonality	188	16.09
Longitude	513	18.25	Annual Mean Temp.	733	19.12	Temp. Seasonality	199	15.19
Longitude	524	22.08	Annual Mean Temp.	891	16.5	Temp. Seasonality	238	35.2
Longitude	551	17.74	Annual Mean Temp.	895	16.56	Temp. Seasonality	267	16.9
Longitude	593	22.66	Annual Mean Temp.	960	15.3	Temp. Seasonality	291	16.4
Longitude	671	32.31	Annual Mean Temp.	975	16.48	Temp. Seasonality	370	17.33
Longitude	673	17.7	Annual Mean Temp.	1258	17.42	Temp. Seasonality	375	15.53
Longitude	708	16.6	Annual Mean Temp.	1303	16.82	Temp. Seasonality	460	19.99
Longitude	791	20.06	Annual Mean Temp.	1336	23.97	Temp. Seasonality	468	16.07
Longitude	883	20.9	Annual Mean Temp.	1663	15.53	Temp. Seasonality	469	18.73
Longitude	985	18.09	Annual Mean Temp.	1871	29.8	Temp. Seasonality	677	18.2
Longitude	1016	16.84	Annual Mean Temp.	1956	15.39	Temp. Seasonality	847	25.68
Longitude	1125	21.86	Annual Mean Temp.	2089	15.28	Temp. Seasonality	996	20.69
Longitude	1133	18.84	Annual Mean Temp.	2094	15.85	Temp. Seasonality	1065	16.05
Longitude	1198	17.61	Annual Mean Temp.	2170	16.3	Temp. Seasonality	1156	16.6
Longitude	1208	16.46	Annual Mean Temp.	2449	15.5	Temp. Seasonality	1241	15.46
Longitude	1276	18.21	Mean Diurnal Range	299	52.96	Temp. Seasonality	1313	21.16
Longitude	1320	21.84	Mean Diurnal Range	595	17.46	Temp. Seasonality	1322	15.12
Longitude	1336	23.37	Mean Diurnal Range	625	16.09	Temp. Seasonality	1398	18.98
Longitude	1513	20.43	Mean Diurnal Range	1052	22.46	Temp. Seasonality	1406	18.61
Longitude	1523	17.76	Mean Diurnal Range	1073	21.41	Temp. Seasonality	1549	19.97
Longitude	1528	17.74	Mean Diurnal Range	1225	21.07	Temp. Seasonality	1615	19.89
Longitude	1575	18.16	Mean Diurnal Range	1874	25.3	Temp. Seasonality	1674	22.8
Longitude	1658	18.27	Mean Diurnal Range	2068	18.16	Temp. Seasonality	1693	20.43
Longitude	1672	15.16	Mean Diurnal Range	2213	16.86	Temp. Seasonality	1814	16.78
Longitude	1700	23.29	Isothermality	88	21.02	Temp. Seasonality	1953	17.74
Longitude	1771	15.55	Isothermality	127	25.84	Temp. Seasonality	2098	17.66

Table 6 (cont.): List of SNPs with associations to environmental variables

Environmental Variable	SNP name	BF (dB)	Environmental Variable	SNP name	BF (dB)	Environmental Variable	SNP name	BF (dB)
Temp. Seasonality	2151	20.16	Mean Temp. of Driest Quarter	366	17.7	Annual Precip.	357	15.55
Temp. Seasonality	2253	18.52	Mean Temp. of Driest Quarter	391	18.16	Annual Precip.	527	20.99
Temp. Seasonality	2492	32.93	Mean Temp. of Driest Quarter	442	16.88	Annual Precip.	571	29.31
Temp. Seasonality	2506	16.56	Mean Temp. of Driest Quarter	497	52.96	Annual Precip.	711	18.57
Temp. Seasonality	2530	15.92	Mean Temp. of Driest Quarter	618	17.38	Annual Precip.	861	31.46
Temp. Seasonality	2539	15.21	Mean Temp. of Driest Quarter	625	20.85	Annual Precip.	1012	20.16
Min Temp. of Coldest Month	140	21.11	Mean Temp. of Driest Quarter	638	15.59	Annual Precip.	1019	18.03
Min Temp. of Coldest Month	145	17.21	Mean Temp. of Driest Quarter	657	15.48	Annual Precip.	1229	16.07
Min Temp. of Coldest Month	238	18.29	Mean Temp. of Driest Quarter	757	17.25	Annual Precip.	1253	26.03
Min Temp. of Coldest Month	597	17.1	Mean Temp. of Driest Quarter	771	20.55	Annual Precip.	1260	16.21
Min Temp. of Coldest Month	651	16.54	Mean Temp. of Driest Quarter	773	16.46	Annual Precip.	1261	18.32
Min Temp. of Coldest Month	825	19.38	Mean Temp. of Driest Quarter	892	27.77	Annual Precip.	1506	18.47
Min Temp. of Coldest Month	913	15.72	Mean Temp. of Driest Quarter	910	17.55	Annual Precip.	1513	15.07
Min Temp. of Coldest Month	1188	17.17	Mean Temp. of Driest Quarter	976	16.03	Annual Precip.	1693	15.79
Min Temp. of Coldest Month	1381	15.32	Mean Temp. of Driest Quarter	1007	17.94	Annual Precip.	1749	15.53
Min Temp. of Coldest Month	1621	18.29	Mean Temp. of Driest Quarter	1171	16.9	Annual Precip.	1776	38.76
Min Temp. of Coldest Month	1623	17.29	Mean Temp. of Driest Quarter	1181	16.62	Annual Precip.	1959	17.38
Min Temp. of Coldest Month	1646	20.79	Mean Temp. of Driest Quarter	1258	26.87	Annual Precip.	2162	16.19
Min Temp. of Coldest Month	2094	15.81	Mean Temp. of Driest Quarter	1267	17.57	Annual Precip.	2195	16.76
Min Temp. of Coldest Month	2218	15.57	Mean Temp. of Driest Quarter	1338	16.76	Annual Precip.	2263	18.94
Min Temp. of Coldest Month	2264	16.07	Mean Temp. of Driest Quarter	1391	15.25	Annual Precip.	2539	16.4
Mean Temp. of Wettest Quarter	16	15.23	Mean Temp. of Driest Quarter	1479	19.54	Precip. of Wettest Month	209	17.23
Mean Temp. of Wettest Quarter	76	20.43	Mean Temp. of Driest Quarter	1542	25.84	Precip. of Wettest Month	362	16.98
Mean Temp. of Wettest Quarter	171	17.15	Mean Temp. of Driest Quarter	1565	30	Precip. of Wettest Month	387	18.5
Mean Temp. of Wettest Quarter	321	42.94	Mean Temp. of Driest Quarter	1626	15.16	Precip. of Wettest Month	711	18.21
Mean Temp. of Wettest Quarter	558	19.4	Mean Temp. of Driest Quarter	1654	15.61	Precip. of Wettest Month	779	15.98
Mean Temp. of Wettest Quarter	797	15.46	Mean Temp. of Driest Quarter	1742	17.55	Precip. of Wettest Month	860	16.15
Mean Temp. of Wettest Quarter	960	15.81	Mean Temp. of Driest Quarter	1874	23.53	Precip. of Wettest Month	1002	16.64
Mean Temp. of Wettest Quarter	975	17.81	Mean Temp. of Driest Quarter	1876	16.24	Precip. of Wettest Month	1279	17.81
Mean Temp. of Wettest Quarter	1145	15.41	Mean Temp. of Driest Quarter	1898	16.76	Precip. of Wettest Month	1344	20.43
Mean Temp. of Wettest Quarter	1320	20.55	Mean Temp. of Driest Quarter	1973	15.37	Precip. of Wettest Month	1646	18.79
Mean Temp. of Wettest Quarter	1342	21.86	Mean Temp. of Driest Quarter	2032	15.64	Precip. of Wettest Month	1693	17.63
Mean Temp. of Wettest Quarter	1612	21.54	Mean Temp. of Driest Quarter	2073	19.75	Precip. of Wettest Month	1882	15.35
Mean Temp. of Wettest Quarter	1779	18.09	Mean Temp. of Driest Quarter	2126	24.23	Precip. of Wettest Month	2102	20.53
Mean Temp. of Wettest Quarter	1998	15.79	Mean Temp. of Driest Quarter	2178	20.62	Precip. of Wettest Month	2195	16.4
Mean Temp. of Wettest Quarter	2037	15.79	Mean Temp. of Driest Quarter	2184	19.1	Precip. of Driest Month	2	17.85
Mean Temp. of Wettest Quarter	2085	16.07	Mean Temp. of Driest Quarter	2272	26	Precip. of Driest Month	33	25.53
Mean Temp. of Wettest Quarter	2244	19.61	Mean Temp. of Driest Quarter	2287	20.43	Precip. of Driest Month	39	22.16
Mean Temp. of Wettest Quarter	2327	17.79	Mean Temp. of Driest Quarter	2463	15.09	Precip. of Driest Month	77	18.96
Mean Temp. of Wettest Quarter	2427	17.34	Mean Temp. of Driest Quarter	2485	18.84	Precip. of Driest Month	97	18.94
Mean Temp. of Wettest Quarter	2452	24.54	Mean Temp. of Driest Quarter	2540	15.37	Precip. of Driest Month	146	15.3
Mean Temp. of Wettest Quarter	2495	17.29	Mean Temp. of Warmest Quarter	183	16.66	Precip. of Driest Month	214	18.01
Mean Temp. of Driest Quarter	16	19.17	Mean Temp. of Warmest Quarter	299	52.96	Precip. of Driest Month	230	27.26
Mean Temp. of Driest Quarter	85	19.31	Mean Temp. of Warmest Quarter	302	15.21	Precip. of Driest Month	450	20.79
Mean Temp. of Driest Quarter	196	15.48	Mean Temp. of Warmest Quarter	891	15.12	Precip. of Driest Month	490	18.55
Mean Temp. of Driest Quarter	200	17.81	Mean Temp. of Warmest Quarter	895	17.81	Precip. of Driest Month	507	15.41
Mean Temp. of Driest Quarter	203	16.84	Mean Temp. of Warmest Quarter	1515	16.09	Precip. of Driest Month	519	23.33
Mean Temp. of Driest Quarter	237	19.87	Mean Temp. of Warmest Quarter	1888	17.66	Precip. of Driest Month	527	16.84
Mean Temp. of Driest Quarter	249	17.44	Mean Temp. of Warmest Quarter	2178	15.48	Precip. of Driest Month	538	15.61
Mean Temp. of Driest Quarter	276	15.57	Mean Temp. of Warmest Quarter	2364	19.36	Precip. of Driest Month	545	20.32
Mean Temp. of Driest Quarter	291	16.34	Annual Precip.	209	22.72	Precip. of Driest Month	611	15.37
Mean Temp. of Driest Quarter	295	15.72	Annual Precip.	262	15.14	Precip. of Driest Month	621	20.72

Table 6 (cont.): List of SNPs with associations to environmental variables

Environmental Variable	SNP name	BF(dB)	Environmental Variable	SNP name	BF(dB)
Precip. of Driest Month	634	20.93	Precip. of Driest Month	2280	31.75
Precip. of Driest Month	665	25.55	Precip. of Driest Month	2282	16.38
Precip. of Driest Month	690	49.95	Precip. of Driest Month	2284	21.25
Precip. of Driest Month	692	15.96	Precip. of Driest Month	2308	16.82
Precip. of Driest Month	704	36.24	Precip. of Driest Month	2311	23.47
Precip. of Driest Month	777	16.24	Precip. of Driest Month	2361	16.21
Precip. of Driest Month	862	17.06	Precip. of Driest Month	2462	15.79
Precip. of Driest Month	957	15.46	Precip. of Driest Month	2485	17.29
Precip. of Driest Month	967	15.92	Precip. of Driest Month	2490	15.02
Precip. of Driest Month	1028	15.44	Precip. of Driest Month	2540	16.76
Precip. of Driest Month	1036	15.79			
Precip. of Driest Month	1110	21.36			
Precip. of Driest Month	1173	15.9			
Precip. of Driest Month	1242	32.48			
Precip. of Driest Month	1271	18.16			
Precip. of Driest Month	1292	17.17			
Precip. of Driest Month	1317	19.07			
Precip. of Driest Month	1382	19.9			
Precip. of Driest Month	1427	15.19			
Precip. of Driest Month	1428	19.82			
Precip. of Driest Month	1435	15.37			
Precip. of Driest Month	1480	20.34			
Precip. of Driest Month	1506	21.66			
Precip. of Driest Month	1589	17.21			
Precip. of Driest Month	1598	18.93			
Precip. of Driest Month	1609	23.68			
Precip. of Driest Month	1646	27.13			
Precip. of Driest Month	1683	16.8			
Precip. of Driest Month	1693	16.52			
Precip. of Driest Month	1733	21.92			
Precip. of Driest Month	1748	20.23			
Precip. of Driest Month	1749	18.68			
Precip. of Driest Month	1795	22.55			
Precip. of Driest Month	1815	16.58			
Precip. of Driest Month	1821	17.64			
Precip. of Driest Month	1856	25.4			
Precip. of Driest Month	1869	19.64			
Precip. of Driest Month	1885	21.9			
Precip. of Driest Month	1920	25.23			
Precip. of Driest Month	1949	15.05			
Precip. of Driest Month	2014	16.56			
Precip. of Driest Month	2042	18.12			
Precip. of Driest Month	2065	17.21			
Precip. of Driest Month	2072	15.35			
Precip. of Driest Month	2081	18.05			
Precip. of Driest Month	2083	16.72			
Precip. of Driest Month	2163	20.08			
Precip. of Driest Month	2170	16.24			
Precip. of Driest Month	2231	16.21			
Precip. of Driest Month	2239	17.96			
Precip. of Driest Month	2251	19.57			
Precip. of Driest Month	2271	19.49			

Table 7: “Risk of non-Adaptedness” values for each sampling site, projected for the three environmental variables with most associations to SNPs.

Model	RPC26				RPC85			
Covar	Precipitation of Driest Month	Mean Temperature of Driest Quarter	Temperature Seasonality	Average	Precipitation of Driest Month	Mean Temperature of Driest Quarter	Temperature Seasonality	Average
#SNPs	**79**	**51**	**33**		**79**	**51**	**33**	
Algeria	0.0085	0.0710	0.0318	*0.0371*	0.0148	0.1254	0.0841	*0.0748*
Catalonia	0.0054	0.1482	0.0432	*0.0656*	0.0905	0.3921	0.0944	*0.1923*
Corsica	0.0036	0.0415	0.0314	*0.0255*	0.0172	0.0871	0.0701	*0.0581*
Haza de Lino	0.0103	0.0545	0.0285	*0.0311*	0.0172	0.1092	0.0754	*0.0672*
Kenitra	0.0000	0.0416	0.0269	*0.0228*	0.0000	0.0834	0.0465	*0.0433*
Landes	0.0356	0.0568	0.0421	*0.0448*	0.1157	0.1084	0.0843	*0.1028*
Monchique	0.0028	0.0355	0.0292	*0.0225*	0.0042	0.0674	0.0497	*0.0405*
Puglia	0.0120	0.0415	0.0366	*0.0300*	0.0437	0.0998	0.0839	*0.0758*
Sardinia	0.0121	0.0501	0.0228	*0.0283*	0.0191	0.0955	0.0574	*0.0574*
Sicilia	0.0007	0.0421	0.0328	*0.0252*	0.0036	0.0761	0.0652	*0.0483*
Sintra	0.0049	0.0411	0.0236	*0.0232*	0.0068	0.0638	0.0405	*0.0370*
Taza	0.0000	0.0464	0.0408	*0.0291*	0.0007	0.0996	0.0795	*0.0599*
Toledo	0.0014	0.0520	0.0690	*0.0408*	0.0133	0.1064	0.1478	*0.0891*
Tuscany	0.0056	0.0383	0.0364	*0.0268*	0.0404	0.0848	0.0773	*0.0675*
Tunisia	0.0048	0.0428	0.0514	*0.0330*	0.0007	0.0963	0.0914	*0.0628*
Var	0.0057	0.0403	0.0388	*0.0282*	0.0257	0.1006	0.0924	*0.0729*
Min R^2	0.0000	0.0000	0.0007	*0.0002*	0.0000	0.0000	0.0007	*0.0004*
Max R^2	0.3410	0.2956	0.3316	*0.3227*	0.3410	0.2956	0.3316	*0.2693*
Average R^2	0.1597	0.1466	0.1545	*0.1536*	0.1597	0.1466	0.1545	*0.1570*

2 Data

Data 1: Parameter files used in the ipyrad analyses. Both the main parameter file and the "popfile" are present in the link.

https://gist.github.com/StuntsPT/399f2957b3af9450c26089c05ee5c037

Data 2: GNU Makefile containing the entire analyses process from the data in this work.

https://gist.github.com/StuntsPT/c3c1f4c1f77f7151f00d168b5a01dced

3 Figures

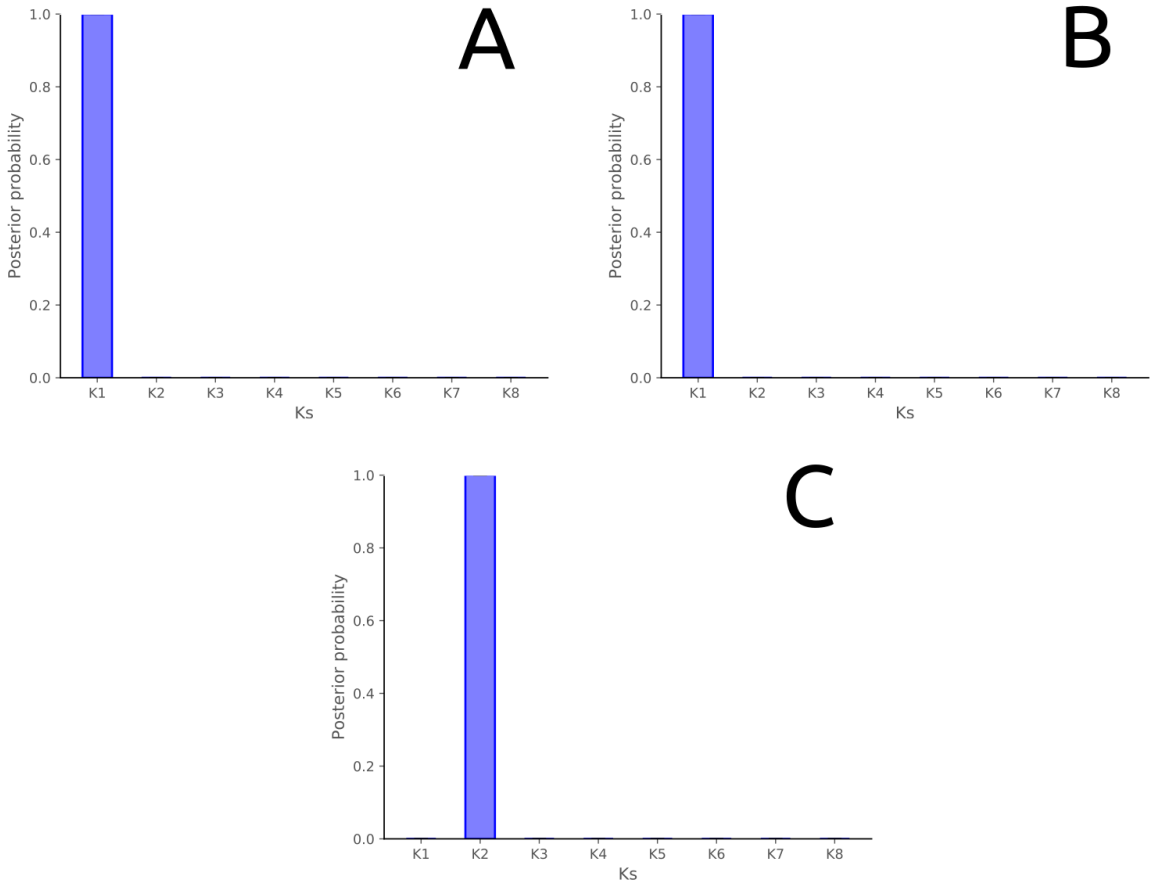

Figure 1: *MavericK* posterior distribution plots. “A” is for the dataset with all loci, “B” is for the dataset with only “neutral” loci, and “C” is for the dataset with only “non-neutral” loci.

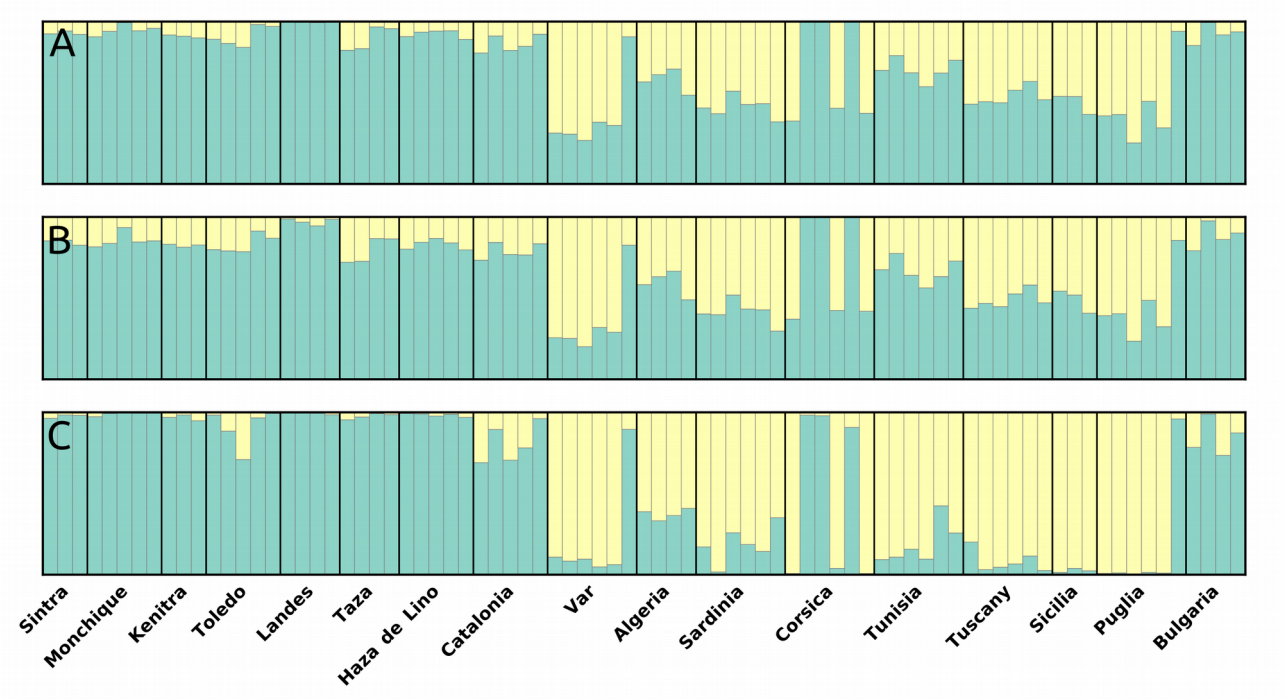

Figure 2: STRUCTURE clustering plots for K=2. Sampling sites are presented from West to East. “A” is the Q-value plot for the dataset with all loci, “B” is for the dataset with only “neutral” loci, and “C” if for the dataset with only “non-neutral” loci.

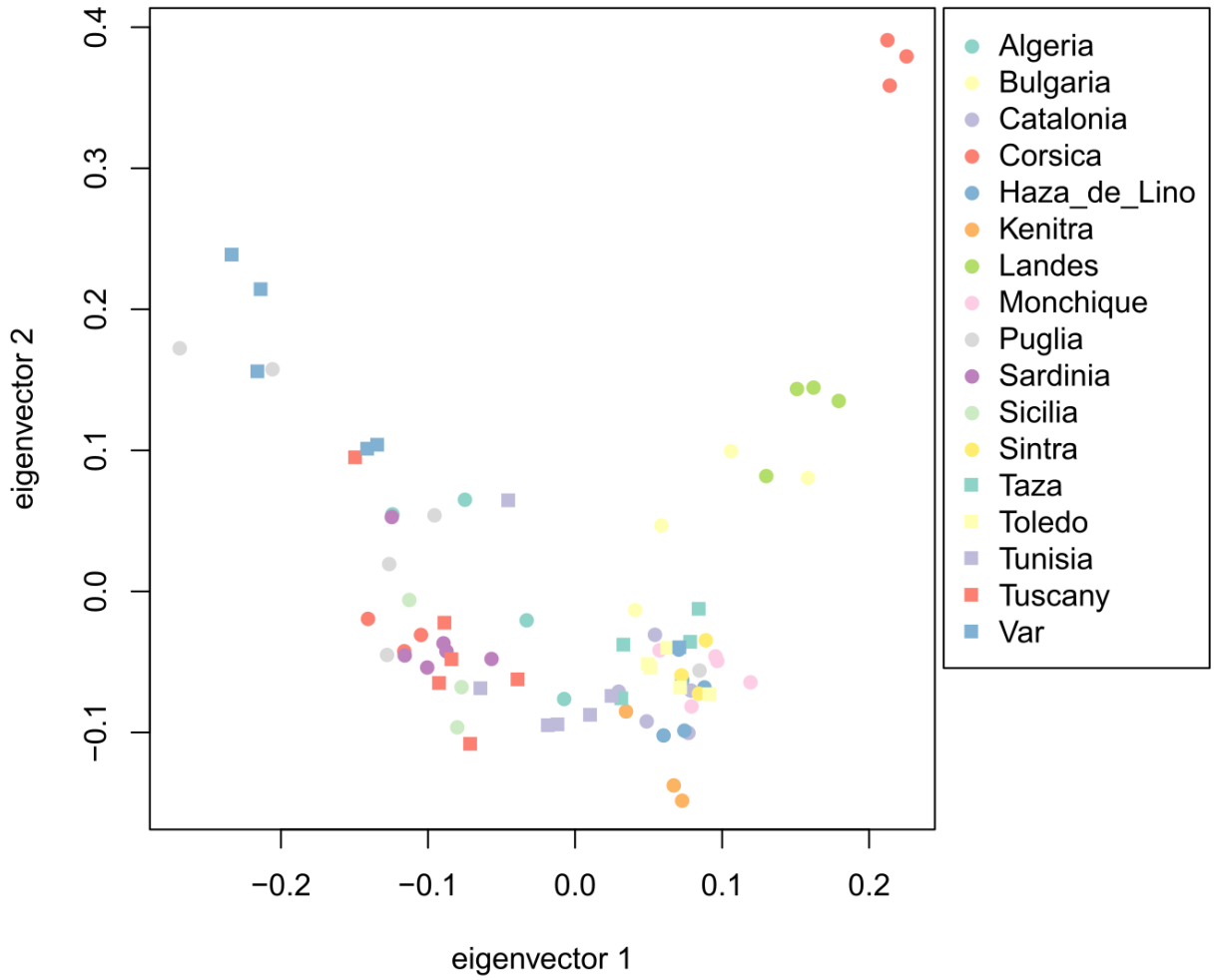

Figure 3: PCA plot for the dataset with all loci. Eigenvector 1 explains 4.31% of the variation and eigenvector 2 explains 2.41% of the variation.