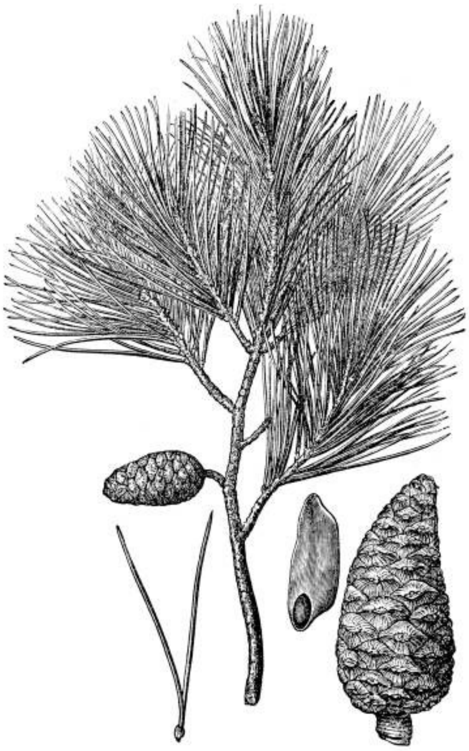
Demography, Selection and Evolution in Conifers



Rose Ruiz Daniels

UNIVERSIDAD AUTÓNOMA DE MADRID

CENTRO DE INVESTIGACIÓN FORESTAL (INIA-CIFOR)

Presented by Rose Ruiz Daniels for the completion of a
PhD in the Autonomous University of Madrid

Supervised by Dr. Delphine Grivet

May 2017

Cover image: Aleppo pine (*Pinus halepensis*) / Vintage illustration from Meyers Konverations-Lexikon 1897.

This thesis is for the shiniest starts in my universe Rowan and the new baby, and my partner Richard that has been there before I became involved in population genetics and is still here, after one child soon to be two, and one thesis and hopefully soon to be two as well.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

Rose Ruiz Daniels

2017

Acknowledgements

I really want to thank my supervisor Delphine Grivet for allowing me to continue exploring evolutionary biology, and an endless supply of patience needed to supervise me for 4 years. I want to also thank Mark Beaumont and Phil Donoghue for the welcome and supervision in Bristol, as well as very insightful talks. A special thanks to my tutor Joaquina de la Torre in the Autonoma. To those people in Madrid and Bristol that during these four years have been part of this journey, you all know who you are, and how great you all are. And last but by all means not least, my parents for giving me and my offspring nurture and love as well as much needed emotional support.

I have left both the names of the co-authors of these studies and the acknowledgements in each chapter, this thesis although a lonely endeavour was also the work of many

But, Mousie, thou art no thy lane,

In proving foresight may be vain;

The best-laid schemes o' mice an' men

Gang aft agley.

From **Tae a Moose, on Turning Her Up in Her Nest with the Plough,**

- Robert Burns (1785)

Overview

This thesis, divided into four chapters, is dedicated to inferring different aspects of molecular adaptation in conifers.

In Chapter one we introduce the work and place it in the context of the current understanding of conifer genetics.

In Chapter two we analyse the demography and selection of the Aleppo pine (*Pinus halepensis* Mill) at the full-scale distribution of the species. We delve into the effect of demographic history, gene surfing due to long range colonisation, and how this affects selection inference.

In Chapter three we consider selection at the local scale, by analysing population pairs contrasted in their climatic conditions from the western part of the range of the Aleppo pine. We then compare different methodologies of inferring selection. We finally apply what has been learnt to infer selection in other conifer species using genomic data.

In Chapter four we investigate microRNAs in conifers as a potential future direction in inferring molecular adaptation, with the aim of developing a way of establishing a gene ontology for these molecules, which could render the transfer of knowledge of these from model species (e.g. *Arabidopsis thaliana* (L.) Heynh) to non-model species such as conifers.

Resumen

La presente tesis doctoral la componen un total de 3 capítulos dedicados a inferir diferentes aspectos de la adaptación molecular en las coníferas.

En el primer capítulo introducimos el trabajo y lo contextualizamos en el actual entendimiento de la genética molecular de las coníferas.

En el segundo capítulo nos centramos en el estudio de la demografía y los procesos de selección natural del pino Carrasco (*Pinus halepensis* Mill) a lo largo de su distribución natural. Ahondamos en el efecto de la historia demográfica y del “gene surfing” (fenómeno estocástico por el que genes de baja frecuencia pueden alcanzar altas frecuencias a causa de colonizaciones a larga distancia) y de cómo esto dificulta la detección de las señales de selección.

En el tercer capítulo estudiamos los procesos de selección a escala local, para lo cual utilizamos pares de poblaciones procedentes de zonas cuyas características ambientales contrastan entre sí, dentro del hábitat del pino Carrasco. Para desarrollar este capítulo comparamos diferentes metodologías estadísticas para inferir SNPs con frecuencias atípicas potencialmente bajo selección. Usamos lo aprendido hasta ahora para inferir selección en otras especies de coníferas, haciendo uso de datos genómicos.

En el cuarto capítulo investigamos los MicroRNAs (un tipo de Pequeño ARN no codificante) en coníferas como una potencial y clave dirección para entender la adaptación molecular. Exploramos cómo establecer una ontología de genes para estas moléculas y así facilitar la transferencia de conocimientos de especies modelo, como es *Arabidopsis thaliana* (L.) Heynh, a otras especies a día de hoy menos estudiadas como son el grupo de coníferas.

Keywords: Molecular Ecology, Selection, Adaptation, Genomics, Conifers.

Contents

List of Figures

List of Tables

Preamble

Two hundred years ago a book was published that changed the way we thought about the natural world. Although natural selection had previously been alluded to, it was not until Darwin and Wallace, and the publication of The Origin of the Species, that the necessary leap was taken to put these ideas concisely and coherently in what became one of the most revolutionary theories in human history;

"As many more individuals of each species are born than can possibly survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be naturally selected. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form" Charles Darwin original abstract from *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (1859)*

Thus, it was conveyed that not all organisms survive and that some will have an advantage over others: those that do will multiply and spread their genes more effectively than that those that do not. Over time this leads to more advantageous traits taking precedence over the not so successful traits. To cite the most well-known example of this one must think of the peppered month during the industrial revolution in England, where the darker individuals became more and more common relative to the lighter individuals, as the bark of the trees it inhabited and camouflaged in turned darker with the fumes of industry (Hof *et al.* 2011)

Natural selection together with the theories of Gregor Mendel demonstrated in pea plants that traits can indeed be passed down from generation to generation and that alleles come in pairs, and can be both dominant and recessive; this elucidated a mechanism in which traits (genes) can pass from generation to generation. This was the birth of Mendelian genetics that formed the backbone of what today we understand as population genetics (Mendel, 1886). The main concepts are the same when dealing with different organisms, as well as data sets from microsatellites to full genomes; population genetics views evolution as changes in the genetic makeup of populations, with the idea that if we can understand the combined action of the forces that change gene frequencies in populations, over many generations we might understand long term trends in evolution. How true this is and how much does microevolution translate into macroevolution is a field of study in itself (Hansen & Martins 1996; Reznick & Ricklefs 2009). So, without losing sight of population genetics, its modern day version was mainly developed by Ronald Fisher, J. B. S. Haldane, and Sewall Wright who established mathematical models with the purpose of understanding how allele frequencies changed over time in populations. This can be very basically summarised as: alleles are influenced by two main forces those deterministic- selection, mutation and migration- and those that are stochastic such as genetic drift. Although this does simplify things a bit, untangling these effects is anything but straightforward, these forces act, in unison, contradictory and sometimes shield the effects from each other. The evidence of natural selection is everywhere, yet it is surprisingly hard to observe, given the time frame it acts on, and the fact that current allele frequencies have been influenced over time by different forces (for a detailed summary of population genetics see Gillespie 2004). Modern day methods are still aiming to untangle selection forces in natural populations parting from the basis of looking for differences in allele frequencies among different populations, or F_{ST}, one way of doing this is by looking at F_{ST} outliers relative to an appropriate neutral population-genetics model indicating selection (Beaumont & Balding 2004). The key is to consider a neutral population model that suitably corrects for forces other than selection that can also affect allelic frequencies between populations (F_{ST}).

Chapter 1 **Introduction**

1.1 Introduction to the systems of study

1.1.1 Conifers

Conifers are the most diverse division of extant gymnosperms, with 627 living species in eight different families, and evolved around 300 million years ago. Modern conifer families are identifiable in the fossil record as early as the Jurassic and Triassic (Rothwell *et al.* 2012), dating them to well before the evolution of the angiosperms and consequently are important as an out-group for comparative evolutionary studies with the angiosperms (e.g. Biffin *et al.* 2012). Conifers have significant ecological importance worldwide, dominating northern hemisphere taiga forest, with many of these northern hemisphere species prospering as invasive species in their non-native southern hemisphere (Allen 2007). The northern coniferous taiga covers vast areas of North America from the Pacific to the Atlantic, and range across northern Europe, Scandinavia, Russia and across Asia through Siberia and Mongolia to northern China and northern Japan, a testament to their ability to survive harsh climatic conditions (www.wwf.com). On the other extreme are the Mediterranean conifers of the Northern Hemisphere that are present in Western America (California), and span vast forested areas in the Mediterranean Basin, most of the time colonists of lands that are characterised by extreme climatic conditions such as high temperatures and low rainfall. They are also economically important in many parts of the world, providing much of the world's prime timber resources (Page 1990).

Owing to their ecological and economic importance, their ability to survive in extreme climatic conditions, and their ubiquity in the northern hemisphere forests, conifers are interesting and important subjects for studies in evolutionary genetics. One thing that has hampered the study of conifers in a molecular evolutionary context is their notoriously large genome size and genome complexity, which has been attributed to polyploidy, highly repetitive tandem elements, transposable elements, size of gene families and abundance of pseudogenes, as well as intron sizes (Ahuja & Neale 2005). Recently, drafts genomes have been made available for three conifers species: Norway spruce (*Picea abies*, Nystedt *et al.,* 2013), white spruce (*Picea glauca*, Birol *et al.* 2013), and Loblolly pine (*Pinus taeda*, Wegrzyn *et al.* 2014). These genomes are approximately 20-30 Giga base-pairs in size, as well as containing multiple repeats further adding to the complexity of sequencing and assembly of these genomes. Not only has this had an impact on the time it has taken to get a full genome, but also on the quality (**Figure 1.1**), making the production of molecular data time consuming and costly.

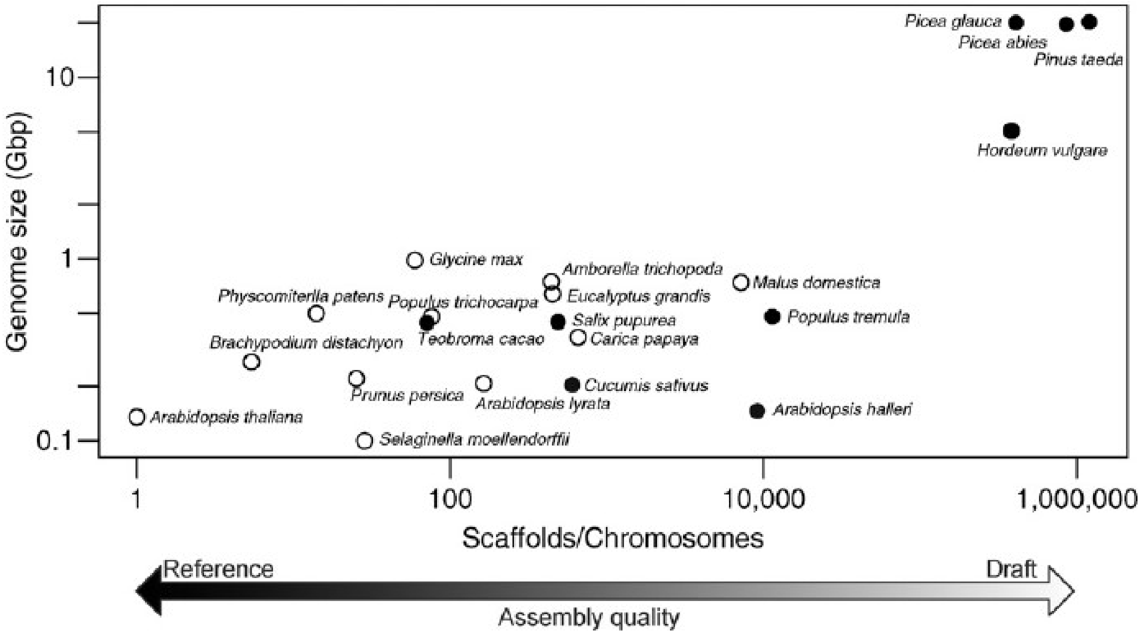


Figure 1.1: Size and assembly of conifer genomes compared with other plant genomes. Genome size is plotted against the number of scaffolds divided by the haploid chromosome number for a range of plant species. From De La Torre *et al. (*2014)

1.1.2 *Pinus halepensis*

Pinus halepensis Mill, known as the Aleppo pine, occupies a vast and fragmented circum-Mediterranean distribution of 3.4 million hectares (**Figure 1.2**). It is wind pollinated with cones of both sexes that flower in the spring; its cones reach maturity two years after pollination. One notable feature of the cones of *P. halepensis* is that mature cones can either open on maturity, or remain closed until exposed to high temperatures these being called serotinous (**Figure 1.3**). This feature is likely to be an adaptation to the constant present natural fires in the Mediterranean to protect the seeds from fire and then to rapidly repopulate newly empty and charred earth (Saracino & Leone 2001; Budde *et al.* 2014). The Aleppo pine lives in almost all bioclimates in the Mediterranean Basin as well as in a range of different substrates. For this reason, *P. halepensis* is regarded as having a high level of adaptability over its substantial range.

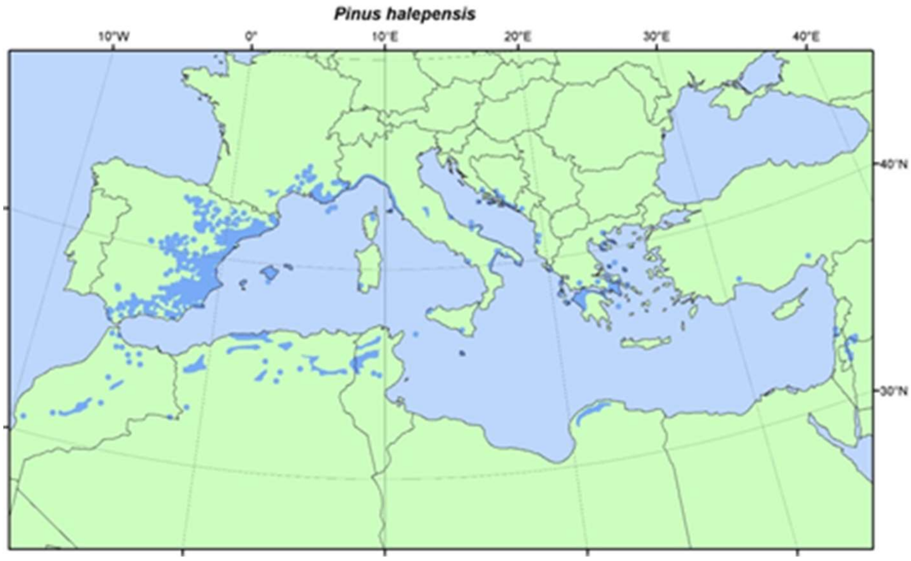


Figure 1.2: The natural range of the Aleppo pine.

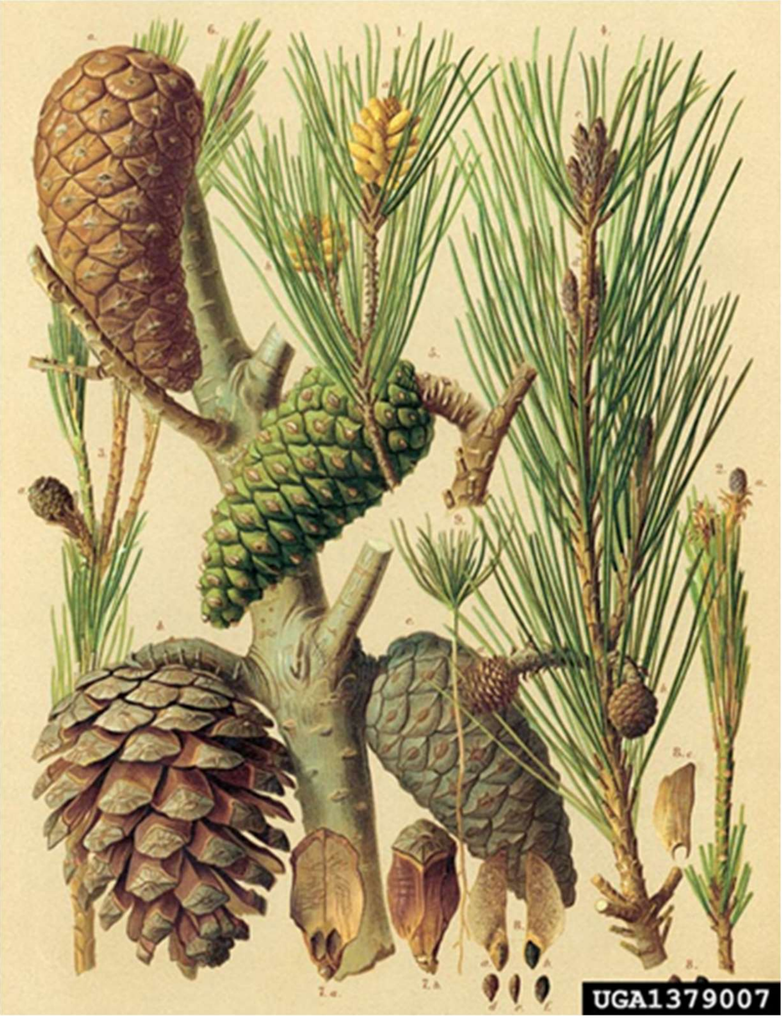


Figure 1.3: Cones of the Aleppo pine at different stages of maturity including a serotinous (bottom right) and non-serotinous (bottom left) mature cone. (Image source https://in.pinterest.com/pin/233765036888343286)

The size of the genome of *P. halepensis* is not known, but is estimated to be typical for conifers at approximately 20-30 Giga base pairs. Prior to this thesis, the most recent work on the genetics of *P. halepensis* used six chloroplast markers (Grivet *et al.* 2009) and eleven candidate genes (Grivet *et al.* 2011). Since then, a transcriptome has also been sequenced (Pinosio 2014), providing more molecular resources. This thesis extends these resources by utilising an unprecedented number of markers: an in-depth exploration of 8 nuclear SSRs and 293 SNPs sampled across the geographic range of the species, and further SNPs sampled at a more local scale. In very recent times, an additional genomic dataset containing 7686 SNPs has become available and is analysed in this thesis.

1.2 Evolution and adaptation in *P. halepensis*

1.2.1 Genetics and demography of *P. halepensis*

Previous studies (Schiller *et al.* 1986; Bucci *et al.* 1998; Grivet *et al.* 2009) have found that *P. halepensis* has low diversity in neural molecular markers. This diversity has been found along a longitudinal gradient, with older populations showing higher levels of genetic diversity than the (likely) more recent and less diverse western populations (Morgante *et al.* 1998; Bucci *et al.* 1998; Grivet *et al.* 2009). This disparity in diversity strongly suggests that the Aleppo pine has undergone a range expansion: Grivet *et al.* (2009) used chloroplast markers of 6 populations and it was determined that the population in Greece is the oldest and more historically stable population in the East of the Mediterranean Basin, strongly suggesting a range expansion from older eastern populations to newer western populations. It is inferred from this that *P. halepensis* expanded westwards around the Mediterranean Basin from a southern Balkan refugia (Grivet *et al*. 2009). This strongly indicates that the Aleppo pine underwent a long-range colonisation from East to West, most likely accompanied by recurrent contractions and expansions, caused by things like forest fires and glaciations. This recent colonisation event combined with the current wide ecological range and scattered distribution across the entire Mediterranean Basin, makes *P. halepensis* especially suited to studies of the impacts of demography and genetic process, as well as the interaction of these on the adaptive potential of tree populations.

Study into the ability of tree species to adapt to a range of climatic conditions is increasingly urgent, due to climatic predictions which indicate that the Mediterranean Basin is likely to shift to the arid domain in the future (Petit *et al.* 2005). Therefore, study into how *P. halepensis* has adapted to its environment in recent times, may give us insight into how it may adapt to and cope with rapidly changing conditions in the future. Its widespread distribution suggests that its populations may well be differentially adapted to local climatic conditions and may show different capabilities to survive water deficient stress. This is reinforced by various studies that show differences in drought tolerance amongst different provenances by looking at both ecophysiological data and gene expression (Atzmon *et al.* 2004; Sathyan *et al.* 2005; Voltas *et al.* 2008). The fact of its success as an invasive species in southern hemispheres (Lavi *et al.* 2005) also suggests that it is highly adaptable. In this thesis, we focus mainly on molecular adaptation, touching briefly on epigenetic adaptation.

1.2.2 Selection and adaptation in general with a special focus on *P. halepensis*

All loci in a genome will be affected by demographic process, whereas only a subset will be influenced by selection (Begun *et al.* 2007). As genetic adaptation takes place, rare but advantageous alleles will increase in frequency, leading to drastic differences in allele frequencies between geographic regions. These differences can be interpreted as signals of positive selection, and indicate adaption to different environmental pressures. However, during range expansions, such as the one undergone by the Aleppo pine, genetic drift between populations increases with geographic distances and spatial expansions into new territories, which can result in the spread of otherwise rare alleles over wide areas where they can reach high frequencies. This stochastic effect known as "gene surfing" is due to strong genetic drift occurring to populations located as the edge of an expansion (Edmonds *et al.* 2004; Klopfstein *et al.* 2006). Such genetic changes leading to drastic differences in allele frequencies between geographic regions can be interpreted as signals of positive selection and thus lead to false positives when inferring selection. It has been shown both experimentally and theoretically that range expansions can generate effects most often attributed to adaptive selection such as allele frequency clines or the observation of drastic gene frequency differences between source and target of an expansion (Excoffier & Ray 2008). In some instances it has been suggested that a big proportion of these allelic differences normally attributed to selection between populations, is in fact due mainly to the actions of gene surfing and not selection (Hofer *et al.* 2009).

During range expansion, genomes are subjected to selection, demography and the increased effect of random genetic drift, which is a challenge when trying to understand how these effects have shaped current allelic frequencies. Only by integrating all knowledge about the demographic history, the population structure and the ecology of the species can we reduce false positives in selective inference. Optimizing sampling design has also been shown to help the inference of selection (Hoban *et al.* 2016). Previous works have found genes under putative selection in the Aleppo pine (Grivet *et al.* 2011), but disentangling the effect of demography from those of selection proved difficult. Since then, new techniques have been developed that correct for population structure when inferring genetic outliers (Gunther & Coop 2013; Gautier 2015; Luu *et al.* 2017). Furthermore, the effect of gene surfing in detecting putative SNPs under selection in the specific case of a forest tree species, which has undergone a range expansion and has long generation times, is a problem. Although methods still do not account for this, it is possible to use simulated data mimicking populations that have undergone a range expansion, and to compare it to real data representing populations that have undergone rnage expansion and potentially selection, and see how much of an effect gene surfing is having on outlier selection methods. Some methods even allow finding outliers associated with different environmental variables (Gunther & Coop 2013; Gautier 2015), giving a hint on the nature of the drivers responsible for selection and if they are the same at the local scale and at the full species distribution.

In this work, we explore selection at the full scale distribution of the Aleppo pine as well as looking at it at a local scale, to explore how selection acts at different spatial scales. Local adaptation is when natural selection locally leads to allele frequency shifts that are uniquely adapted to those local conditions, either on already present allelic frequencies (standing genetic variation) or in novel mutations (Savolainen *et al.* 2013; Hoban *et al.* 2016). As genomic data sets and new methodologies become available (see Lotterhos & Whitlock 2015 for review) it will become easier to untangle these questions. It is very likely that local adaptation is somewhat tempered by the plasticity of the species, and to the extent that they can adapt to the local conditions given their genetic heritage (Gould & Lewontin 1979). These new methodologies paired with experimental design can help understand the mechanisms of local adaptation.

1.3 Beyond SNPs and population genetics. Adaptation through other mechanisms

Evidence of non-genetic inheritance is growing both theoretically and empirically across kingdoms, including plants, this is an interesting avenue for understanding how organisms cope/adapt to changing environmental pressures (Salinas *et al.* 2013). Evidence of epigenetic adaptation is appearing in conifers (Bräutigam *et al.* 2013), the most widely cited mechanisms for this process is DNA methylation (Sáez-Laguna *et al.* 2014) although the methods of epigenetic processes are varied and we will not have time to discuss then in detail in this thesis. The mechanisms and roles of epigenetic processes in rapid adaptation in plants are still largely unknown, and this field is both new and exciting in presenting a wide variety of possibilities for future research. To grasp the time span in which these epigenetic changes can occur one only has to look at one study that shows that *Pinus sylvestris* grown in the sites of the Chernobyl nuclear disaster (1986) showed hypermethylation in their DNA, a sign of a potential adaptation to high dosage of radiation, of course further research is needed to establish a clear cause and effect of these results, (Kovalchuk *et al.* 2003), but it gives an idea on how fast these changes can take place, compared to genetic adaptations that can take many hundreds of generations. Transcriptional changes seem to play a role in epigenetic adaptation of forest tress (Bräutigam *et al.* 2013) Studies are starting to emerge that shed light on how these Transcriptional changes could be mediated; In Norway spruce, conserved miRNAs as well as a large proportion of novel non-conserved miRNAs involved in temperature-dependent epigenetic "memory" have been found. The expression of these being significantly different in progenies showing distinct epigenetic difference in bud set, but not in the progeny from a non-responding family without differences in bud set, These differential expression of miRNAs suggests their putative participation in epigenetic regulation in conifers. (Yakovlev *et al.* 2010).

1.4 microRNAs

1.4.1 Introduction to miRNAs

MicroRNAs (miRNA) are a type of non-coding regulatory small RNA that occur in the genomes of several eukaryotic lineages (including plants) and operate at the post-transcriptional level by mediating gene silencing. As conifer genomics advances and more genomes become sequenced, the discovery and annotation of miRNAs in conifer species becomes easier.

1.4.2 miRNAs and adaption in conifers

There is widespread interest in these molecules given that they are potentially used to counter environmental stresses in plants by altering gene expression (Rajwanshi *et al.* 2014), and have been found to have regulatory roles under the following stresses: disease resistance, abiotic, and biotic stress response, and reproduction (Kruszka *et al.* 2012; May *et al.* 2013; Shriram *et al.* 2016), making them good molecular candidates to explore local adaptation. As pine genomes have only started to be sequenced, it has been hard to discover miRNAs in conifers beyond those that have already been described in the literature, mainly taken from comparative studies with other plant species. This has led to some studies prematurely declaring that there are not as present in the gymnosperms compared to the angiosperms (De La Torre *et al.* 2014). However this lack of information/evidence does not preclude the important role of miRNAs in conifers, as illustrated by differential expression of miRNAS observed in conifers, in different empirical studies (Yakovlev *et al.* 2010; Quinn *et al.* 2014; Qiu *et al.* 2016).

1.4.3 Challenges

The study of miRNAs is further stymied by another issue: although the development of next generation sequencing has led to an expansion of the miRNA repository miRBase (Kozomara & Griffiths-Jones 2014), there exists widespread inconsistencies and errors in annotation resulting in large numbers of incorrectly identified miRNAs due to a failure to adhere to existing annotation criteria. This leads to several negative consequences such as obstructing the functional comparative analyses of miRNAs, hampering evolutionary studies and exaggerating the impact of "species specific" miRNAs (Taylor *et al.* 2014, 2017).Within the study of micro RNAs it has long been recognised that a common language for annotation is required to allow for the identification of functional conservation (Ashburner *et al.* 2000) and this has been attempted with some success in the animal Kingdom (Fromm *et al.* 2015). The most powerful ontology is one that reflects the evolutionary history of the gene family, since a gene name that reveals a common origin for a gene present in several lineages allows for the identification of functional conservation or divergence in those lineages. Currently, miRNAs are assigned to gene families that reflect a common evolutionary origin from a single ancestor gene, along with a lowercase alphabetic suffix to signify paralogue identity within the miRNA family. Meaning that this component of the gene name does not provide an adequate scheme of homology for genes across species i.e. homologous genes in two different species may have different names. miRNA function on the level of the gene and not family, since small sequence differences in the mature sequence of paralogues within a family can lead to different genes being targeted, what this means is that the miRNAs currently classified as belonging in a family will not necessarily have similar functions. Many studies have demonstrated differential expression between paralogues within families, and that individual paralogues have specific roles distinct to other members of the same family (e.g. Thatcher *et al.* 2015). Due to the lack of a scheme of homology that identifies homologues across species, this information cannot be transferred to other species without an ontology that allows for the identification of homologues.

1.5 Aim of this Thesis

The aim of this thesis is to explore and elucidate the genetic basis of adaptation of *P. halepensis* to environmental pressures. The approach adopted is to utilise a wide range of data sampling techniques, both in terms of molecular markers (nuclear microsatellites, hundreds of SNPs up to several thousand SNPs, and miRNA that can potentially identify gene regulatory changes) and populations (full distribution range versus local scale) to identify genetic adaptations to environment, These datasets are combined with the latest development in statistical frameworks to identify selective pressures, and the utility of these methods are rigorously assessed through the use of simulated data as well as comparative empirical data.

1.5.1 Chapter 2

This thesis starts by focusing on a species of conifer, the Aleppo pine (*Pinus halepensis*), which occupies a vast range in the Mediterranean Basin of 3.3 million hectares. We use it as a model system to explore classic population genetics given the long-range colonisation history of this conifer across the Mediterranean Basin. In Europe, fossil records and genetic data indicate that Aleppo pine demographic history would be characterized by an initial ancient colonization (last interglacial before the last glaciation), followed by later re-expansions (after the last glaciation) from eastern and western refugia towards the western Mediterranean Basin. This left the Aleppo pine with a complex genetic pattern, characterized by two main genetic clusters, and an eastern/southern cluster which is both older and more genetically diverse than the western cluster. We use this information to try to tease apart the effect of natural selection, demography and the increase of genetic drift at the edges of range expansions (a process known as gene surfing) and that produces a similar molecular footprint as selection on current allele frequencies (Excoffier & Ray 2008). We make use of molecular markers (SSRs and SNPs) to identify selection footprints driven by environmental pressures at the molecular level while correcting for population structure. We then make use of simulated populations to test for false positives that arise due to gene surfing as a consequence of range expansion, and finally we establish a framework in order to test these effects in any species that has undergone a range expansion. Here we not only unveil a more complex demographic history of Aleppo pine, but find outlier SNPs under putative selection linked to temperature (1 SNP) and precipitation (6 SNPs), as well as explore how problematic is gene surfing when inferring outlier loci in species that have undergone past range expansions, providing one of the few empirical studies to do so.

1.5.2 Chapter 3 (part I)

Does natural selection act differently on different scales? Will selection drivers differ in full scale studies compared to local scale? The study of microgeographic adaptation has been generally unexplored in forest trees and subsequently in conifers (Scotti *et al.* 2016), most likely due to the difficulty in procuring molecular markers in part due to their large genomes, as well as their relative long generation times, making the studies of adaptation in model species with short generations more appealing. By using the Aleppo pine system at the local scale, we can contrast selection in two different spatial scales. To do so we explore different methods to detect outlier loci potentially responsible for local adaptation. We optimised our sampling design using sampling in pairs in order to have the highest power to detect selection, by maximizing the difference in adaptive environment while minimising the differences in evolutionary history – a technique design which has been shown to work well in simulated data sets (Lotterhos & Whitlock 2015). To infer SNPs that are potentially involved in local adaptation we run the most up-to-date outlier detection methods that correct for population structure: two Bayesian methods, Bayenv2 (Gunther & Coop 2013) and Baypass (Gautier 2015), and one method based on PCA (latent factor) implemented in PCAdapt (Luu *et al.* 2017). We review these methods and find interesting loci under putative selection at the local scale.

1.5.3 Chapter 3 (part II)

In the second part of chapter 3, we then use genomic data obtained by exome capture and use the concepts and bioinformatics pipelines used in the first part of this chapter to analyse bigger SNP datasets from three different conifer species sampled at the local scale: Silver fir (*Abies alba*), Atlas cedar (*Cedrus atlantica*), and Aleppo pine (*Pinus halepensis*). We found outlier SNPs for all three species associated with different environmental variables of interest.

When this thesis started no conifer genomes had been sequenced, and since then three came out simultaneously *Pinus taeda*, *Picea abies* and *Picea glauca*. This new era of full genome sequences opens new avenues to understand conifer evolution in a wider context. I got the chance to be involved in a project that was led by the University of Bristol (UK) where I did several doctoral stays, which allowed me to study one aspect of adaptation at the full genome level.

1.5.4 Chapter 4

In chapter four we delve into whole genome analysis to understand better the evolutionary history of plant microRNAs. We attempt to set up a classification framework for these to improve future works into how these affect developmental and physiological processes in plants. In this thesis, we try to infer miRNA homology between species using two different approaches: the first using phylogenetic analysis. Once the miRNAs were established as real in the model organism *Arabidopsis lyrata*, they were grouped into families of paralogues. In order to root the paralogue trees the most evolutionarily distant genotyped species that contained a miRNA paralogue of that family was used, so it represented the lineage of the earliest known occurrence of the paralogue of interest. This included using miRNAs from the following species: *Selaginella moellendorffi*, *Pinus taeda* and *Picea abies*. This approach was found to be not robust enough to establish an ontology although some of the results were interesting.

The second more conservative approach inferencing the miRNA homology between species was based on BLAST that was used to search for all the validated precursor sequence in each of the genomes of species Brassicaceae. Four taxa within the Brassicaceae were chosen to perform a homology search to validate *A. thaliana* miRNAs: *Arabidopsis lyrata, Capsella grandiflora,* and C*apsella rubella*. So far, this method has yielded good results within the Brassicaceae but it is unclear how far it can work and if it can extend to a kingdom wide level, but initial results are promising and ongoing.

1.6 References

Ahuja MR, Neale DB (2005) Evolution of genome size in conifers. *Silvae Genetica*, **54**, 126–137.

Allen TFH (2007) A Natural history of equations. *BioScience*, **57**, 286.

Ashburner M, Ball CA, Blake JA *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Atzmon N, Moshe Y, Schiller G (2004) Ecophysiological response to severe drought in *Pinus halepensis* Mill. trees of two provenances. *Plant Ecology (formerly Vegetatio)*, **171**, 15–22.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular ecology*, **13**, 969–80.

Begun DJ, Holloway AK, Stevens K *et al.* (2007) Population Genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology*, **5**, e310.

Biffin E, Brodribb TJ, Hill RS, Thomas P, Lowe AJ (2012) Leaf evolution in Southern Hemisphere conifers tracks the angiosperm ecological radiation. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 341–348.

Birol I, Raymond A, Jackman SD *et al.* (2013) Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics*, **29**, 1492–1497.

Bräutigam K, Vining KJ, Lafon-Placette C *et al.* (2013) Epigenetic regulation of adaptive responses of forest tree species to the environment. *Ecology and evolution*, **3**, 399–415.

Bucci G, Anzidei M, Madaghiele A, Vendramin GG (1998) Detection of haplotypic variation and natural hybridization in halepensis -complex pine species using chloroplast simple sequence repeat (SSR) markers. *Molecular Ecology*, **7**, 1633–1643.

Budde KB, Heuertz M, Hernández-Serrano A *et al.* (2014) In situ genetic association for serotiny, a fire-related trait, in Mediterranean maritime pine (*Pinus pinaster*). *New Phytologist*, **201**, 230–241.

Edmonds CA, Lillie AS, Cavalli-Sforza LL (2004) Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences*, **101**, 975–979.

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, **23**, 347–351.

Fromm B, Billipp T, Peck LE *et al.* (2015) A Uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu Rev Genet*, **49**, 213–242.

Gautier M (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, **201**, 1555–1579.

Gillsepie JH (2010) *Population genetics: a concise guide.* Johns Hopkins University Press, Baltimore, Maryland, USA.

Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London Series B*, **205**, 581–598.

Grivet D, Sebastiani F, Alia R *et al.* (2011) Molecular footprints of local adaptation in two Mediterranean conifers. *Molecular Biology and Evolution*, **28**, 101–116.

Grivet D, Sebastiani F, González-Martínez SC, Vendramin GG (2009) Patterns of polymorphism resulting from long-range colonization in the Mediterranean conifer Aleppo pine. *New Phytologist*, **184**, 1016–1028.

Gunther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.

Hansen TF, Martins EP (1996) Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution*, **50**, 1404.

Hoban S, Kelley JL, Lotterhos KE *et al.* (2016) Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist*, **188**, 379–397.

Hofer T, Ray N, Wegmann D, Excoffier L (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Annals of Human Genetics*, **73**, 95–108.

Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, **23**, 482–490.

Kovalchuk O, Burke P, Arkhipov A *et al.* (2003) Genome hypermethylation in *Pinus silvestris* of Chernobyl--a mechanism for radiation adaptation? *Mutation research*, **529**, 13–20.

Kozomara A, Griffiths-Jones S (2014) MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, **42**, D68-73.

Kruszka K, Pieczynski M, Windels D *et al.* (2012) Role of microRNAs and other sRNAs of plants in their changing environments. *Journal of Plant Physiology*, **169**, 1664–1672.

De La Torre AR, Birol I, Bousquet J *et al.* (2014) Insights into conifer giga-genomes. *Plant physiology*, **166**, 1724–32.

Lavi A, Perevolotsky A, Kigel J, Noy-Meir I (2005) Invasion of *Pinus halepensis* from plantations into adjacent natural habitats. *Applied Vegetation Science*, **8**, 85.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

Luu K, Bazin E, Blum MGB (2017) pcadapt : an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, **17**, 67–77.

May P, Liao W, Wu YJ *et al.* (2013) The effects of carbon dioxide and temperature on microRNA expression in Arabidopsis development. *Nature Communications*, **4**.

Mendel G (1865) Experiments in Plant Hybridization. *Journal of the Royal Horticultural Society*, **IV**, 3–47.

Morgante M, Felice N, Vendramin GG (1998) Analysis of hypervariable chloroplast microsatellites in *Pinus halepensis* reveals a dramatic genetic bottleneck. In: *Molecular Tools for Screening Biodiversity*, pp. 407–412.

Nystedt B, Street NR, Wetterbom A *et al.* (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**, 579–584.

Page CN (1990) Economic Importance and Conifer Conservation. In: *Pteridophytes and Gymnosperms*, pp. 293–294. Springer Berlin Heidelberg, Berlin, Heidelberg.

Petit RJ, Hampe A, Cheddadi R (2005) Climate changes and tree phylogeography in the Mediterranean. *Taxon*, **54**, 877–885.

Qiu Z-B, Yuan M-M, Hai B-Z, Wang L, Zhang L (2016) Characterization and expression analysis of conserved miRNAs and their targets in *Pinus densata*. *Biologia Plantarum*, **60**, 427–434.

Quinn CR, Iriyama R, Fernando DD (2014) Expression patterns of conserved microRNAs in the male gametophyte of loblolly pine (*Pinus taeda*). *Plant Reproduction*, **27**, 69–78.

Rajwanshi R, Chakraborty S, Jayanandi K, Deb B, Lightfoot DA (2014) Orthologous plant microRNAs: microregulators with great potential for improving stress tolerance in plants. *Theoretical and Applied Genetics*, **127**, 2525–2543.

Reznick DN, Ricklefs RE (2009) Darwin's bridge between microevolution and macroevolution. *Nature*, **457**, 837–842.

Rothwell GW, Mapes G, Stockey RA, Hilton J (2012) The seed cone Eathiestrobus gen. nov.: Fossil evidence for a Jurassic origin of Pinaceae. *American Journal of Botany*, **99**, 708–720.

Sáez-Laguna E, Guevara M-Á, Díaz L-M *et al.* (2014) Epigenetic variability in the genetically uniform forest tree species *Pinus pinea* L. *PLoS ONE*, **9**, e103145.

Salinas S, Brown SC, Mangel M, Munch SB (2013) Non-genetic inheritance and changing environments. *Non-Genetic Inheritance*, **1**, 38–50.

Saracino A, Leone V (2001) Survival strategies and recovery mechanism after fire in the Mediterranean environment: the case of Aleppo pine forests. *Monti e Boschi*, **52**, 38–46.

Sathyan P, Newton RJ, Loopstra CA (2005) Genes induced by WDS are differentially expressed in two populations of aleppo pine (*Pinus halepensis*). *Tree Genetics & Genomes*, **1**, 166–173.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature reviews. Genetics*, **14**, 807–20.

Schiller G, Conkle MT, Grunwald C (1986) Local diefferentitation among Mediterranean populations of Aleppo pine in their isoenzymes. *Silvae Genetica*, **35**, 11–19.

Scotti I, González-Martínez SC, Budde KB, Lalagüe H (2016) Fifty years of genetic studies: what to make of the large amounts of variation found within populations? *Annals of Forest Science*, **73**, 69–75.

Shriram V, Kumar V, Devarumath RM, Khare TS, Wani SH (2016) MicroRNAs As Potential Targets for Abiotic Stress Tolerance in Plants. *Frontiers in Plant Science*, **7**.

Taylor RS, Tarver JE, Foroozani A, Donoghue PCJ (2017) MicroRNA annotation of plant genomes − Do it right or not at all. *BioEssays*, **39**, 1600113.

Taylor RS, Tarver JE, Hiscock SJ, Donoghue PCJ (2014) Evolutionary history of plant microRNAs. *Trends in Plant Science*, **19**, 175–182.

Thatcher SR, Burd S, Wright C, Lers A, Green PJ (2015) Differential expression of miRNAs and their target genes in senescing leaves and siliques: insights from deep sequencing of small RNAs and cleaved target RNAs. *Plant Cell and Environment*, **38**, 188–200.

Van't Hof AE, Edmonds N, Dalikova M, Marec F, Saccheri IJ (2011) Industrial melanism in british peppered moths has a singular and recent mutational origin. *Science*, **332**, 958–960.

Voltas J, Chambel MR, Prada MA, Ferrio JP (2008) Climate-related variability in carbon and oxygen stable isotopes among populations of Aleppo pine grown in common-garden tests. *Trees - Structure and Function*, **22**, 759–769.

Wegrzyn JL, Liechty JD, Stevens KA *et al.* (2014) Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics*, **196**.

Yakovlev IA, Fossdal CG, Johnsen Ø (2010) MicroRNAs, the epigenetic memory and climatic adaptation in Norway spruce. *New Phytologist*, **187**, 1154–1169.

Hieronymus Bosch, *The Garden of Earthly Delights*, Museo del Prado, Madrid

Chapter 2 **Inferring selection in instances of long range colonisation: the Aleppo pine (*Pinus halepensis*) in the Mediterranean Basin**

Rose Ruiz Daniels[1], Richard S. Taylor[2], María Jesús Serra Varela[1,3,4], Giovanni G. Vendramin[5], Santiago González-Martínez[4,6]& Delphine Grivet[1,4]

[1] Department of Forest Ecology and Genetics, Forest Research Centre, INIA, Carretera A Coruña km 7.5, 28040 Madrid, Spain.

[2] School of Earth Sciences, University of Bristol, Wills Memorial Building, Queen's Road, Bristol BS8 1RJ, UK

[3] Department of Plant Production and Forest resources, University of Valladolid, Avda. de Madrid 44, 34004 Palencia, Spain

[4] Sustainable Forest Management Research Institute, INIA, University of Valladolid, 34004 Palencia, Spain

[5] Institute of Biosciences and Bioresources, National Research Council, I-50019 Sesto Fiorentino (FI), Italy

[6] BIOGECO, INRA, Univ. Bordeaux, 33610 Cestas, France

2.1 Abstract

Teasing apart the effect of natural selection and demography, more specifically the increase of genetic drift at the edges of range expansions (a process known as gene surfing), on current allele frequencies is a challenging task, especially when considering that both selection and gene surfing leave a similar molecular footprint. We address this question by assessing with molecular markers (SSRs and SNPs) the long-range colonisation history of Aleppo pine (*Pinus halepensis* Mill) across the Mediterranean Basin, and by identifying putative SNPs under selection using genetic-environment association (*Bayenv2*) and F_{ST}-related (*PCAdapt*) methods. Furthermore, we incorporate effects of allele surfing by simulating populations that mimic the species past range expansions, discarding thereby loci displaying false signal of selection. It was found that Aleppo pine shows a previously unsuspected complex genetic structure across its range, as well as evidence of local selection at few SNPs in response to environmental variables mostly related to drought. This study contributes to the increased evidence supporting that plant populations are able to adapt to new environments despite the expected accumulation of deleterious mutations that takes place during long-range colonisations.

2.2 Introduction

The establishment of a species in new territories will depend on its capacity of adapting to new biotic and abiotic conditions. During range expansions, populations will be confronted with selection and adaptation to different environments, as well as to new conspecific and heterospecific interactions (Hewitt 2000). Spatial expansion into new territories results in loss of genetic diversity, as populations are subjected to a series of founder effects following fragmentation and the colonisation of new areas by a smaller number of individuals (Slatkin & Excoffier 2012). During this process, rare alleles can spread by chance over wide areas where they can reach very high frequencies, due to the increase effect of genetic drift (Edmonds *et al.* 2004). This stochastic effect is known as gene surfing and is a direct result of increased genetic drift at populations located at the edge of an expansion (Edmonds *et al.* 2004; Klopfstein *et al.* 2006; Hallatschek & Nelson 2008; Travis *et al.* 2010). This increase in the frequency of rare alleles could be confounding with the increase of allelic frequency of a beneficial mutation propagated by selection/adaptation (Klopfstein *et al.* 2006; Excoffier & Ray 2008; Hofer *et al.* 2009). The confounding outcome of selection and gene surfing is further exacerbated in that gene surfing, unlike most other demographic processes, does not affect all loci in a genome, but only affects a subset of loci (Excoffier *et al.* 2009) like directional selection (Begun *et al.* 2007). Accounting for the effect of gene surfing is therefore critical when investigating loci underlying the response to adaptive constraints in species that have undergone long-range expansions. Very few studies have explored the effect of allele surfing on inferring selection, although the importance of doing so has been stated (De Mita *et al.* 2013; Lotterhos & Whitlock 2015: Hoban *et al*. 2016). Empirical works on the genetic dynamics in range expansions have been documented in various organisms (e.g. Graciá *et al.* 2013; White *et al.* 2013; Pierce *et al.* 2014; Swaegers *et al.* 2015; Gralka *et al.* 2016; Jezkova *et al.* 2016; Peischl *et al.* 2014 and 2016), but are still scarce in plants, where long distance dispersal (LDD) events play a pivotal role in shaping genetic diversity during recolonization (Excoffier *et al.* 2009), with higher levels of LDD leading to higher levels of genetic diversity because the homogenizing effect of surfing is prevented (Le

Corre & Kremer 1998). Therefore, depending on the plant characteristics (i.e. capacity for LDD), as well as the environment it has colonized, range expansion could be accompanied or not with gene surfing – and disentangling the neutral and adaptive processes operating during range expansion is challenging.

Two main approaches are used to detect signals of natural selection, correlation methods that look for correlations of spatially diverging populations with environmental variables (Rellstab *et al.* 2015), or outlier tests that look for variants with unexpectedly large differences of allele frequencies among populations (usually done based on F_{ST}-related test statistics; see François *et al.* 2016). Correlation methods aim at assessing spatial adaptive genetic variation by correlating loci with allele frequency changes in geographic space with changes in environmental variables. These methods are appealing in that they can not only identify loci subject to adaptive constraints, but can also factor in the ecological variables that differentiate amongst populations and therefore be potentially responsible for selective constraints. Several approaches have been recently proposed in order to evaluate the association of ecological variables with genetic differentiation (e.g. Frichot *et al.* 2015; Gautier 2015; and see Rellstab *et al.* 2015 for a review). Because these methods sometimes generate high false-positive rates (see review in Pannell & Fields 2014), it is critical to account for the underlying neutral correlation structure across populations (De Mita *et al.* 2013; Frichot *et al.* 2015; Rellstab *et al.* 2015). Among the different approaches developed to detect selection along environmental gradients, the Bayesian method developed by Coop *et al.* (2010), and implemented in *Bayenv* (Gunter & Coop 2013), has been shown to have good statistical power compared to other approaches (De Mita *et al.* 2013; Gunter & Coop 2013), and it performs well when identifying loci under spatially divergent selection (Lotterhos & Whitlock 2014), even under instances of Isolation by Distance (De Mita *et al*. 2013). However, although this method explicitly incorporates neutral spatial autocorrelation by integrating a covariance matrix, it does not take into account explicitly historical demographic change (e.g. range expansion, contraction).

Approaches based on F_{ST}-related measures are facing several challenges (accounting for hierarchical population structure, grouping individuals into populations, and integrating multilocus datasets from next generation sequencing; see Luu *et al.* 2017), and to address them, a recent method based on Principal Component Analysis (PCA) was developed, implemented in the *PCAdapt* R package (Luu et *al.* 2017). This method does not require to group individuals into populations and allelic variants can be related to different evolutionary events that correspond to the different principal components (Duforet-Frebourg *et al.* 2016; Luu *et al.* 2017). At the genome wide level there is a clear relationship between F_{ST} and PCA (McVean 2009), and simulations have shown that this relationship is also applicable at the level of single variant (Duforet-Frebourg *et al.* 2016). Furthermore, this method is particularly well adapted in scenarios of range expansion where population structure is continuous, and it has been shown to be more powerful compared to other genome scan methods tested (Luu *et al.* 2017).

In this study, we use molecular and environmental data to disentangle the neutral and adaptive processes operating during range expansion in Aleppo pine (*Pinus halepensis* Mill.), a conifer with circum-Mediterranean distribution that is geographically fragmented and spans 3.5 million hectares. Its specific distribution translates into a different colonization history compared to most European forest trees (Hewitt 2000), but in line with the longitudinal genetic diversity pattern detected for woody taxa in the Mediterranean Basin (Fady & Conord 2010), with eastern and more likely older populations showing higher levels of genetic diversity than the less diverse and possibly newer western populations (Morgante *et al* 1998; Bucci *et al.* 1998; Grivet *et al.* 2009). Genetic data together with paleoecological data and fossil record suggest that the actual distribution of Aleppo pine would have resulted from several range expansions, including an old one in the late Pleistocene (Jaramillo *et al.* 2010), and a more recent one after the Last Glacial Maximum (Nahal 1962; Schiller *et al.* 1986; Pons 1992; Grivet *et al.* 2009). This colonization is likely to have followed an East to West axis, accompanied by a combination of bottlenecks and series of founder events, leading possibly to the spatial spread of alleles at the edge of the expansion range (the so-called gene surfing effect). Additionally, admixture zones present along its distribution range suggest that Aleppo pine populations are connected by recurrent gene flow (Serra-Varela et al. 2017).

During its range expansion across Europe and North Africa, Aleppo pine would have encountered a range of new ecological selective pressures likely to result in adaptations to these conditions, especially regarding the West-East climatic dipole in the Mediterranean (Van Andel, 2002). Although this pine is highly plastic (De Luis *et al.* 2013; Santos-del-Blanco *et al.* 2013), presently living in a wide ecological range of moisture and thermal conditions, it is highly adapted to dry Mediterranean climates, and presents adaptive variation in response to climatic conditions such as precipitation and duration of dry season (Voltas *et al.* 2008). Molecular work has also suggested that some genes commonly linked to conifer adaptation have been the target of selection in this species (Grivet *et al.* 2009, 2011). While several lines of evidence have pointed to Aleppo pine as being the target of natural selection during its recolonization of the western Mediterranean, until now it has been difficult to fully address this issue due to the lack of genomic resources and the confounding effects of demography and positive selection.

This study uses Aleppo pine to explore how demographic processes, including gene surfing, and selection have shaped current allele frequencies in a widespread forest tree. To realise that aim our objectives are threefold: (i) Adding new insights to the recolonization history of Aleppo pine by increasing substantially the number of populations and genetic markers analysed; (ii) Identifying putative loci under selection based on genetic-environment association methods and on differentiation outlier methods; (iii) Assessing with simulated range expansion data sets whether the SNP candidates for local adaptation are not spurious due to the effect of gene surfing during range expansion.

2.3 Materials and Methods

2.3.1 Sampling and DNA preparation

2.3.1.1 Plant material and DNA extraction

Aleppo pine populations were extensively sampled across the natural range of the Aleppo pine across the Mediterranean Basin. In total, 1,046 individuals were collected from 44 populations for nuclear microsatellites (SSRs) (see details in Table S1 and Figure S1, Supporting Information), while 1,326 individuals were collected from 49 populations for Single Nucleotide Polymorphisms (SNPs, see details in Table S2 and Figure S2). About 74% of the populations were common to both datasets.

For each individual genotyped, 50 mg of pine needles was dried with silica gel and grinded in QIAwell (Qiagen, Venlo, The Netherlands) plate homogenized with mixer mill MM300 (RETSH, Haan, Germany) under liquid nitrogen. DNA extractions were carried out with the kit Invisorb DNA plant HTS 96 kit (Invitek, Hayward, CA) following the manual instructions. DNA was quantified with Nanodrop 10000 (Thermo Fisher Scientific, Wilmington, DE).

2.3.1.2 Environmental variables

We characterized each population with 27 environmental variables in order to assess SNP-environment associations. These variables were representative of the period 1950-2000 and included: (i) 19 bioclimatic variables available in WORLDCLIM (Hijmans *et al.* 2005); (ii) four compound variables created following Zimmermann et al. (2007) to characterize water availability: summer and spring potential evapotranspiration (ETPTsummer; ETPTspring), as well as summer and spring moisture indexes (MINDsummer; MINDspring); (iii) four aridity indexes (one per quarter of the year - aridity Q1-Q4) were calculated following Eckert *et al.* (2010) (see Table S3 for more details on bioclimatic data, Supporting Information). We also included in the analysis latitude and longitude to explore the effects of gene surfing in the real compared to simulated data sets. When more than one environmental variable was significantly associated with a SNP, Pearson correlation among pairs of environmental variables were calculated in order to determine their statistical dependence.

2.3.2 Genotyping

2.3.3 Nuclear microsatellites (SSRs)

Aleppo pine populations were genotyped at nine SSRs: epi3 (Budde *et al.* 2014), FRPP94 and ITPH4516 (Mariette *et al.* 2001), NZPR544 and B4F08 (Guevara *et al.* 2005), pEST2669 and pEST8 (Steinitz *et al.* 2011), PtTX3030 and PtTX3116 (González-Martínez *et al.* 2004). Forward primers were labelled with fluorochromes on the 5' end and amplifications were performed using the Qiagen Multiplex PCR kit (Qiagen, Venlo, The Netherlands). Amplified fragments were separated using an ABI 3730 genetic analyser (Applied Biosystems, Carlsbad, Ca), and fragment sizes were identified using the GeneScan™ LIZ ® Size Standard (Applied Biosystems) using the software GeneMapper version 4.0 (Applied Biosystems). For more details on SSR markers and the amplification protocol, see Budde et al. (2014). The scored loci were then analysed with Micro-Checker 2.2. (Van Oosterhout *et al.* 2004) to test for null alleles and scoring errors: all but one locus (NZPR544) did not display any signal of null alleles and the corresponding eight loci were used for further analyses.

2.3.3.1 Single Nucleotide Polymorphisms (SNPs)

Aleppo pine populations were successfully genotyped at 294 SNPs (conversion rate of 76.56 %) using a newly-generated 384-plex SNP assay with Illumina VeraCode technology as described in Pinosio *et al.* (2014). SNP markers in this assay originated from two sources: (i) 144 SNPs were taken from 117 polymorphic amplicons obtained from direct sequencing of haploid (megagametophytes) tissue sampled from the full distribution of the species (www.evoltree.eu;

CRIEC initiative and unpublished results); (ii) 240 SNPs were taken from 28,236 SNPs discovered through alignment of the transcriptome of two trees with contrasted phenotypes. See Pinosio *et al*. (2014) for further details. Loci for which SNPs were found putatively under selection were annotated from homology with other plant species using Geneious version 6.1 (Biomatters, available from http://www.geneious.com/), searching against known plant EST contigs and the NCBI reference protein database.

2.3.4 Genetic variation within and among populations

For both SSR and SNP datasets, global genetic differentiation (F_{ST}) was estimated as well as its 95% confidence interval (bootstrap of 1,000 replications), using GDA (Weir 1996). Different genetic diversity parameters were estimated across all populations using SPAGeDi (Hardy & Vekemans 2002): number of alleles (*NA*), effective number of alleles (*NAe*; Nielsen *et al.* 2003), allelic richness (*Ae*), and expected heterozygosity or gene diversity corrected for sample size (*He*; Nei 1972). The same genetic parameters were also calculated for distinct gene pools identified at the smallest cluster level (*K*=2) with the Bayesian approach STRUCTURE (see 'Population structure and demography' section below). In order to contextualise genetic diversity results across the Mediterranean Basin, a spatial representation of effective number of alleles (*NAe*) per population from the SSR data was drawn in ArcGIS, using an Inverse Distance Weighted (IDW) technique, with a variable search on 44 points corresponding to the populations sampled for these markers.

2.3.5 Population structure and demography

To infer population structure, the Bayesian clustering method STRUCTURE (Pritchard *et al.* 2000) was used on both SSR and SNP datasets, with the following parameters: number of clusters (*K*) set from 1 to 20; number of iterations set to 10; number of steps set to 100,000; with a burn in period of 10,000 to minimize the effect of the starting configuration, and with an admixture model of ancestry. The number of genetic groups (*K*) for each marker was explored following the Evanno's method (Evanno *et al.* 2005), using STRUCTURE HARVESTER (Earl & vonHoldt 2011) that plots the rate of change of the mean likelihood distribution and its variance as a function of *K*. To visualise how clusters were distributed across the natural range of Aleppo pine, the STRUCTURE outputs for the best *K*s were plotted on a map of the Mediterranean Basin, using the program ArcGIS. A Mantel test was used to coarsely explore the relationship of genetic relatedness and geographical distance, using a Matrix of pairwise F_{ST} with our SNP data calculated in GENEPOP V4.3 (Rousset 2008) against a matrix of geographical distances, calculated using the package *ade4* in R.

We finally applied, suing SNP data, the directionality index ψ as described in Peter & Slatkin (2013) to infer possible origin and direction of range expansion. This statistic uses the clines in the frequencies of neutral low-frequency alleles created during range expansion (increasing frequency in the direction of the expansion) to detect patterns created by successive range expansions.

2.3.6 Range expansion simulations

False signals of selection can be produced by the 'gene surfing' effect on the colonisation fronts of recent range expansions (Excoffier & Ray 2008). In order to test whether the spatial pattern of SNPs under selection was due to range expansion rather than (or in addition to) selection, we simulated various range expansion scenarios without selection with a modified version of SPLATCHE 2 (Ray *et al.* 2010). This simulation programme uses a coalescent-based approach to simulate different population histories forward in time. A spatial representation of the Mediterranean Basin was modified to be able to fit the program requirement (with one 'deme' comprising of 200 square km) using ArcGis. Land territory was classified as suitable for colonisation and areas of sea as unsuitable, and resistance across the natural range of Aleppo pine was standardised in order to recreate a range expansion with no barriers, based on the biology of the species (wide ecological niche, great capacity for dispersal, broad spatial distribution; see Fady *et al.* 2003). The simulation was run considering 29 simulated populations that were analogues to our real populations, and that covered most of the natural range of the Aleppo pine. Simulations (294 SNPs) were performed for a range expansion from East to West that lasted 9,000 years as an approximation of the historical time of the Aleppo's pine range expansion (Nahal 1962; Schiller *et al*. 1986; Pons 1992). In order to find a virtual simulation analogous to our potential real situation, simulations were run until a colonisation scenario was achieved that emulated that of the Aleppo pine in the Mediterranean Basin. The quality of the match between simulated and real populations was checked by comparing their heatmaps of pairwise F_{ST} (see Figure S9, Supporting Information) using a Mantel test with 100,000 permutations carried out in R with the packet *ade4*. The following final parameters were then selected at 30 years generation time (Brown *et al.* 2004; Willyard *et al.* 2007), migration rate (fraction of the deme population that will emigrate at each generation) = 0.18, number of generations = 300, and growth rate = 0.18. We then varied the migration rate parameter to include values of 0.10, 0.18, 0.26 and 0.34 to simulate differing levels of admixture between the demes.

Each of these four simulated datasets was then analysed using the two selection tests controlling for shared history and gene flow (*Bayenv2* and *PCAdapt*) in the same manner described as for their analogous real populations (see Selection tests controlling for shared history and gene flow). The distribution of Bayes factors obtained from *Bayenv2,* and that of the *p*-value obtained from *PCAdapt* were then compared with those obtained with real populations (see Selection tests controlling for shared history and gene flow). These simulations, in addition to give some insights on the effect of gene surfing in detecting outliers, allow assessing the levels of admixture between populations.

2.3.7 Selection tests controlling for shared history and gene flow

In order to identify selection acting on SNPs, we used the Bayesian method described in Coop *et al.* (2010) to evaluate the association of environmental variables (i.e. climatic variables in the present study) with genetic marker differentiation (i.e. SNPs in the present study). This method is implemented in the *Bayenv2* package, which estimates the empirical pattern of covariance in allele frequencies between populations, and uses this as the null model to test each individual SNP for selection. This methodology accounts for the neutral correlation of allele frequencies across populations, a critical issue when there is evidence of shared history and gene flow among populations as in the Aleppo pine (see Results below). To minimize the stochasticity between runs, 20 covariance matrices were produced, from 2 million iterations each, and the mean of these matrices was used. Our confidence in this covariance matrix representing the true variance of allele frequencies across populations was tested by comparing its heatmap with that of pairwise F_{ST} calculated using Genepop vs 4.5 (Raymond & Rousset 1995; Rousset 2008), for the SNP dataset (Figure S6, Supporting Information). Covariance matrices were built using the *heatmap* function in R, and then compared using a mantel test using the R package *ade4*. Detecting environmental associations was then performed testing, by means of Bayes factors, whether a correlation between the allele frequencies at a SNP and an environmental variable is greater than expected given the null model. This approach was repeated three times for each SNP-environmental variable tested for association, in order to account for instability between independent runs (Blair *et al.* 2014). The mean of the three runs was then used to infer the final Bayes factors. A cut-off point of minor allele frequency (MAF) was not applied for this analysis, where all SNPs were considered.

The *PCAdapt* (version 2.0 of the R package) method (Luu *et al.* 2017) was run on SNPs with minor allele frequencies higher than 5% to detect selection events based on principal component analysis (PCA). *PCAdapt* was run with a number of principal components *K*=10, and the number *K* explaining most of the variation was selected using the graphical approach based on the scree plot (Jackson 1993), as recommended by the authors of the software. Then the Mahalanobis distances were used as the statistic test to look for outlier SNPs, and transformed into p-values to perform multiple hypothesis testing. The distribution of the p-values was checked using a Q-Q plot of the expected p-values versus observed p-values. Finally, the cut-off for detecting outliers was chosen based on the q-value procedure implemented in the *qvalue* R package (R Core Team 2013), using 0.01% as false discovery rate threshold.

2.3.8 Quantifying SNPs under putative selection, taking into account gene surfing

Determining a suitable cut-off for significant Bayes factors is not trivial due to the difficulty in obtaining well calibrated statistics as outlined in Coop *et al*. (2010). Different methods were used to determine Bayes factors of significance, i.e. the top highest Bayes factors (e.g. Alberto *et al.* 2013), the Jeffreys' scale alone (Pujolar *et al.* 2014) or in combination with Spearman's rank correlation (e.g. De La Torre *et al.* 2015; Jaramillo-Correa *et al.* 2015), or the percentage cut-off methods based on p-values (e.g. Hancock *et al.* 2011; Pyhäjärvi *et al.* 2013; White *et al.* 2013) or q-values (Frichot *et al.* 2015; Ring 2015). In the present study, we combined three complementary approaches to validate the SNPs potentially under selection obtained with *Bayenv2*: (i) we first took the top 0.1% of all Bayes factors obtained from all SNP/Environmental variable interactions. The p-value (0.001) was calculated using the following conservative formula: $p = (r+1)/(n+1)$ (North *et al.* 2002; Davison & Hinkley 1997), where r is the number of tests that have produced a test statistic greater than or equal to the result calculated for the actual data ($r = 7$), and n is the total number of tests ($n = 8032$); (ii) these top Bayes factors were then compared to those obtained in the simulated data to confirm that they are not a product of gene surfing, as if they were to be equal or lower in value than those obtained in the simulated data, it would suggest they could be false positives. To establish how values and distributions of Bayes factors varied between the real and the four simulated datasets, Q-Q plot were produced for each of the five colonisation scenarios using R, where the quantiles of Bayes factors were plotted against the quantiles of a chi2 distribution with one degree of freedom in order to compare visually the different simulated date sets with the real data sets. (iii) The cut-off for the Bayes factors was based on a cut of the continuous distribution for the highest values. These top Bayes factors (2ln*K*) were then interpreted with the Jeffreys' scale (1961) (K>$10^{1/2}$). Similarly to point (ii), the Q-Q plot of the *p*-values based on real data and obtained with *PCAdapt* was compared to the Q-Q plots for the four simulated data in order to discard any false positive due to range expansion.

2.4 Results

2.4.1 Genetic variation within and among populations

As expected, genetic diversity corrected for sample size (*He*) was significantly higher in the SSR data than in SNP data, both for the gene diversity corrected for sample size (*He* = 0.47 versus 0.23, respectively) and for the effective number of alleles (*NAe* = 2.16 versus 1.40, respectively) (**Table 2.1**). When using the more diverse SSR markers to examine the effective number of alleles in a spatial context, an East-West split across the natural range of the Aleppo pine was observed, the Western part of the range presenting much lower diversity, especially for the Iberian Peninsula (**Figure 2.1** and Table S2, Supporting Information). The genetic diversity statistics at the minimum number of clusters level detected by STRUCTURE (*K*=2) show that the eastern cluster displays higher levels of genetic diversity than the western cluster (**Table 2.1**) This difference is significant for the SSR data (no overlap of the 95% CI for both *NAe* and *He*), whereas the SNP data exhibit a similar trend but the results were not statistically significant at the 95% level (**Table 2.1**).

Global F_{ST} for SSRs (0.167) and for SNPs (0.168) were not significantly different (**Table 2.1**). At *K*=2, the F_{ST} for the SNPs was significantly lower in the West (0.066) than in the East (0.195), while the F_{ST} for the SSRs was also lower in the West (0.107) than the East (0.131) but not at a statistically significant level. The heatmap of the pairwise population F_{ST} confirms the genetic closeness of the western populations (Figure S6, Supporting Information), with populations on the Iberian Peninsula exhibiting much lower genetic differentiation than populations from the other locations.

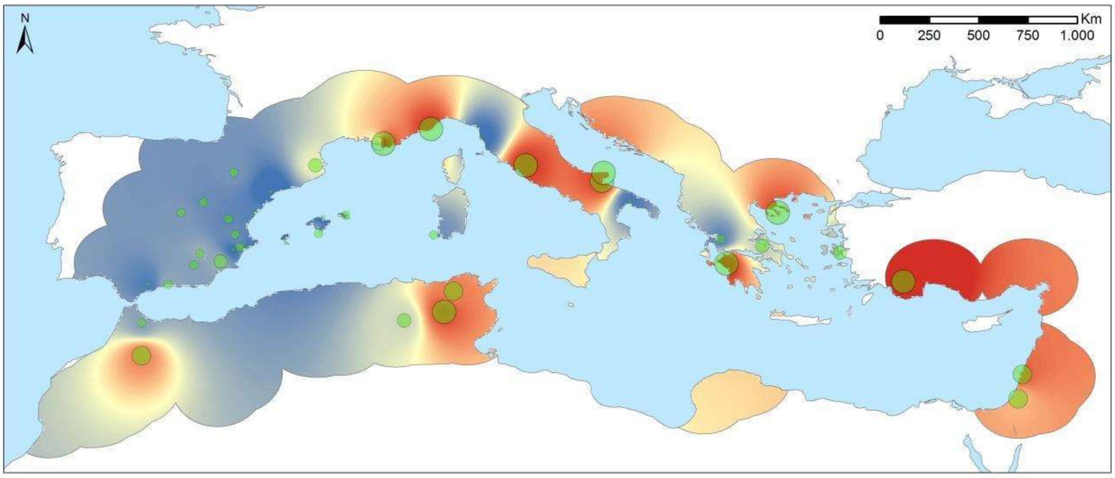


Figure 2.1: Spatial representation of the effective number of alleles (*NAe*) from the SSR dataset. The dots are proportional to the amount of effective number of alleles, with the colour scheme representing the magnitude of *NAe* (red: higher effective number of alleles; blue: lower effective number of alleles).

	N_{ind}	N_{pop}	*NA*	*NAe*	95% CI	*He*	95% CI	F_{ST}	95% CI
SSRs									
All	1045	45	10.63	2.16	2.00-2.32	0.47	0.44-0.49	0.167	0.128-0.211
West	716	29	6.88	1.80	1.77-2.13	0.40	0.39-0.45	0.107	0.082-0.133
East	310	15	10.13	2.92	2.50-2.93	0.53	0.53-0.59	0.131	0.116-0.145
SNPs									
All	1325	49	1.94	1.40	1.39-1.41	0.23	0.23-0.24	0.168	0.155-0.180
West	1019	30	1.87	1.39	1.38-1.41	0.23	0.22-0.24	0.066	0.061-0.070
East	306	15	1.92	1.41	1.39-1.44	0.24	0.23-0.26	0.195	0.183-0.207

Table 2.1: Summary statistics for SSRs and SNPs across all populations, as well as for the two western and eastern gene pools as defined with STRUCTURE. Missing data was lower than 3% in all cases. N_{ind}: number of individuals; N_{pop}: number of populations; *NA*: number of alleles; *NAe*: mean per population of effective number of alleles (Nielsen *et al.* 2003); *He*: mean per population gene diversity corrected for sample size (Nei 1978); F_{ST}: genetic differentiation.

2.4.2 Population structure and demography

Bayesian clustering revealed significant population structure across Aleppo pine natural range (**Figure 2.2**). The Evanno *et al.* (2005) method indicates K=2 as the optimal number of genetic clusters for both molecular datasets, defining a western and an eastern gene pool (**Figure 2.2** and Figure S4 and S5, Supporting Information). However, the curve of ln$K$ plateaus around K=7 for both molecular datasets (Figure S4 and S5, Supporting Information), indicating a potential higher number of clusters, with the SNP data, which seems to indicate a finer scale substructure of the main eastern and western gene pools (**Figure 2.2** and S3, Supporting Information, and Serra-Valera *et al*. under review). This pattern is less obvious in the SSRs (Figure S1 and S2, Supporting Information), and reflects a lower resolution power in inferring population structure of this type of markers (Swaegers *et al*. 2015). Membership coefficients (*Q*) to more than one gene pool indicates admixture areas, i.e. populations shared between gene pools (Figure S3, Supporting Information, and Serra-Varela *et al*. under review). The Mantel test revealed a significant relationship between the genetic distances between sampling sites and their geographic distances (R^2=0.549; *p*-value = 0.001), translating the isolation-by-distance neutral pattern of genetic variation due to limited dispersal across space (Wright, 1943). The directionality index ψ pointed at an eastern origin of range expansion (the exact location being uncertain), with gene flow mainly directed from East towards West.

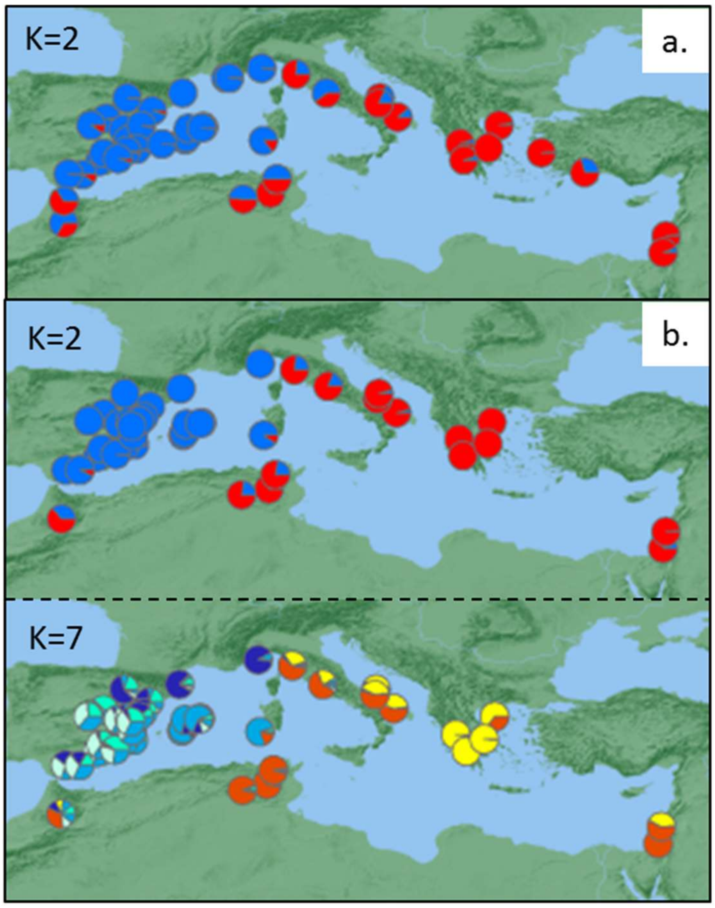


Figure 2.2: Bayesian clustering analysis of SSRs (a) and SNPs (b) performed with STRUCTURE.

2.4.3 Natural selection during long-range colonisation

2.4.3.1 Genetic-environment association

Independent covariance matrices (*N*=20) used to compute the null model in *Bayenv2* had a low standard deviation (0.102), indicating little variation among them. Confidence in the mean covariance matrix representing the true variance of allele frequencies across populations was also reinforced by the similarity between its heatmap and that of the pairwise population F_{ST} as shown by the significant correlation (*p*-value = 0.0001) of the Mantel test (R^2=0.903) (Figure S6, Supporting Information). Following the three criteria used to establish the cut-off value for the Bayes factors, we found that: (i) the top 0.1% of the Bayes factors in our data set gives a cut-off point of 7.90; (ii) of these top 0.1% Bayes factors none was lower than Bayes factors obtained in the simulated data (**Figure 2.3**); (iii) the top Bayes factors based on the cut of the continuous distribution corresponds to BF>8 (**Figure 2.3**). Based on these three criteria, seven SNP-environmental variable associations were identified as having the potential of being due to selection (**Table 2.2**). Two of these SNPs (SNP183 and SNP375) were found to be associated to more than one environmental variable, but these variables were highly correlated (Pearson correlation r>0.95).

Our confidence that these seven outlier SNPs are the outcome of selection and not of recolonization is based upon the comparison between real and simulated datasets. The simulated data allowed us to account for the effect of migration rate, and therefore the intensity of a potential gene surfing effect, in detecting significant SNP-environment associations, by establishing the upper threshold of Bayes factors expected by drift alone. The simulated data with the smallest migration rate (m=0.1) presented a pattern of population structure most similar to that of the real data, as shown by the Mantel test comparing the matrix of pairwise F_{ST} (R^2 = 0.687; p-value<0.0001; Figure S8), as well as more closely mimicking the value and distribution of the observed Bayes factors (**Figure 2.3**). Under this scenario, higher Bayes factors were found compared to the other scenarios but still lower than those observed for the seven outlier SNPs, suggesting that these SNPs may be under selection.

SNPs	Name	Environmental variable	K	Environmental variable	K	Strength of evidence	Correlation environmental variables
SNP 54	seq-8115-2383	aridityQ4	8.08			substantial	
SNP 151	seq-0_8992_01-119	BIO14	9.16			substantial	
SNP 183	seq-35072-1606	BIO12	8.62	aridityQ4	8.6	substantial	0.96
SNP 186	seq-0_6924_02-89	BIO2	9.22			substantial	
SNP 374	seq-11269-769	aridityQ4	8.09			substantial	
SNP 375	seq-51080-670	BIO18	11.91	aridityQ3	13.91	strong	0.99
SNP 378	seq-2_3941_01-381	BIO14	9.88			substantial	

Table 2.2: SNPs under high likelihood of being under selection and the environmental variables associated, for Bayes factor *K* (with their interpretation following Jeffreys' scale). When more than one environmental variable were associated for a given SNP, Pearson correlation between the environmental variables is indicated (with 0 indicating no correlation and 1 indicating high correlation). Aridity Q3: aridity of quarter 3; Aridity Q4: aridity of quarter 4; BIO2: mean diurnal range temperature; BIO7: annual temperature range; BIO12: annual precipitation; BIO14: precipitation of driest month; BIO18: precipitation of the warmest quarter.

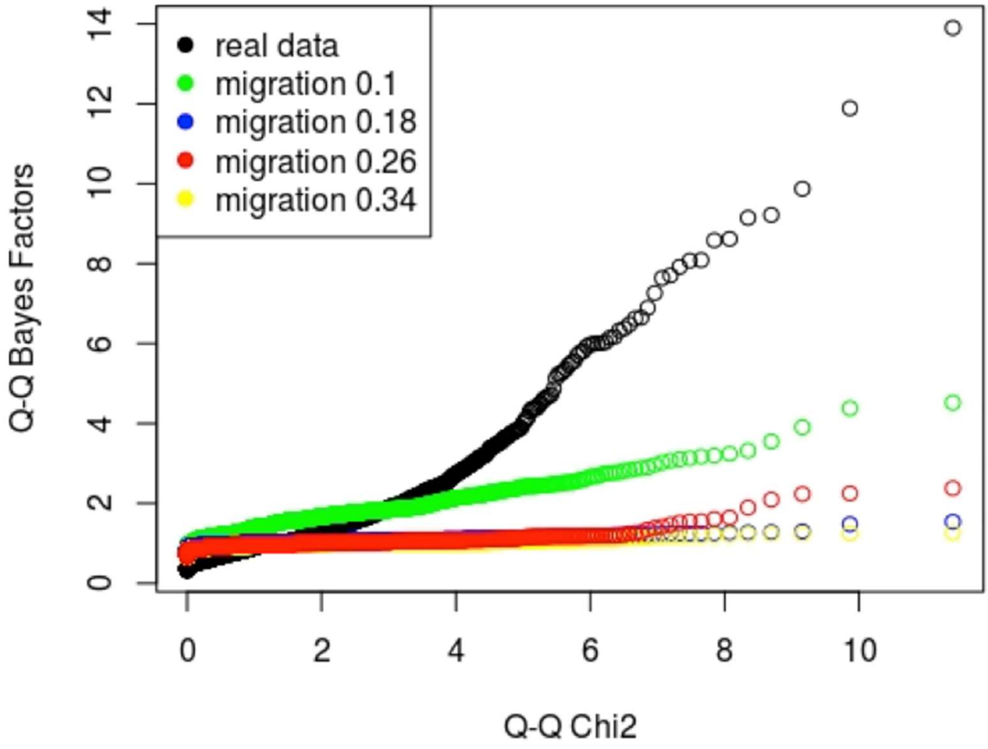


Figure 2.3: Quantile-quantile plots comparing the quantile chi2 distribution with the quantile of Bayes factors for real data and simulated data (for migration rates 0.10, 0.18, 0.26 and 0.34).

2.4.3.2 Principal Component Analysis

The scree plot indicates *K*=2 as explaining most of the genetic variation in the dataset, and was used in subsequent *PCAdapt* analyses to detect outlier SNPs. The Q-Q plot of expected *vs*. observed *p*-values indicates that most of the *p*-values follow the expected uniform distribution, while the smallest *p*-values (<0.3) are smaller than expected confirming the presence of outliers. Based on the false discovery rate of 0.01%, seven SNPs were detected as outliers (**Figure 2.4**).

Comparison between Q-Q plots of expected *vs*. observed *p*-values between real and simulated data reveals the following pattern: simulations with the smallest migration rate (m=0.1) give the closest pattern to the real data, with most of the *p*-values following the expected uniform distribution, while the FDR of 0.01% detects the presence of five outliers (**Figure 2.4**). Then, no outliers were detected for m=0.18, and many outliers were detected for m=0.26 and m=0.34 (Supporting Information). Because outliers were detected both in the real and simulated data (with m=0.1), we cannot discard that the outlier SNPs detected with *PCAdapt* are false positive. None of the outlier SNPs detected with *PCAdapt* corresponds to those detected with *Bayenv2*.

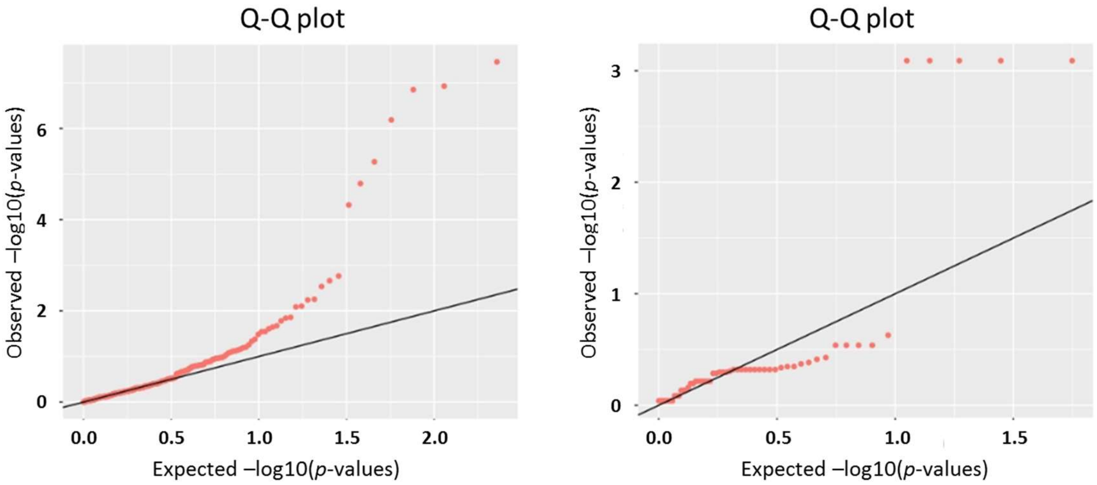


Figure 2.4: Quantile-quantile plots showing outliers from the *PCAdapt* analysis from the real data (7 outliers) (a) and simulated data for migration rate of 0.10 (5 outliers) (b).

2.5 Discussion

This study not only confirms previous demographic inferences, but also provides a more complete and complex picture of the evolutionary history of Aleppo pine than previous work based on only few molecular markers and populations. Furthermore, we addressed the issue of detecting outlier SNPs taking into account population structure and gene surfing that can lead to incorrect inference, providing one of the few empirical studies showing the importance of controlling for demography when identifying specific alleles involved in local adaptation in expanded populations.

2.5.1 Spatial genetic structure and demography.

The two sets of analysed markers clearly point to an East/North Africa (Israel, Turkey, Greece, central and southern Italy, Algeria, Morocco) - West (Northern Italy, Sardinia, southern France, Balearic Islands, Iberian Peninsula) spatial disjunction, with populations from the eastern gene pool harbouring more alleles, translating to higher level of genetic diversity and differentiation. The zones located at the junction of the two gene pools (northern Italy, Algeria, and Morocco) present genetic admixture, translating substantial levels of gene flow. Furthermore, the directionality index ψ clearly pointed at an eastern origin of range expansion, although with an uncertain exact location probably due to the lack of some sampling coverage in the East.

Combining the results obtained in this study with previous population genetic studies based on molecular markers (Bucci *et al*. 1998; Morgante *et al* 1998; Gómez *et al.* 2005; Grivet *et al.* 2009; Jaramillo-Correa *et al.* 2010; Grivet *et al.* 2011), as well as with palaeoecological and fossil records (Nahal 1962; Pons 1992), the following putative recolonization scenario can be hypothesised: Aleppo pine occupied a larger distribution in the Tertiary (reaching much higher latitude), and the climatic conditions at the end of the Tertiary and beginning of the Quaternary would have confined its distribution to the southern Europe (Nahal 1962). Probably during the last interglacial before the last glaciation, a very old long-range expansion would have happened, as suggested by molecular works pointing at a separation of African and Iberian gene pools (Jaramillo-Correa *et al*. 2010) as well as of western and eastern gene pools (González-Martínez personal communication) around that time. During the Last Glacial Maximum (25,000-18,000 yr BP) the harsher climatic conditions, particularly in the western Mediterranean (Van Andel

2002), would have translated into bottlenecks (i.e. in Italy; Grivet *et al*. 2011) and the persistence of populations in local glacial refugia (both in the West and in the East). Consecutively to the last glaciation, climatic conditions would have become gradually more favourable towards the western Mediterranean (Van Andel 2002) giving rise to a more recent expansion wave leading to the current distribution of Aleppo pine around 10,000 years ago (Nahal 1962; Schiller *et al.* 1986; Pons 1992). Range expansion would have been accompanied by gene flow, resulting in admixture zones all along the distribution range of the species and a complex population genetic structure.

During the various waves of Aleppo pine range expansions, some rare alleles could have reached very high frequencies because of strong genetic drift occurring in populations located at the front wave of the expansion (Edmonds *et al.* 2004; Klopfstein *et al.* 2006). This is the case for some of the SNPs analysed in the present study (see some examples in Figure S8, Supporting Information) that display a longitudinal cline of allele frequency, translating the potential action of gene surfing or/and the action of natural selection, both processes leaving similar genomic footprints (Excoffier & Ray 2008). Therefore, in the specific case of Aleppo pine, the effect of allele surfing should be taken into account when seeking SNPs involved in local adaptation during long-range colonization.

2.5.2 Footprint of natural selection during range expansion

In the present study, while looking for footprint of selection, we controlled for two confounding effects: (i) we used the relatedness among populations (*Bayenv*) or individuals (*PCAdapt*) to correct for neutral population/individual structure in the data; (ii) we captured the effects of allele surfing by simulating various scenarios of recolonization with different intensities of migration rates, and therefore different admixture intensities among populations during range expansion. Simulations indicated the occurrence of a range expansion which can lead to high false-positive rates depending on the statistical method used (Lotterhos & Whitlock 2014; Lotterhos & Whitlock 2015). Although both the genetic-environment association and PCA-based outlier tests identified some SNPs potentially under selection, our simulations indicated that only SNPs from the former method might be involved in adaptation, while those detected with the latter may result from the confounding effects of demography and drift. Supporting this finding, it has been shown with the use of simulated data, that methods based on latent factors (estimated by considering the statistical model and the data simultaneously) performed less robustly in scenarios of long range colonisation and Isolation by Distance compared to a Bayesian linear approach (Lotheros & Witlocks 2015). Furthermore, it is still unclear to what extent do principal components represent population structure (Galinsky *et al*. 2015).

When looking for footprints of local adaptation in Aleppo pine, we incorporated gene surfing by simulating a relatively recent (i.e. just after the LGM) wave of range expansion from one eastern refugia. Our simulations present some limitations, i.e. they do not take into account old recolonization processes nor multiple refugia, and thus they may overestimate the range of expansion rate in Aleppo pine. Potential demographic scenarios for the Aleppo pine suggests an initial ancient colonization (last interglacial before the last glaciation), followed by later re-expansions (after the last glaciation) from eastern and western refugia towards the western Mediterranean Basin. The relative importance of these events for explaining patterns of genetic variation is difficult to assess, and the genetic signal we see nowadays potentially reflects a mix of old and recent expansion waves. While simulations comparing the distribution of *Bayenv* statistics under 1-Refugia (1R) and 2-Refugia (2R) scenarios reveal a similar reduced variance for neutral markers along the BF axes, they also indicate a greater variance for selected loci in the 2R scenario. This outcome suggests that our 1R simulations would allow a good separation between neutral and selected loci, although not providing the best statistical separation among loci with different strength of selection (Lotterhos & Witlocks 2015). Finally, it may be that the simulated data followed a more severe range expansion than that experienced in Aleppo pine (i.e. m=0.1 would be higher than the real migration rate), thus making our tests conservative.

Since the probability of a mutation establishing itself on the expansion front and spreading with the expanding wave is largely driven by genetic drift, the rate at which beneficial, neutral and deleterious alleles fix on the wave front is relatively similar to each other relative to the core population (Peischl *et al.* 2013). In expanding species, beneficial mutations at the front wave allow rapid adaptation during early stages of expansion, but as deleterious mutations occur at a higher rate than beneficial ones, a decrease of fitness along the expansion axis is expected (Peischl *et al.* 2016). Distinguishing among the different mutations requires fitness data, and assessing the phenotypic effect of mutations in the target species. Although for the time being we cannot link the phenotypic (i.e. fitness related traits) and genomic (i.e. the 7 SNPs found under selection) data, common-garden experiments (provenance trials covering most of the species natural area) indicate that Aleppo pine populations are differently adapted across their distribution, in relation to a geographic pattern that mimics the climatic conditions specific to the western and southern edges of the distributional range (Climent *et al.* 2008). This suggests that Aleppo pine has been able to adapt to new environments in its westernmost and southernmost expansion fronts, despite the expected accumulation of deleterious mutations (the so call expansion load) on the wave front during expansion range (Peischl *et al.* 2013).

Here we specifically focus on the confounding effect of allele surfing in detecting putative SNPs under selection in the specific case of a forest tree species that presents long-range expansion and long generation time. Other processes, outside the scope of this study, can of course mimic the molecular footprint left by selection at specific loci (see Hobban *et al*. 2016), such as background selection or hybridization and introgression from related species (when *P. halepensis* occurs with *P. brutia* they form natural hybrids; Richardson 1998; Quézel 2000; Schiller 2000).

2.5.3 Outlier SNPs

The seven SNPs significant in the Bayesian environmental association analysis were correlated with bioclimatic variables related to temperature and precipitation ( ), known to potentially affect growth and reproduction in Aleppo pine (Baquedano *et al.* 2008; Climent *et al.* 2008; Voltas *et al.* 2008 and 2015; Girard *et al.* 2011 and 2012; Santos-del-Blanco *et al*. 2013; De Luis *et al.* 2013; Hernández-Serrano *et al.* 2014).

Five out of the seven SNPs putatively under selection are in genes found to be relevant for adaptation in other tree species, in response to a wide range of selective pressures (Table 2.3). Interestingly, SNP151, which frequency is associated with the precipitation of the driest month (BIO14), is located within a gene which expression was induced by drought stress in two other Mediterranean pines. This locus presents some homology with a peroxisomal membrane protein that could play a role in the establishment of a ROS scavenging mechanism in response to drought stress (Cruz de Carvalho 2008).

SNPs	Sequence	Annotation	E-value	Evidence in other tree species	Drivers
SNP24	seq-0_12329_02-271	polyadenylate-binding protein 3-like	6.05E-12		
SNP54	seq-8115-2383	ABC transporter G family member 11-like	2.04E-97	*Pinus massonanai*[a] ; *Populus pruinosa*[b]	phosphorus, salinity
SNP 151	seq-0_8992_01-119	peroxisomal membrane protein 11D-like	4.38E-13	*Pinus pinaster*[c]; *Pinus pinea*[c]	water stress
SNP 183	seq-35072-1606	"Lig v 1-like" endoglucanase 6-like	8.84E-08		
SNP 186	seq-0_6924_02-89	polygalacturonase inhibitor 1-like	2.68E-34	*Populus tremula*[d]	nitrogen
SNP 374	seq-11269-769	alpha-galactosidase-like	1.02E-29	*Pinus radiata*[e]	growth, wood formation
SNP 375	seq-51080-670	RING-H2 finger protein ATL48-like	5.50E-13	*Picea glauca*[f]; *Picea engelmannii*[f]	herbivory stress
SNP 378	seq-2_3941_01-381	B3 domain-containing protein At3g19184-like	1.62E-66		

Table 2.3: Annotations of the loci for which SNPs were found under selection. [a]Fan *et al*. (2014); [b]Zhang *et al*. (2014); [c]Perdiguero *et al*. (2013); [d]Wei *et al*. (2013); [e]Li *et al*. (2009); [f]Verne *et al*. (2011)

2.6 Conclusions

In this study, we propose a framework to infer selection in instances of long range colonisation taking into account important sources of false positives: shared history and gene flow, as well as gene surfing. Although the first source of false positive is widely acknowledged and population structure is included routinely in different approaches to detect gene polymorphisms under selection, the impact of gene surfing during long range colonisation has been studied theoretically but seldom empirically, and even more rarely in a scenario of adaptation. Our results comparing observed and simulated data with different migration rates, indicate that the lower the migration rate, the closest the allele frequencies mirrored those of our real data, leading to the identification of 7 SNPs potentially under selection with the Bayesian correlative approach. Most of the SNP-environment associations are related to drought, an environmental driver of crucial importance for the adaptation of Aleppo pine that inhabits the Mediterranean Basin.

Acknowledgements

We would like to thank E. Ballesteros Burgos, D. Barba Egido, A. Blanco Asenjo, F. del Caño Prieto, M.R. Chambel, D. Lafuente Laguna, Z. Lorenzo, and Sergio San Segundo González for sample collecting, C. García-Barriga and Z. Lorenzo for DNA extraction and genotyping, as well as D. Sánchez De Ron and J. Martínez-Fernández for GIS insights, and M.A. Beaumont for insightful advice and discussions during the analysis stage. This project has been supported by the Spanish National Research Plan (AdapCon CGL2011-30182-C02-01, and RTA2013-00048-C03-01), as well as from the ERA-Net BiodivERsA (TipTree BiodivERsA 2012-15). R.R.D. was supported by a FPI from the Spanish Ministry of Economy and Competitiveness; R.S.T was funded through an Engineering and Physical Sciences Research Council (EPSRC) doctoral training grant (DTG) through the Bristol Centre for Complexity Science; M.J.S.V. was supported by a FPU from the Spanish Ministry of Education, Culture and Sport; D.G. was supported by a 'Ramón y Cajal' fellowship from the Spanish Ministry of Science and Innovation.

2.7 References

Alberto FJ, Derory J, Boury C *et al.* (2013) Imprints of natural selection along environmental gradients in phenology-related genes of *Quercus petraea*. *Genetics*, **195**, 495–512.

Baquedano FJ, Valladares F, Castillo FJ (2008) Phenotypic plasticity blurs ecotypic divergence in the response of *Quercus coccifera* and *Pinus halepensis* to water stress. *European Journal of Forest Research*, **127**, 495–506.

Begun DJ, Holloway AK, Stevens K *et al.* (2007) Population Genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology*, **5**, e310.

Blair LM, Granka JM, Feldman MW (2014) On the stability of the Bayenv method in assessing human SNP-environment associations. *Human Genomics*, **8**, 1.

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences*, **101**, 15255–15260.

Bucci G, Anzidei M, Madaghiele A, Vendramin GG (1998) Detection of haplotypic variation and natural hybridization in *halepensis* -complex pine species using chloroplast simple sequence repeat (SSR) markers. *Molecular Ecology*, **7**, 1633–1643.

Budde KB, Heuertz M, Hernández-Serrano A *et al.* (2014) In situ genetic association for serotiny, a fire-related trait, in Mediterranean maritime pine (*Pinus pinaster*). *New Phytologist*, **201**, 230–241.

Climent J, Prada MA, Calama R *et al.* (2008) To grow or to seed: ecotypic variation in reproductive allocation and cone production by young female Aleppo pine (*Pinus halepensis,* Pinaceae). *American Journal of Botany*, **95**, 833–842.

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

Cruz de Carvalho MH (2008) Drought stress and reactive oxygen species. *Plant Signaling & Behavior*, **3**, 156–165.

Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge, United Kingdom.

De Luis M, Čufar K, Di Filippo A *et al.* (2013) Plasticity in dendroclimatic response across the distribution range of Aleppo pine (*Pinus halepensis*). *PLoS ONE*, **8**, e83550.

Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB (2016) Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. *Molecular Biology and Evolution*, **33**, 1082–1093.

Earl DA, VonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Edmonds CA, Lillie AS, Cavalli-Sforza LL (2004) Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences of USA*, **101**, 975–979.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Foll M, Petit RJ (2009) Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 481–501.

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, **23**, 347–351.

Fady B, Conord C (2010) Macroecological patterns of species and genetic diversity in vascular plants of the Mediterranean basin. *Diversity and Distributions*, **16**, 53–64.

Fady B, Semerci H, Vendramin GG (2003) EUFORGEN Technical guidelines for genetic conservation and use for Aleppo pine (*Pinus halepensis*) and Brutia pine (*Pinus brutia*) p.p 6. *International Plant Genetic Resources Institute*, Rome, Italy.

François O, Martins H, Caye K, Schoville SD (2016) Controlling false discoveries in genome scans for selection. *Molecular Ecology*, **25**, 454–469.

Frichot E, Schoville SD, de Villemereuil P, Gaggiotti OE, François O (2015) Detecting adaptive evolution based on association with ecological gradients: Orientation matters! *Heredity*, **115**, 22–28.

Galinsky KJ, Bhatia G, Loh P-R *et al.* (2016) Fast Principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *The American Journal of Human Genetics*, **98**, 456–472.

Gautier M (2015) Genome-wide scan for adaptive divergence and association with oopulation-specific covariates. *Genetics*, **201**, 1555–1579.

Girard F, Vennetier M, Ouarmim S, Caraglio Y, Misson L (2011) Polycyclism, a fundamental tree growth process, decline with recent climate change: the example of *Pinus halepensis* Mill. in Mediterranean France. *Trees*, **25**, 311–322.

Girard F, Vennetier M, Guibal F *et al.* (2012) *Pinus halepensis* Mill. crown development and fruiting declined with repeated drought in Mediterranean France. *European Journal of Forest Research*, **131**, 919–931.

Gómez A, Vendramin GG, González-Martínez SC, Alía R (2005) Genetic diversity and differentiation of two Mediterranean pines (*Pinus halepensis* Mill. and *Pinus pinaster* Ait.) along a latitudinal cline using chloroplast microsatellite markers. *Diversity and Distributions*, **11**, 257–263.

González-Martínez SC, Robledo-Arnuncio JJ, Collada C *et al.* (2004) Cross-amplification and sequence variation of microsatellite loci in Eurasian hard pines. *Theoretical and Applied Genetics*, **109**, 103–111.

Gracia E, Botella F, Anadon JD *et al.* (2013) Surfing in tortoises? Empirical signs of genetic structuring owing to range expansion. *Biology Letters*, **9,** 1091–2012.

Gralka M, Stiewe F, Farrell F *et al.* (2016) Allele surfing promotes microbial adaptation from standing variation. *Ecology Letters*, **19**, 889–898.

Grivet D, Sebastiani F, Alia R *et al.* (2011) Molecular footprints of local adaptation in two Mediterranean conifers. *Molecular Biology and Evolution*, **28**, 101–116.

Grivet D, Sebastiani F, González-Martínez SC, Vendramin GG (2009) Patterns of polymorphism resulting from long-range colonization in the Mediterranean conifer Aleppo pine. *New Phytologist*, **184**, 1016–1028.

Guevara MA, Chagne D, Almeida MH *et al.* (2005) Isolation and characterization of nuclear microsatellite loci in *Pinus pinaster* Ait. *Molecular Ecology Notes*, **5**, 57–59.

Gunther T, Coop G (2013) Robust Identification of Local Adaptation from Allele Frequencies. *Genetics*, **195**, 205–220.

Hallatschek O, Nelson DR (2008) Gene surfing in expanding populations. *Theoretical Population Biology*, **73**, 158–170.

Hancock AM, Witonsky DB, Alkorta-Aranburu G *et al.* (2011) Adaptations to climate-mediated selective pressures in humans (MW Nachman, Ed,). *PLoS Genetics*, **7**, e1001375.

Hardy OJ, Vekemans X (2002) Spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Hernandez-Serrano A, Verdu M, Santos-del-Blanco L *et al.* (2014) Heritability and quantitative genetic divergence of serotiny, a fire-persistence plant trait. *Annals of Botany*, **114**, 571–577.

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Hoban S, Kelley JL, Lotterhos KE *et al.* (2016) Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist*, **188**, 379–397.

Hofer T, Ray N, Wegmann D, Excoffier L (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Annals of Human Genetics*, **73**, 95–108.

Jackson DA (1993) Stopping rules in principal components analyses: a comparison of heuristical and statistical approaches. *Ecology,* **74**, 2204–2214.

Jaramillo-Correa JP, Grivet D, Terrab A *et al.* (2010) The Strait of Gibraltar as a major biogeographic barrier in Mediterranean conifers: A comparative phylogeographic survey. *Molecular Ecology*, **19**, 5452–5468.

Jaramillo-Correa JP, Prunier J, Vázquez-Lobo A, Keller SR, Moreno-Letelier A (2015) Molecular signatures of adaptation and selection in forest trees. *Advances in Botanical Research*, **74**. 265–306.

Jezkova T, Jaeger JR, Oláh-Hemmings V *et al.* (2016) Range and niche shifts in response to past climate change in the desert horned lizard *Phrynosoma platyrhinos*. *Ecography*, **39**, 437–448.

Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, **23**, 482–490.

De La Torre A, Ingvarsson PK, Aitken SN (2015) Genetic architecture and genomic patterns of gene flow between hybridizing species of Picea. *Heredity*, **115**, 153–164.

Fan F, Cui B, Zhang T *et al.* (2014) The temporal transcriptomic response of *Pinus massoniana* Seedlings to Phosphorus Deficiency. *PLoS ONE*, **9**, e105068.

Le Corre V, Kremer A (1998) Cumulative effects of founding events during colonisation on genetic diversity and differentiation in an island and stepping-stone model. *Journal of Evolutionary Biology,* **11**, 495–512

Li X, Wu HX, Dillon SK, Southerton SG (2009) Generation and analysis of expressed sequence tags from six developing xylem libraries in *Pinus radiata*. *BMC Genomics*, **10**, 41.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of F ST outlier tests. *Molecular Ecology*, **23**, 2178–2192.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology,* **24**, 1031–1046.

Luu K, Bazin E, Blum MGB (2017) pcadapt : an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, **17**, 67–77.

Mariette S, Chagné D, Decroocq S *et al.* (2001) Microsatellite markers for *Pinus pinaster* Ait. *Annals of Forest Science*, **58**, 203–206.

Michele Morgante, Nicoletta Felice GGV (1998) Analysis of hypervariable chloroplast microsatellites in *Pinus halepensis* reveals a dramatic genetic bottleneck. pp. 407–412 *Molecular Tools for Screening Biodiversity*, Springer Press, The Netherlands.

De Mita S, Thuillet A-C, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

Nahal I (1962) Le pin d'Alep (*Pinus halepensis* Mill.). Etude taxonomique, phytogéographique, écologique et sylvicole. *Annales de l'Ecole Nationale Des Eaux et Forêts et de la Station de Recherches et Expériences Forestières*, **4**, 7–207

Nei M (1972) Genetic distance between populations. *The American Naturalist*, **106**, 283–292.

Nielsen R, Tarpy DR, Reeve HK (2003) Estimating effective paternity number in social insects and the effective number of alleles in a population. *Molecular Ecology*, **12**, 3157–3164.

North B V, Curtis D, Sham PC (2002) A note on the calculation of empirical P values from Monte Carlo procedures. *American Journal of Human Genetics*, **71**, 439–41.

Pannell JR, Fields PD (2014) Evolution in subdivided plant populations: concepts, recent advances and future directions. *New Phytologist*, **201**, 417–432.

Peischl S, Dupanloup I, Bosshard L, Excoffier L (2016) Genetic surfing in human populations: from genes to genomes. *Current Opinion in Genetics & Development*, **41**, 53–61.

Peischl S, Dupanloup I, Kirkpatrick M, Excoffier L (2013) On the accumulation of deleterious mutations during range expansions. *Molecular Ecology*, **22**, 5972–5982.

Perdiguero P, Barbero M del C, Cervera MT, Collada C, Soto Á (2013) Molecular response to water stress in two contrasting Mediterranean pines (*Pinus pinaster* and *Pinus pinea*). *Plant Physiology and Biochemistry*, **67**, 199–208.

Peter BM, Slatkin M (2013) Detecting range expansions from genetic data. *Evolution*, **67**, 3274–3289.

Pierce AA, Zalucki MP, Bangura M *et al.* (2014) Serial founder effects and genetic differentiation during worldwide range expansion of monarch butterflies. *Proceedings of the Royal Society B: Biological Sciences*, **281**, 20142230.

Pinosio S, González-Martínez SC, Bagnoli F *et al.* (2014) First insights into the transcriptome and development of new genomic tools of a widespread circum-Mediterranean tree species, *Pinus halepensis* Mill. *Molecular Ecology Resources*, **14**, 846–56.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–59.

Pons A (1992) Les enseignements des données historiques concernant le pin d'Alep. *Forêt Méditerranéenne, 3,155-157.*

Pujolar JM, Jacobsen MW, Als TD *et al.* (2014) Genome-wide single-generation signatures of local selection in the panmictic European eel. *Molecular Ecology*, **23**, 2514–2528.

Pyhäjärvi T, Hufford MB, Mezmouk S, Ross-Ibarra J (2013) Complex patterns of local adaptation in teosinte. *Genome Biology and Evolution*, **5**, 1594–609.

Ray N, Currat M, Foll M, Excoffier L (2010) SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. *Bioinformatics,* **26**, 2993–4.

Raymond M, Rousset F (1995) GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R (2015) A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, **24**, 4348–4370.

Ring KH, (2015) PyBayenv: A framework for interpreting, testing and optimizing Bayenv analyses. https://www.duo.uio.no/handle/10852/44752

Rousset F, (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Santos-del-Blanco L, Bonser SP, Valladares F, Chambel MR, Climent J (2013) Plasticity in reproduction and growth among 52 range-wide populations of a Mediterranean conifer: adaptive responses to environmental stress. *Journal of Evolutionary Biology*, **26**, 1912–1924.

Santos-Del-Blanco L, Climent J, González-Martínez SC, Pannell JR (2012) Genetic differentiation for size at first reproduction through male versus female functions in the widespread Mediterranean tree *Pinus pinaster*. *Annals of Botany*, **110**, 1449–1460.

Serra-Varela MJ, Alía R, Ruíz Daniels R, Zimmermann NE, Gonzalo-Jiménez J, & Grivet D (2017) Assessing vulnerability of two Mediterranean conifers to support genetic conservation management. *Diversity and Distributions*, **5**, 507–516

Schiller G, Conkle MT (1986) Local differentiation among Mediterranean populations of Aleppo pine in their isoenzymes. *Silvae Genetica*, **35**, 11–19.

Slatkin M, Excoffier L (2012) Serial founder effects during range expansion: a spatial analog of genetic drift. *Genetics*, **191**, 171–81.

Steinitz O, Troupin D, Vendramin GG, Nathan R (2011) Genetic evidence for a Janzen-Connell recruitment pattern in reproductive offspring of *Pinus halepensis* trees. *Molecular Ecology*, **20**, 4152–64.

Stephan W (2016) Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular Ecology*, **25**, 79–88.

Swaegers J, Mergeay J, Van Geystelen A et al. (2015) Neutral and adaptive genomic signatures of rapid poleward range expansion. *Molecular Ecology*, **24**, 6163–6176.

Travis JMJ, Munkenmuller T, Burton OJ (2010) Mutation surfing and the evolution of dispersal during range expansions. *Journal of Evolutionary Biology*, **23**, 2656–2667.

Van Andel TH (2002) The climate and landscape of middle part of Weichselian glaciation in Europe: the stage 3 project. *Quaternary Research*, **57**, 2–8.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) micro-checker: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes*, **4**, 535–538.

Verne S, Jaquish B, White R, Ritland C, Ritland K (2011) Global transcriptome analysis of constitutive resistance to the white pine weevil in spruce. *Genome Biology and Evolution*, **3**, 851–867.

Voltas J, Lucabaugh D, Chambel MR, Ferrio JP (2015) Intraspecific variation in the use of water sources by the circum-Mediterranean conifer *Pinus halepensis*. *New Phytologist*, **208**, 1031–1041.

Voltas J, Chambel MR, Prada MA, Ferrio JP (2008) Climate-related variability in carbon and oxygen stable isotopes among populations of Aleppo pine grown in common-garden tests. *Trees*, **22**, 759–769.

Wegmann D, Currat M, Excoffier L (2006) Molecular diversity after a range expansion in heterogeneous environments. *Genetics*, **174**, 2009–20.

Wei H, Yordanov YS, Georgieva T, Li X, Busov V (2013) Nitrogen deprivation promotes Populus root growth through global transcriptome reprogramming and activation of hierarchical genetic networks. *New Phytologist*, **200**, 483–497.

White TA, Perkins SE, Heckel G, Searle JB (2013) Adaptive evolution during an ongoing range expansion: The invasive bank vole (*Myodes glareolus*) in Ireland. *Molecular Ecology*, **22**, 2971–2985.

Willyard A, Ann W, Syring J *et al.* (2007) Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for Pinus. *Molecular Biology and Evolution*, **24**, 90–101.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114–138.

Zhang J, Jiang D, Liu B *et al.* (2014) Transcriptome dynamics of a desert poplar (*Populus pruinosa*) in response to continuous salinity stress. *Plant Cell Reports*, **33**, 1565–1579.

Appendix A: Supplementary Information

"Now, here, you see, it takes all the running you can do to keep in the same place" Lewis Carroll through the looking glass, contextualised in evolutionary biology by Van Valen (1973)"

Credit: Illustration by Sir John Tenniel from Lewis Carroll's Through the Looking-Glass, 1871

Chapter 3 **(Part I) Looking for local adaptation: a case study of the Aleppo pine (*Pinus halepensis*) in the Mediterranean Basin.**

Rose Ruiz Daniels[1], Richard S. Taylor[2], Giovanni G. Vendramin[3], Santiago C. González-Martínez[4], Delphine Grivet[1,5] & Mark A. Beaumont[2]

[1] Department of Forest Ecology and Genetics, Forest Research Centre, INIA, Carretera A Coruña km 7.5, 28040 Madrid, Spain.

[2] School of Biological Sciences, University of Bristol, Wills Memorial Building, Queen's Road, Bristol BS8 1RJ, UK

[3] Institute of Biosciences and Bioresources, National Research Council, I-50019 Sesto Fiorentino, Florence, Italy [4]
BIOGECO, INRA, Univ. Bordeaux, 33610 Cestas, France

[5] Sustainable Forest Management Research Institute, INIA, University of Valladolid, 34004 Palencia, Spain

3.1 Abstract

Finding outlier loci responsible for local adaptation is challenging, and is best approached by suitable sampling design and rigorous method selection. In this study, we aim to detect outlier loci (single nucleotide polymorphisms) by using a paired sampling technique that aims to maximise environmental differences between populations, while minimising differences in evolutionary history. We then apply three different statistical methodologies - two Bayesian outlier methods and one latent factor principal component method - to identify outliers in common amongst these methods, some of which are associated with environmental conditions. We review the performance of these methods with the aim of facilitating future work on outlier detection using empirical data sets.

3.2 Introduction

When organisms expand their geographical range, they are inevitably faced with new selective pressures that result in natural selection acting on phenotypic variation and consequently genotypic variation. This causes an allelic shift at the local level that maximises the fitness of that organism and leads to local adaptation: that is, when an individual's fitness is higher than average in its local environment compared with an individual from elsewhere. Finding the genetic basis of local adaptation is a topic of crucial importance in evolutionary biology as it allows for the study of the mechanisms of natural selection, which happen on a local scale (Savolainen *et al.* 2013; Hoban *et al.* 2016). Two main approaches are widely used to identify genomic regions under selection. One approach identifies correlations between allele frequencies and environmental variables, while the other look for loci with unusually high levels of differentiation (Narum & Hess 2011; De Mita *et al.* 2013; Lotterhos & Whitlock 2014, 2015; Frichot *et al.* 2015; Rellstab *et al.* 2015). Since the available methods for detecting outlier loci are based on underlying hypotheses that take into account different signals left by natural selection at the molecular level, there are inevitably discrepancies in the set of loci identified as under selection (Gunther & Coop 2013; Gautier 2015; Hoban *et al.* 2016; Luu *et al.* 2017). Deciding what method is most suitable to detect markers under selection is challenging, as is choosing the significance threshold for a marker to be under selection.

In this study we aim to identify single nucleotide polymorphism (SNP) loci that may be under local selection in Aleppo pine (*Pinus halepensis Mill.*), by combining various recently developed statistical methods with a paired sampling design. Aleppo pine has a circum-Mediterranean distribution that is both geographically fragmented and vast, spanning 3.5 million hectares. This pine lives in a wide range of environmental conditions (both in terms of climates and substrates), which reflects a high degree of adaptability over its natural distribution under dry Mediterranean climates, as also illustrated by its capacity to become invasive in other Mediterranean regions in the southern hemisphere (Rouget *et al.* 2001; Lavi *et al.* 2005). Its environmental versatility lies both in its high plasticity (Santos-del-Blanco *et* *al.* 2012; de Luis *et al.* 2013) as well as potentially to its molecular adaptation (Grivet *et al.* 2011; Ruiz Daniels *et al.* under review) to the Mediterranean region. In Europe, fossil records and genetic data indicate that Aleppo pines demographic history would be characterized by an initial ancient colonization (last interglacial before the last glaciation), followed by later re-expansions (after the last glaciation) from eastern and western refugia towards the western Mediterranean Basin (Ruiz Daniels *et al.* under review). This left Aleppo pine with a complex genetic pattern, with two main genetic clusters: an eastern/southern cluster that is both older and more genetically diverse than a western cluster (Grivet *et al.* 2009; Ruiz Daniels *et al.* under review), both being connected by population admixture (Serra-Varela *et al.* 2017; Ruiz Daniels *et al*. under review).

To detect local adaptation in Aleppo pine, we explore whether alleles have reached fixation in a particular local population by using a specific sampling design: we select paired/triplet populations in sites of contrasting environmental conditions, by sampling populations in different altitudes in order to reflect differences in precipitation and temperature, two climatic variables known to affect this conifer distribution/adaptation (Atzmon *et al.* 2004; Sathyan *et al.* 2005; Voltas *et al.* 2008; Serra-Valera *et al*. 2017 ; Ruiz Daniels *et al*. Under review). To gain insights on the action of selection on putative outliers, we compare and contrast three different statistical methods that correct for population structure, an attribute that allows these methods to outperform methods that do not (De Mita *et al.* 2013; Blair *et al.* 2014; Lotterhos & Whitlock 2015; Rellstab *et al.* 2015). Two Bayesian outlier approaches, Bayenv2 (Coop *et al.* 2010; Gunther & Coop 2013) and Baypass (Gautier 2015) are used. Both methods infer a scaled genetic covariance matrix which can be used to obtain the so-called *XtX* statistics. This statistic can be viewed as a SNP specific F_{ST} corrected for the scaled covariance of population allele frequencies which takes non-independence of population frequencies into account (Gunther & Coop 2013). Bayesian approaches have been shown to perform well in situations where there is non-uniform population structure (De Mita *et al.* 2013; Lotterhos & Whitlock 2014). The Baypass method is based on a very similar model to Bayenv but is suggested to be an improvement in two aspects: i) a higher accuracy in the estimation of the baseline (ancestral) allele frequency through the use of a hierarchical Bayesian model; ii) the use of simulated pseudo observed data sets (PODs) from the posterior predictive distribution in order to calibrate the *XtX* statistics and determine a cut-off point of potentially significant outliers by applying a false discovery rate threshold. The two methodologies are used in conjunction, with the intention of using the significance threshold for the *XtX* in Baypass to help defining a cut-off point for the Bayenv outputs, for which no standardised method exists either for the *XtX* statistics or for the Bayes factors.

The third approach we adopt is a method based on principal component analysis (PCA), implemented in the software PCAdapt (Luu *et al.* 2017). In this case the genotype matrix is decomposed into its principal components and outlier SNPs are detected as those with unusually large loadings on one or more components. Indeed, the commonly used F_{ST} index of genetic variation can also be view as the proportion of variance explained by principal components (McVean 2009; Duforet-Frebourg *et al.* 2016). The differences in PCA loadings contributed by different loci reflect the differences in allele frequencies between populations triggered by positive selection. This approach focuses on an individual basis, instead of populations therefore eliminating the need of grouping individuals into populations, as populations are a subjective category, and inadequate choice of these can lead to signals of selection being missed (Yang *et al.* 2012). This approach has been shown to work well for outlier detection with simulated data sets under different demographic scenarios (Lotterhos & Whitlock 2015).

In the present study, our objectives are as follows: 1) To infer instances of local adaptation in Aleppo pine using pairs of natural populations from contrasted environments; 2) To explore what method or combination of these work best to infer local adaptation in this conifer adapted to the Mediterranean climate.

3.3 Materials and Methods

3.3.1 Sampling

3.3.1.1 Plant material and environmental data

Adult trees from three countries were sampled in pairs or triplets, in sites characterized by contrasted climatic conditions along altitudinal gradients: three French populations (356 individuals), two Italian populations (50 individuals) and two Spanish populations (70 individuals). In total 476 Individuals were collected from 7 populations (see Table S1 for more details). The aim of this sampling design, was to maximise the differentiation between environments due to natural selection, while minimising the differences in evolutionary history (gene flow should reduce differentiation of loci not under selection) (Hoban *et al.* 2016). This approach has been shown to be effective in comparative studies using simulated data (Lotterhos & Whitlock 2015), as well as when inferring allelic outliers in empirical studies (e.g. Nadeau *et al.* 2016).

For each population 19 bioclimatic variables for the period 1950-2000 were taken from WORLDCLIM (Hijmans *et al.* 2005), to explore SNP-environment associations (see Table S2). The accuracy of this coarse grain climatic dataset was checked with local climatic datasets for the environmental variables based on precipitation and temperature. Furthermore altitude was added as a potential driver of selection, bringing the total of environmental variables up to 20 (see Table S3 for more details).

3.3.1.2 DNA extraction, SNP genotyping and gene annotation

For every individual genotyped tree, 50 mg of needles were dried with silica gel and grinded in QIAwell (Qiagen, Venlo, the Netherlands), plate homogenized with mixer mill MM300

(RETSH, Haan, Germany) under liquid nitrogen. DNA extractions were carried out with the kit Invisorb DNA plant HTS 96 kit (Invitek, Hayward, CA) following the manual instructions. DNA was quantified with Nanodrop 10000 (Thermo Fisher Scientific, Wilmington, DE).

Aleppo pine populations were successfully genotyped at 294 SNPs (conversion rate of 76.56 %) using a 384-plex SNP assay with Illumina VeraCode technology as described in Pinosio *et al.* (2014). SNP markers in this assay originated from two sources: (i) 144 SNPs were taken from 117 polymorphic amplicons obtained from direct sequencing of haploid (megagametophytes) tissue sampled from the full distribution of the species (www.evoltree.eu; CRIEC initiative and unpublished results) these were taken from candidate genes for selection found in the Loblolly pine (*Pinus taeda*); (ii) 240 SNPs were taken from 28,236 SNPs discovered through alignment of the transcriptome of two trees with contrasted phenotypes.

Loci for which SNPs were found putatively under selection were annotated from homology with tree species using BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi), by using a megablast to identify similar sequences.

3.3.2 Statistical analysis

3.3.2.1 Population structure

To confirm that the population pairs/triplets were sampled from areas of similar genetic origins, population genetic structure was estimated in two ways. First, pairwise F_{ST} was computed among all seven populations with Genepop v4.1 (Rousset 2008) . Second, population genetic structure was inferred using the Bayesian clustering method STRUCTURE (Pritchard *et al.* 2000), with the following parameters: number of clusters (*K*) set to 10; number of iterations set to 10; number of steps set to 100,000 with a burn in period of 10,000 to minimize the effect of the starting configuration, and with an ancestry model of admixture.

3.3.2.2 Looking for outliers using Bayesian approaches

Bayenv2

In order to identify selection acting on SNPs, the method described in Coop *et al.* (2010) was used, this estimates the population differentiation statistic *XtX* as well as evaluates the association of 20 ecological variables (see Material and Method Section) with genetic marker differentiation (i.e. SNPs in the present study). This method is implemented in the Bayenv 2.0 package (Gunther & Coop 2013), which estimates the empirical pattern of covariance in allele frequencies between populations, and uses this as the null model to test each individual SNP for selection. This methodology accounts for the neutral correlation of allele frequencies across populations, a critical consideration when there is evidence of shared history and gene flow among populations, as in Aleppo pine. To minimize the stochasticity in estimating the null model, three covariance matrices were produced, from 100.000 million iterations each, and the mean of these matrices was used. In order to test that this covariance matrix represented well the true variance of allele frequencies across populations it was compared to a pairwise F_{ST} matrix using a Mantel test in R (R development core team) with 1000 permutations.

This covariance matrix, by taking into account correlations in population structure, was then used as the null model in two analyses aiming at detecting SNPs under selection: (i) detecting outliers through differences in overall genetic differentiation as in classic outlier tests, by estimating the *XtX* averages across multiple samples from the MCMC. Outlier *XtX* values from the main *XtX* distribution were considered as potentially under selection. Finally, these values were compared to the ones estimated with Baypass in order to ascertain significance (ii) Finding SNPs that show a significant correlation between an environmental variable and allele frequencies. This was carried out by comparing a model which allows for a linear relationship between allele frequency (i.e. from the 294 SNPs) and environmental variable (i.e. the 20 environmental variables described in Material and Methods) with a null model in this case using the covariance matrix, using Bayes factors. This approach was repeated three times for each SNP-environmental variable tested for association, in order to account for instability between independent runs (Blair *et al.* 2014). The mean of the three runs was then used to infer the final Bayes factors (BF). The Kass & Raftery (1995) criteria was then used to determine the probability of these being under selection, with 2lnK values above 6 (BF>20) classified as strong.

Baypass

Baypass (Gautier 2015) was run using the core model, and R (R Core Team 2013) was used to analyse and visualise outputs. First, from the SNP data the correlation matrix $\widehat{\Omega}$ was computed and estimates of *XtX* differentiation were produced. The correlation matrix was then visualised as a hierarchical cluster tree in order to be compared to a neighbour joining tree of pairwise F_{ST} done in genepop (V4.1)(Rousset 2008) distances between populations where computed using the r package APE (Paradis *et al.* 2004). Second, a pseudo observed dataset (POD) sample was constructed using the *simulate_baypass* function in R with 1,000 SNPs. This was done by estimating the posterior mean of the hyperparameters a_pi and b_pi, which specify the beta priors for Pi, the baseline (ancestral) frequency at each locus and then simulating samples with these parameter values. The POD data was then analysed in the same way as the real data to produce values of *XtX*. The quantiles of these empirical *XtX* distributions were used to calibrate the *XtX* observed for each locus in the original data, and the 99% quantile of the *XtX* distribution from the POD analysis provided a 1% threshold *XtX* value (the default cut-off stated by the program designers), which was then used as a decision criterion to discriminate between outlier and neutral SNPs.

Baypass was also used to detect outlier SNPs that show a significant correlation between an environmental variable and allele frequencies. This was carried out by comparing a model which allows for a linear relationship between allele frequency (i.e. from the 294 SNPs) and environmental variable (i.e. the 20 environmental variables described in Material and Method) with a null model using the correlation matrix. For this purpose, Baypass was run making use of an importance sampling estimator (IS) model inspired by Coop *et al.* (2010). The empirical Bayesian p-value (eBPis) was calculated for each SNP environmental relationship, allowing evaluating the support in favour of a non-null regression coefficient when the eBPis is above 3 (Gautier 2015). When using this feature in Bayenv2 it is known that there is high variability between runs, it is advisable to run analyses multiple times and average the Bayes factors (Blair 2013). This is not the case with Baypass where no such variability exists.

In the two Bayesian analysis SNPs with a minimum allele frequency (MAF) below 5% were not removed as in the PCA based method, as it is possible that some of these are under selection, and the programs do not recommend their removal.

3.3.2.3 Looking for outliers using PCA

First, the number of principal components (K) was established by running the PCA analysis in PCAdapt (Luu *et al.* 2017) with a large enough number of components (10 in this case). K was chosen by performing a scree plot, which gives the percentage of variance explained by each PC in decreasing order. The choice of K was also corroborated by plotting individuals on the first two components (called a score plot by PCAdapt) to see if the level of clustering is consistent with the chosen value for K. To account for missing data, the correlation matrix between individuals was computed using only the markers available for each individual. The Mahalanobis statistic was used to identify outliers, by measuring how distant a data point is from the multivariate space's centroid (overall mean), considering the covariance structure of all the data points in the sample. By default, the program removed the alleles displaying MAF< 5%.

Both Manhattan and QQ-plots were used to visualise the distribution of the data. The presence of outliers was confirmed by applying a false discovery rate (FDR), defined as the percentage of false positives among the list of candidate SNPs. FDR was set at 10% as recommended by the PCAdapt authors (Luu *et al.* 2017) using the R package *qvalue* (R Core Team 2013), which transforms the p-values into q-values.

3.4 Results

3.4.1 Population structure

Pairwise F_{ST} indicates that, as expected from previous studies on this species (Serra-Valera *et al.* 2017; Ruiz Daniels *et al*. in review), geographically close populations (i.e. pair/triplet populations) display little genetic differentiation, while populations from different countries are differentiated, with France and Spain being the least differentiated, while Italy is very differentiated from Spain and France (Table S4). These results are confirmed with the Bayesian clustering that revealed that Aleppo pine populations show evidence of structure across the three groups of populations sampled (**Figure 3.1**). A K=3 was chosen as being the most appropriate with the results of the Bayesian clustering, in agreement with that of the PCA used when running PCAdapt (**Figure 3.2**).

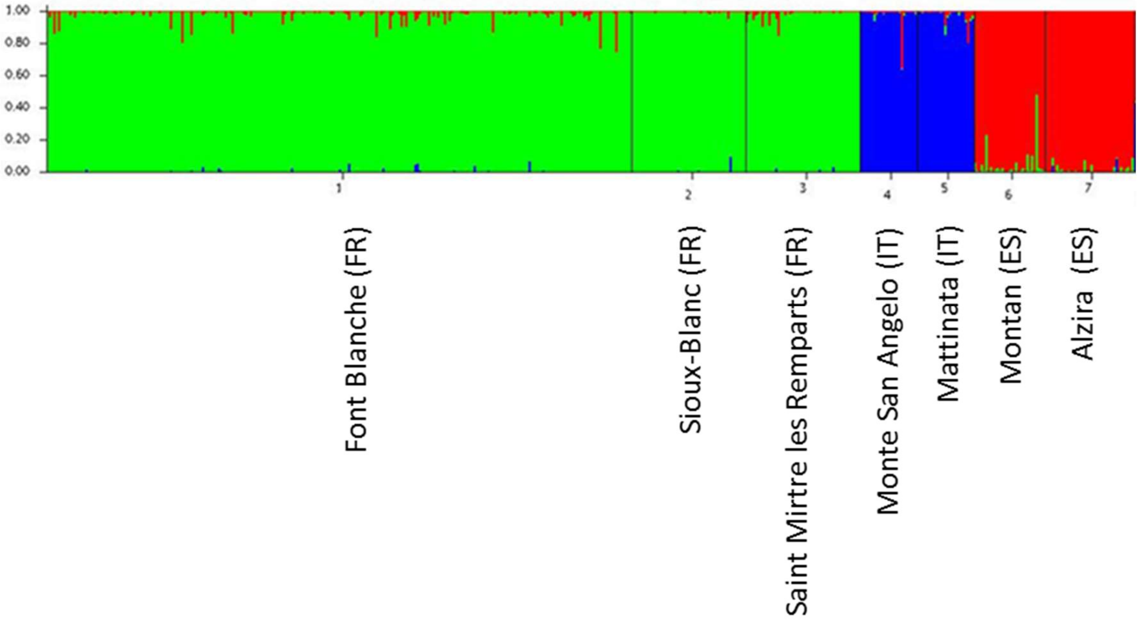


Figure 3.1: Results on the Bayesian clustering performed in STRUCTURE, for the seven populations.

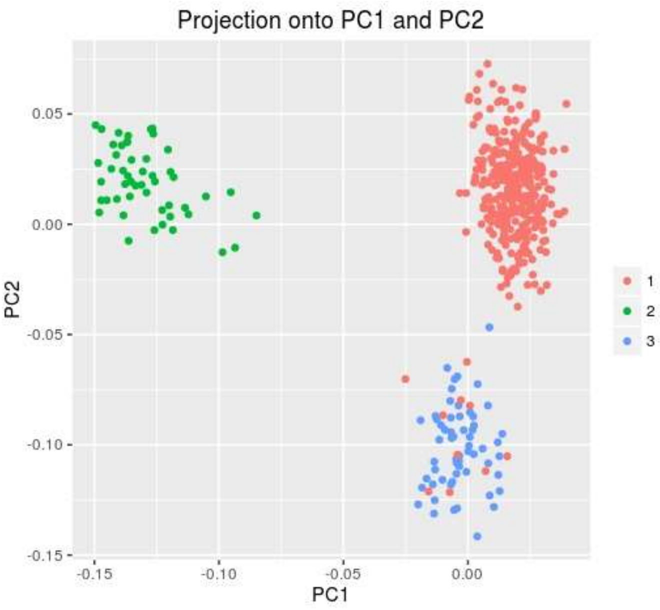


Figure 3.2: Score plot of SNP data where each provenance is colour coded (1 France, 2 Italy and 3 Spain). This plot displays the projections of the individuals onto the specified principal components.

3.4.2 Looking for outliers using Bayesian approaches

Using Bayenv2

Confidence in the mean covariance matrix (resulting from three independent covariance matrices) representing the true variance of allele frequencies across populations was assessed by the similarity between its heatmap and that of the pairwise population F_{ST} (Figure S1) that was found significant, with the Mantel test indicating r2= 0.801 and a simulated p-value of 0.009.

After identifying outlier SNPs in Baypass based on *XtX*, we compared them with the top outliers in Bayenv2 based on *XtX*: top outliers coincided between the two methods, corresponding to a cut off in the *XtX* distribution of 20 for Bayenv2. Therefore, SNPs with XtX above 20 were considered as outliers. The four top outliers from the *XtX* statistical analysis were as follows: SNP 378 (seq-2_3941_01-381), SNP 316 (seq-10373-2483), SNP 4 (seq-9882-801) and SNP 149 (seq-8671-529) (Table 3.1; Figure S2, left). When the Bayesian linear model was used with the 20 different environmental variables, four potential outlier SNPs were detected, four classified as strong candidates for selection (20<BF<150), based on the Kass and Raftery (1995) criteria: SNP 169 (seq-0_10162_01-244), SNP 312 (seq-UMN_3408_01-293), SNP 316 (seq-10373-2483) and SNP 378 (seq-2_3941_01-381). Seven environmental variables were involved in these correlations: altitude, geography, BIO 2 (mean diurnal range), BIO 9 (mean temperature of driest quarter), BIO 12 (annual precipitation), BIO 16 (precipitation of wettest quarter) and BIO 19 (precipitation of coldest quarter) (Table S5).

Bayenv2		
SNP number	sequence	*XtX* value
149	seq-8671-529	20,42
4	seq-9882-801	20,83
316	seq-10373-2483	21,52
378	seq-2_3941_01-381	23,76
Baypass		
SNP number	sequence	*XtX* value
378	seq-2_3941_01-381	13,4401295
149	seq-8671-529	14,1439954
PCAdapt		
SNP number	sequence	p-value
4	seq-9882-801	7.23E-004
10	seq-44358-1615	3.51E-003
19	seq-55383-1485	2.05E-007
94	seq-55383-900	1.74E-007
113	seq-44358-2515	3.80E-003
258	seq-9882-2209	4.74E-004
269	seq-16094-1379	5.90E-003
281	seq-16094-410	3.36E-003
331	seq-55383-141	2.95E-007
335	seq-1_6493_01-100	2.64E-009
384	seq-0_3073_01-92	1.21E-003

Table 3.1: Summary of the results from all three statistical outlier methods

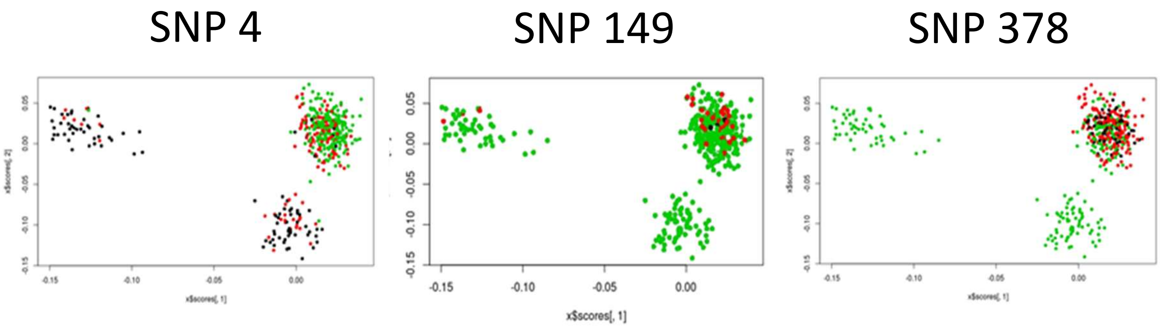


Figure 3.3: Highlights on the score plot for the three main SNP outliers common to the outlier methods based on the *XtX* statistics (Bayenv2 and Baypass) and PCA (PCAdapt). SNPs monomorphic for allele 1 are in green, monomorphic for allele 2 are in black, and heterozygote alleles are in red. The populations are the same as in **Figure 3.2**.

Using Baypass

Confidence in the correlation matrix $\widehat{\Omega}$ produced for correcting for pre-existing population structure in order to produce *XtX* statistics on our SNP data, was firstly assessed visually by comparing the $\widehat{\Omega}$ values amongst the populations (Figure S3). The correlation matrix can be viewed as a hierarchical cluster tree (Figure S4, left), in which the relationships of the seven populations were similar to those found with a neighbour joining tree based on pairwise F_{ST} (Figure S4, right), with populations from a same country clustering together (the exception being the French population from Font Blanche in the hierarchical cluster based on $\widehat{\Omega}$ that includes much more individuals – from 5 to 10 times more than the other studied populations). Two outliers SNPs were observed: SNP 149 (seq-8671-529) and SNP 378 (seq-2_3941_01-381) (**Table 3.1**; Figure S2, right).

Using the Bayesian outliers approach simultaneously with the Baypass addition, which calculates the empirical Bayesian P value (eBPis) and thus allowing to evaluate the support in favour of a non-null regression coefficient eBPis>3, sixteen SNP outliers were found (see Table S6). These were then compared to the ones in Bayenv2 and only the outliers that coincided in both methods for the same environmental variable were included as high candidates for selection, leading to a total of three SNPs (**Table 3.2**).

SNP	Sequence	Env.	BF Bayenv	eBPis Baypass
169	seq-0_10162_01-244	BIO9	41.97	5.48337617
316	seq-10373-2483	Altitude	20.90	3.76859495
378	seq-2_3941_01-381	BIO12	47.38	3.65435041

Table 3.2: SNPs that coincided in being outliers for the same environmental variables for both Bayesian linear models perform with Bayenv 2 and Baypass.

Looking for outliers using PCA

The scree plot showed clearly that after K=3, most of the variation is accounted for and it is unnecessary to use more PCs (Figure S5). This was further confirmed using a score plot (**Figure 3.2**). A Manhattan and a QQ-plot were used to visualise the distribution of the p-values before their conversion to q-values (Supplementary materials Figure S6), eleven outliers were detected (**Table 3.1**) using q-values at false discovery rate of 10%.

3.4.3 Results of all outlier tests

As illustrated in the Venn diagrams, a total of 8 SNPs were identified as outliers in at least two methods (**Figure 3.4**): SNP 4, SNP 149, SNP 169, SNP 258, SNP 269, SNP 281, SNP 316 and SNP 378. The Bayesian outlier method making use of *XtX* statistics showed that the two outliers detected in Baypass correspond to two of the four detected in Bayenv2 (SNP 149 and

SNP 378; **Figure 3.4**a): this result is logical as Baypass is based on Bayenv2 with some modifications that are meant to refine the detection of outliers. One of the SNP outliers (SNP 378) was also found in both Bayesian linear methods in association with environmental variables in both Bayenv2 and Baypass (**Figure 3.4**b). The PCA-based method performed in PCAdapt found eleven outliers; of these one outlier was common with the *XtX* outliers found in Bayenv2 (SNP4; **Figure 3.4**a).

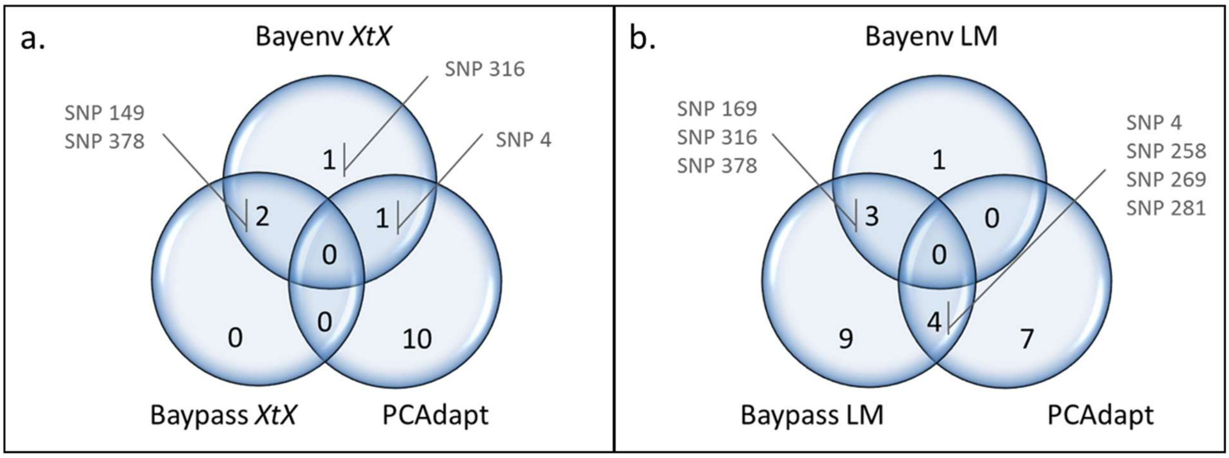


Figure 3.4: Venn diagrams comparing the outliers detected using PCAdapt with those identified under selection using the *XtX* statistics from Bayenv2 and Baypass (a), and with those identified using the Bayesian linear model (LM) from Bayenv2 and Baypass (b). Only the SNPs found under selection in more than one method are indicated.

The Bayesian linear models performed in Bayenv2 and Baypass to assess if outlier SNPs were found in association with different environmental variables revealed three outlier SNPs (**Figure 3.4**b), correlated with the same environmental variables in both methods: SNP 169 was found in association with BIO9 (mean temperature driest quarter), SNP 316 in association with altitude and 378 with BIO12 (annual precipitation). Out of the eleven SNPs detected with PCAdapt, four were in common with the Bayesian linear models performed in Baypass (SNP

4, SNP 258, SNP 269, and SNP 281; Figure 4b), all associated with BIO12 (Annual precipitation) (**Table 3.3**). Finally, one SNP detected with Bayenv based on the *XtX* statistic was in common with the two Bayesian linear models performed in Baynv2 and Baypass (SNP 316; Figure 3.4 a and b). A BLAST search for the sequences corresponding to all detected SNP outliers returned no results.

SNP	Contig	p-value PCAdapt	eBPis Baypass	Env.
4	seq-9882-801	7.23E-004	5.37596943	BIO12
258	seq-9882-2209	4.74E-004	5.0545209	altitude
258	seq-9882-2209	4.74E-004	6.33377503	BIO12
258	seq-9882-2209	4.74E-004	3.01268994	BIO19
269	seq-16094-1379	5.90E-003	4.85579828	BIO12
269	seq-16094-1379	5.90E-003	3.77450901	BIO16
269	seq-16094-1379	5.90E-003	4.72441191	BIO19
281	seq-16094-410	3.36E-003	3.2116434	altitude
281	seq-16094-410	3.36E-003	5.20973055	BIO11
281	seq-16094-410	3.36E-003	3.02821708	BIO12
281	seq-16094-410	3.36E-003	4.64057127	BIO16
281	seq-16094-410	3.36E-003	4.91602902	BIO19

Table 3.3: SNPs found under selection in both PCAdapt and Baypass using the linear model to find SNPs associated with different environmental variables

3.5 Discussion

To detect SNPs under putative selection it is paramount to select both a good sampling technique and a suitable methodology. The aim of this study was to find instances of local adaptation in Aleppo pine by considering pairs of populations from genetically similar groups that were exposed to contrasting climatic conditions important for the species adaptation. This sampling design was combined with various recently developed statistical methods to optimise genetic outlier detection. We discuss afterwards the main results, as well as assess the utility of these statistical methods in empirical study.

3.5.1 Local adaptation in Aleppo pine

Both the F_{ST} analysis and the Bayesian clustering confirmed that the population pairs/triplets have had distinct evolutionary histories from each other: French and Spanish populations were genetically less differentiated possibly because they belong to the more recently colonised Western cluster of Aleppo pine, while Italian populations were more differentiated from the two others most likely because they are part of the older Eastern/Southern cluster (Ruiz Daniels *et al.* in review). Our sampling covered therefore not only contrasted environmental conditions within population pairs, but also distinct evolutionary genetic units. Sampling has been found to be one of the most influential factors when performing outlier detection tests in studies with simulated data (Lotterhos & Whitlock 2015; Hoban *et al.* 2016). It is only recently, that empirical analyses have integrated pairwise sampling strategy to increase the power of outlier detection (e.g. Nadeau *et al.* 2016). Other studies have made used of different sampling strategies to maximise outlier detection, such as African origin population compared to expanding human populations (Peischl *et al.* 2016), introduction compared to expansion populations in the invasive bank vole (White *et al.* 2013), and core compared to edge populations in the damselfly (Swaegers *et al.* 2015).

Three SNPs under putative selection were detected by at least two outlier methods with clear significant threshold, SNP 4, SNP 149 and SNP 378. Unsurprisingly, it was not possible to identify the genes corresponding to these SNPs because of the limited genomic resources available for conifers (De La Torre *et al.* 2014). SNP4 was found under selection with the PCAdapt method, and with Baypass linear model in association with annual precipitation, while SNP 378 was detected under selection in the two Bayesian linear models in association with annual precipitation. Finally, SNP 378 was detected independently under selection in relation to the precipitation of driest month in full-scale Europe wide study of this species (Ruiz Daniels *et al*. in review), using the Bayenv2 linear model. Altogether these results point to SNP 378 as a very good candidate for selection and indicate that precipitation is a potential predominant driver of selection at two distinct spatial scales. Although our population pairs were sampled at contrasted altitudes to reflect differences in temperature and precipitation, two climatic factors known to be potential pressures for Aleppo pine (Santos-del-Blanco *et al.* 2013; Voltas *et al.* 2015; Ruiz Daniels *et al.* in review), it may well be that the outlier SNPs may be correlated to other environmental factors linked to the ones under study but that were not contemplated (because they are unavailable or impossible to test).

When observing in more detail the allelic frequency distribution of the three candidate SNPs, several patterns emerged: the allelic frequency distribution of SNP 4 indicated that allele A was in higher frequency in sites of low altitude (i.e. dry sites) and in lower frequency in sites of high altitude (i.e. wet sites) across the three countries, while its frequency was intermediate for the French site located at intermediate altitude (Figure S7). The allelic frequency distribution was somehow similar for SNP 149 and SNP 378 as the higher proportion of one allele was associated with one site (Figure S7), but this was true for only a subset of pairs (only the French and the Italian pair for SNP 149, and only the French pair for SNP 378; the other pairs being monomorphic), while the allelic frequency for the French population located at intermediate altitude was not intermediate. These results point to the importance of taking into account several points when interpreting the outputs of the selection tests at the local scale: i) the evolutionary history of the species (for Aleppo pine, the Spanish populations lack polymorphism due to their recolonization history); ii) the sampling size in terms of individuals per population (much more precise allelic frequencies could be obtained for the French population composed of 356 individuals compared to the Italian and Spanish populations comprised between 25-40 individuals); iii) the sampling size in terms of population repetition (three population pairs are somewhat limited to draw general conclusion, especially if some populations are monomorphic for the targeted SNPs).

3.5.2 Comparison of statistical methods

Discovering genes under local adaptation is challenging, and because methods differ in the way that they summarise complex data, they will be subjected to noise (Lotterhos & Whitlock 2015), which one must take into account when looking for SNPs under putative selection. Lotterhos and Whitlock (2015) found that it is to be expected that the use of different statistical frameworks would yield different genes under selection, which is the case in both theoretical (Lotterhos & Whitlock 2014, 2015) and empirical (Schweizer *et al.* 2016; Yeaman *et al.* 2016; Nadeau *et al.* 2016) works, including the present study. As of yet there is no single widely accepted approach in order to infer genetic outliers (Rellstab *et al.* 2015). The one main message is that, when present, population structure must be corrected for (De Mita *et al.* 2013; Blair *et al.* 2014; Lotterhos & Whitlock 2014, 2015; Rellstab *et al.* 2015). We approached the challenge of adopting a suitable methodology by contrasting three statistical outlier methods able to correct for population structure: two Bayesian methods and one based on PCA.

The different methods use different statistics and associated cut-off thresholds to identify outlier loci. In this study, we chose the cut-off threshold recommended by the different programs when available. Methods with a clear statistical threshold for outliers (Baypass and PCAdapt) are more straightforward to interpret than those that do not have any (Bayenv2). This is especially true when assessing *XtX* outliers in Bayenv2, where the cut-off threshold is hard to establish (see supplementary materials Figure S2, left). Also, when using the Bayesian linear model to look for SNPs significantly associated with climatic variables, Bayenv2 comes with no predetermined way of assessing significant outliers, and therefore the criterion of the Kass and Ratfery (1995) table was used for this purpose. Baypass by contrast comes with a predetermined way of inferring outliers by calculating posterior predictive p-values, and this criterion led to a much less conservative estimate of the number of outliers (19 SNPs) compared with those found in Bayenv2 (4SNPs) mainly because we chose the default parameters in the program with an eBPis above 3 (Gautier 2015). This threshold can of course be adjusted differently and be made more or less conservative.

Previous empirical studies on Aleppo pine at the full distribution scale that compared simulated data sets to actual data to infer outlier SNPs, found that PCAdapt yielded more false positives that Bayenv2 (Ruiz Daniels et al. under review). It has previously been observed in comparative studies that make use of simulated data involving varied demographic scenarios, that latent factor and Bayesian outlier models which make use of a covariance matrix tend to outperform other methods that do not correct for population structure when detecting outlier loci (De Mita et al. 2013; Blair et al. 2014; Lotterhos & Whitlock 2014; Rellstab et al. 2015). These two approaches tended to perform differently under different demographic conditions, although the difference was not too marked (Lotterhos & Whitlock 2015). Furthermore, in the aforementioned work, although it was found that these two methods worked similarly well in detecting outliers when compared by means of empirical p-values, it did indicate that the false positive rates for latent factor models (such as those used in PCadapt) are undesirably high (Lotterhos & Whitlcock 2015). In this study, we found that the latent factor method yielded a higher number of outliers, but among them some corresponded to those detected with Bayesian linear models, adding more confidence to these. Finally, it must be noted that the PCAdapt method and the Bayesian linear mode in Baypass found three common outliers (SNP 258, SNP 269 and SNP 281), that were not detected by either the XtX statistics analysis or the Bayesian linear model in Bayenv2, suggesting that these are missed out by Bayenv2 altogether.

It is disputed what to do in case where loci are found to be under selection for one method but not another. One solution is to consider only outliers in common between the different methodologies, as done in this study as well as other empirical works (Schweizer et al. 2016; Nadeau et al. 2016). However, looking only at outliers in common between different methodologies may lead to loci under selection being missed, especially those under weak selection. There are some indications that many SNPs are under weak selection in conifers (Yeaman et al. 2016): local adaptation could result mainly from small, potentially undetectable, covarying shifts in frequency at many loci (Hoban et al. 2016). Looking for polygenic adaptation adds an extra layer of complexity when trying to disentangle the effects of demographic history and selection. Some methods have been developed to address that issue by looking at subtle allelic frequency shifts at many loci, in order to detect polygenic adaptation from genomic data (Turchin et al. 2012; Daub et al. 2013; Berg & Coop 2014).

3.5.3 Conclusions

Detecting outlier loci in empirical studies can be greatly assisted using a paired sampling technique as well as the comparison of various statistical methods that aid data exploration, in addition to add confidence to the detected outlier loci. With this combined approach, three outliers were found (SNP 4, SNP 149 and SNP 378) in Aleppo pine at the local scale, each detected by two outlier methods using different statistical approaches. However, the allelic frequency pattern associated to the contrasted environmental conditions was present across the three population pairs for SNP 4 only, indicating that cautious should be taken when analysing the outputs of the selection methods. In addition to an adequate sampling design and the combination of multiple statistical methods, the knowledge of the species evolutionary history

is prerequisite for effectively detecting SNP subjected to selection. Finally, annual precipitation was detected as an important driver of selection for several SNPs detected with Bayesian linear models. This study provides candidate loci that may be responsible for local adaption in a conifer highly adapted to the Mediterranean environment.

3.6 References

Atzmon N, Moshe Y, Schiller G (2004) Ecophysiological response to severe drought in *Pinus halepensis* Mill. trees of two provenances. *Plant Ecology (formerly Vegetatio)*, **171**, 15–22.

Berg JJ, Coop G (2014) A population genetic signal of polygenic adaptation. *PLoS genetics*, **10**, e1004412.

Blair LM, Granka JM, Feldman MW (2014) On the stability of the Bayenv method in assessing human SNP-environment associations. *Human Genomics*, **8**, 1.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–23.

Daub JT, Hofer T, Cutivet E *et al.* (2013) Evidence for polygenic adaptation to pathogens in the human genome. *Molecular Biology and Evolution*, **30**, 1544–1558.

Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB (2016) Detecting genomic signatures of natural selection with principal component analysis: Application to the 1000 genomes data. *Molecular Biology and Evolution*, **33**, 1082–1093.

Frichot E, Schoville SD, de Villemereuil P, Gaggiotti OE, François O (2015) Detecting adaptive evolution based on association with ecological gradients: Orientation matters! *Heredity*, **115**, 22–28.

Gautier M (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, **201**, 1555–1579.

Grivet D, Sebastiani F, Alía R *et al.* (2011) Molecular footprints of local adaptation in two Mediterranean conifers. *Molecular biology and evolution*, **28**, 101–16.

Grivet D, Sebastiani F, González-Martínez SC, Vendramin GG (2009) Patterns of polymorphism resulting from long-range colonization in the Mediterranean conifer Aleppo pine. *New Phytologist*, **184**, 1016–1028.

Gunther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hoban S, Kelley JL, Lotterhos KE *et al.* (2016) Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *The American Naturalist*, **188**, 379–397.

Kass RE, Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association*,

90, 773–795.

De La Torre AR, Birol I, Bousquet J *et al.* (2014) Insights into conifer giga-genomes. *Plant physiology*, **166**, 1724–32.

Lavi A, Perevolotsky A, Kigel J, Noy-Meir I (2005) Invasion of *Pinus halepensis* from plantations into adjacent natural habitats. *Applied Vegetation Science*, **8**, 85.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, **23**, 2178–2192.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

de Luis M, Čufar K, Di Filippo A *et al.* (2013) Plasticity in Dendroclimatic Response across the Distribution Range of Aleppo Pine (*Pinus halepensis*). *PLoS ONE*, **8**, e83550.

Luu K, Bazin E, Blum MGB (2017) pcadapt : an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, **17**, 67–77.

McVean G (2009) A Genealogical Interpretation of Principal Components Analysis. *PLoS Genetics*, **5**, e1000686.

De Mita S, Thuillet A-C, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

Nadeau S, Meirmans PG, Aitken SN, Ritland K, Isabel N (2016) The challenge of separating signatures of local adaptation from those of isolation by distance and colonization history: The case of two white pines. *Ecology and Evolution*, **6**, 8649–8664.

Narum SR, Hess JE (2011) Comparison of F ST outlier tests for SNP loci under selection. *Molecular Ecology Resources*, **11**, 184–194.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Peischl S, Dupanloup I, Bosshard L, Excoffier L (2016) Genetic surfing in human populations: from genes to genomes. *Current Opinion in Genetics & Development*, **41**, 53–61.

Pinosio S, González-Martínez SC, Bagnoli F *et al.* (2014) First insights into the transcriptome and development of new genomic tools of a widespread circum-Mediterranean tree species, *Pinus halepensis* Mill. *Molecular ecology resources*, **14**, 846–56.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–59.

Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R (2015) A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, **24**, 4348–4370.

Rouget M, Richardson DM, Milton SJ, Polakow D (2001) Predicting invasion dynamics of four alien *Pinus* species in a highly fragmented semi-arid shrubland in South Africa. *Plant Ecology*, **152**, 79–92.

Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Santos-del-Blanco L, Bonser SP, Valladares F, Chambel MR, Climent J (2013) Plasticity in reproduction and growth among 52 range-wide populations of a Mediterranean conifer: adaptive responses to environmental stress. *Journal of Evolutionary Biology*, **26**, 1912–1924.

Santos-del-Blanco L, Climent J, González-Martínez SC, Pannell JR (2012) Genetic differentiation for size at first reproduction through male versus female functions in the widespread Mediterranean tree *Pinus pinaster*. *Annals of botany*, **110**, 1449–60.

Sathyan P, Newton RJ, Loopstra CA (2005) Genes induced by WDS are differentially expressed in two populations of aleppo pine (*Pinus halepensis*). *Tree Genetics & Genomes*, **1**, 166–173.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature reviews. Genetics*, **14**, 807–20.

Schweizer RM, vonHoldt BM, Harrigan R *et al.* (2016) Genetic subdivision and candidate genes under selection in North American grey wolves. *Molecular Ecology*, **25**, 380–402.

Serra-Varela MJ, Alía R, Ruiz Daniels R *et al.* (2017) Assessing vulnerability of two Mediterranean conifers to support genetic conservation management in the face of climate change. *Diversity and Distributions*, **5**, 507–516.

Swaegers J, Mergeay J, Van Geystelen A *et al.* (2015) Neutral and adaptive genomic signatures of rapid poleward range expansion. *Molecular Ecology*, **24**, 6163–6176.

Turchin MC, Chiang CW, Palmer CD *et al.* (2012) Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics*, **44**, 1015–1019.

Voltas J, Chambel MR, Prada MA, Ferrio JP (2008) Climate-related variability in carbon and oxygen stable isotopes among populations of Aleppo pine grown in common-garden tests. *Trees - Structure and Function*, **22**, 759–769.

Voltas J, Lucabaugh D, Chambel MR, Ferrio JP (2015) Intraspecific variation in the use of water sources by the circum-Mediterranean conifer *Pinus halepensis*. *New Phytologist*, **208**, 1031–1041.

White TA, Perkins SE, Heckel G, Searle JB (2013) Adaptive evolution during an ongoing range expansion: The invasive bank vole (*Myodes glareolus)* in Ireland. *Molecular Ecology*, **22**, 2971–2985.

Yeaman S, Hodgins KA, Lotterhos KE *et al.* (2016) Convergent local adaptation to climate in distantly related conifers. *Science*, **353**, 1431-1433

(Part II) Using a Bayesian linear model for outlier detection in three different conifer species

Rose Ruiz Daniels[1] & Delphine Grivet[1,2]

[1] Department of Forest Ecology and Genetics, Forest Research Centre, INIA, Carretera A Coruña km 7.5, 28040 Madrid, Spain.

[2] Sustainable Forest Management Research Institute, INIA, University of Valladolid, 34004 Palencia, Spain

3.7 Introduction and aims of this study

As we saw in the first part of this chapter, sampling in paired populations that are under divergent ecological conditions with similar evolutionary history, is an effective way to identify outlier genes potentially under local adaptation (Hoban *et al.* 2016). Current sequencing technologies, combined with well-designed field sampling of natural populations, can in theory lead to the identification of selectively significant loci using population genomic approaches (Lotterhos & Whitlock 2015). Being able to sequence thousands of single nucleotide polymorphisms (SNP) for a significant number of populations has currently become technically feasible and economically possible, even for non-model species such as conifers. In the present study, we used recently developed SNPs from exome capture in several conifer species within the framework of an European project (FLAG: Forest tree ecological genetics: interplay of gene flow and environmental variability in shaping local adaptation and genetic adaptive potential; ANR-Bioadapt, France) to look for footprints of selection at a local scale in three species of conifers, using a paired sampling technique from close sites with contrasting environmental conditions but similar evolutionary histories. A Bayesian linear model (Gunther & Coop 2013) was applied to identify SNP loci under divergent selection in these populations. This allows us to extend previous works and previous bioinformatic knowledge and apply it to large genomic datasets, in order to look for signatures of microgeographic differentiation between populations.

3.8 Materials and Methods

3.8.1 Sampling

Three species for which molecular data were available at this stage of the European project (FLAG) were analysed: silver fir (*Abies alba*), Atlas cedar (*Cedrus atlantica*) and Aleppo pine (*Pinus halepensis*). In this study, each site is characterized by a pair of geographically adjacent populations (**Figure 3.5**), with each population composed of 25 mature individuals sampled at least 20m apart from each other. Thus, for each of the three species there were *n* sites and *2n* populations (**Error! Reference source not found.**). Paired populations within sites were characterized with contrasted environmental conditions but similar evolutionary histories (**Error! Reference source not found.**). 50 mg of needles was dried with silica gel and grinded in QIAwell (Qiagen, Venlo, The Netherlands) plate homogenized with mixer mill MM300 (RETSH, Haan, Germany) under liquid nitrogen. DNA extractions were carried out with the kit Invisorb DNA plant HTS 96 kit (Invitek, Hayward, CA) following the manual instructions. DNA was quantified with Nanodrop 10000 (Thermo Fisher Scientific, Wilmington, DE).

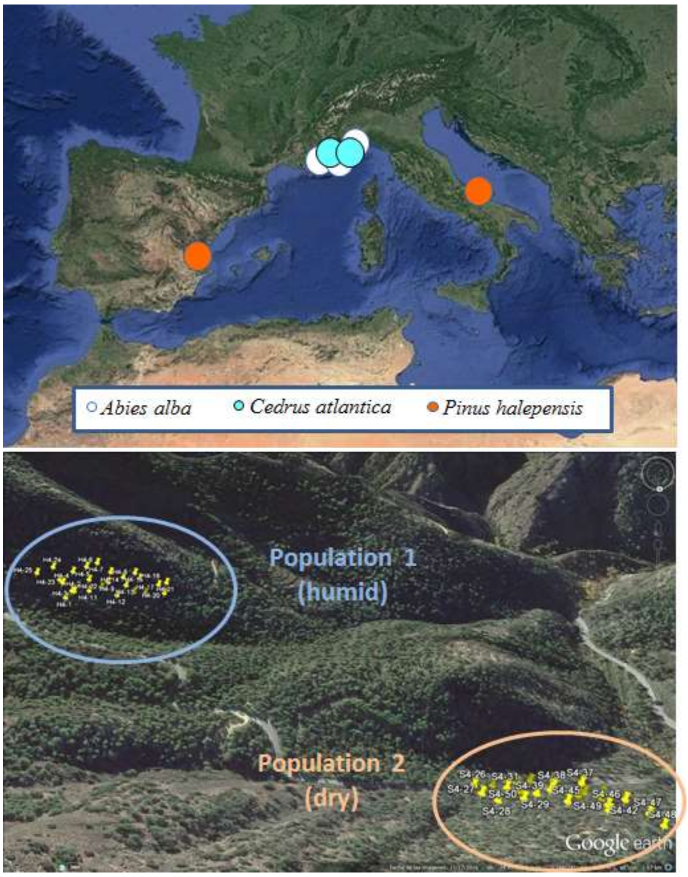


Figure 3.5: Sampling sites for the three species (upper figure), along with an example of a paired populations characterized by contrasted environmental conditions (lower figure).

3.8.2 Genotyping

DNA samples were submitted to sequence capture using a subset of the sequence capture probes designed for *Pinus taeda* by Rapid Genomics (Gainesville, FL). Exome capture was carried out by Rapid Genomics (Gainesville, FL). Captured fragments were paired-end sequenced with Illumina GA2X technology, producing 100-bp reads. Raw sequencing data were then trimmed with CUTADAPT (http://cutadapt.readthedocs.io/en/stable/index.html), with minimum read length = 70, and trimming both at the 5' and at the 3' (quality threshold = 20). For each species, a *de novo a*ssembly was then performed on a single individual, chosen based on the total number of available reads and on average per-base read quality. VELVET v. 1.2.07 (Zerbino & Birney 2008; Zerbino 2010) was used for assembly. K-mer length was allowed to vary between 31 and 89 while running *velveth,* and then *velvetg* was run on each output with coverage cut-off = 8, minimum contig length = 100 and expected coverage = 20. The number of contigs obtained with each K-mer length value was compared to the number of sequences we expected to retrieve through sequence capture (i.e. the number of probes, see above) and the assembly whose number of contigs more closely approximated the expectation was retained. The chosen K-mer length was 37, 85 and 57. For all individuals (including the one used for assembly), read mapping was performed with BOWTIE2 v 2.0.6 (Langmead *et al.* 2009; Langmead & Salzberg 2012). Following read mapping, variant calling and filtering were carried out as follows: SAMTOOLS MPILEUP was used to build an *mpileup* file from all individual alignments within each species; VARSCAN v. 2. 3.9 (Koboldt *et al.* 2009) was applied to the *mpileup* file (with parameters: minimum coverage = 20; minimum average quality = 30, minimum variant frequency = 0.3, minimum frequency for homozygote call = 0.7); the *vcf* file obtained with VARSCAN was submitted to VCFTOOLS (https://vcftools.github.io/index.html), with parameters: minimum number of alleles = 2; maximum number of alleles = 2; maximum fraction of missing data = 0.2).

Following these filtering steps based upon read and base call quality and quantity, further filters were applied based on the biological properties of variants and their distribution, using the ad-hoc R (R Development Core Team 2008) Ffff script (Developed by the Scotti lab). The Ffff script takes as input a *vcf* file and filters out contigs based on a SNP density (variants per base) and loci based on a heterozygosity threshold, plus it allows the removal of targeted individuals, and returns genotype tables; conversion from *vcf* to genotype tables is performed by internally calling PGDSPIDER v. 2.1.0.3 (Lischer & Excoffier 2012). Contigs with variant density > 0.05 (i.e. having more than one variant every 20 bases) and heterozygosity = 1 (I. e. carrying only heterozygote genotypes) were excluded, to remove contigs potentially containing paralogs for the characterisation of variant distribution in each species' data set. After filtering with Ffff and removal of trailing monomorphic loci, the final data set contained 6197 SNPs, 7246 SNPs and 7686 SNP contigs respectively for *A. alba*, *C. atlantica*, and *P. halepensis*.

3.8.3 Running Bayenv2 for outlier detection

The basic idea is to identify loci where the environmental variable of interest has a linear effect on the allele frequencies across populations. This method is carried out in a two-step procedure. First a covariance matrix is computed with the SNPs and then this matrix is used to identify loci where environmental variables of interest have a linear effect on the allele frequencies across populations, this way the methodology accounts for the neutral correlation of allele frequencies across populations. This method is implemented in the Bayenv 2.0 package (Gunther & Coop 2013), which estimates the empirical pattern of covariance in allele frequencies between populations, and uses this as the null model to test each individual SNP for selection.

Input files were produced using Python script (R. Ruiz Daniels), which also removed the monomorphic SNP present in the data. As it is common with big databases, there is always the issue that a few monomorphic SNPs escape the rigorous process of cleaning, and will not allow the program to run. Furthermore, the script automated the Bayenv 2.0 program so it runs all SNPs without having to manually rerun the program for each SNP, making these analyses more feasible for big data sets. Bayenv 2.0 was run using Python with 100.000 interactions to produce the covariance matrix. Then in order to identify selection acting on SNPs we used the method described in Coop et al. (2010) to evaluate the association between ecological variables and SNP frequencies.

The ecological variables of interest in each species are listed in **Table 3.4** and populations were selected so as to reflect contrasted environmental conditions for these variables. Instead of using the real environmental variables, we used dud ones (i.e.by coding areas of low humidity as 0 and high humidity as 1) as we did not have precise climatic measures for each species and sites, the aim was to compare the outputs across species. In addition, for *Pinus halepensis* we also used numbers ranking populations from East to West, as a strategy to identify outliers using this method. Unfortunately, due to time constraints, the analysis could not be repeated for the other two species.

Detecting environmental adaptation was then performed by means of Bayes factors (*K*), to see wherever a correlation between the allele frequencies at a SNP, and an environmental variable is greater than expected given the null model. That this relationship showed a particular high Bayes factor and thus deviated from the null model was verified by using the Kass & Raftery (1995) table where a Bayes factor of higher than 20 was considered of interest (i.e. corresponding to 2ln*K*>6) (**Table 3.5**).

Species	**Population pair number**	**SNP number**	**Environmental variable**
Abies alba	3	5992	humidity
Cedrus atlantica	2	6456	soil condition
Pinus halepensis	2	7695	humidity and East/West

Table 3.4: Sampling, genomic and environmental datasets for each species

2 ln *K*	***K***	**Strength of evidence**
0 to 2	1 to 3	not worth more than a bare mention
2 to 6	3 to 20	positive
6 to 10	20 to 150	strong
>10	>150	very strong

Table 3.5: The Kass and Ratfery table

3.9 Results and Discussion

Although the number of populations pairs were more limited than in our previous work (two for *C. atlantica* and *P. halepensis*, and three for *A. alba*), as well as the number of environmental variables tested (only one for *A. alba* and *C. atlantica*, and two for *P. halepensis*), we did find outliers that are potential candidates for selection. These results may be attributed to the higher number of SNPs tested combined with the specific sampling design that was optimising contrasting environmental conditions. The number of outlier SNPs with *K* ≥20 for each species was as follows: five for *A. alba* (Table 3.6; Figure 3.6(left)), one for *C. atlantica* (Table 3.6; Figure 3.6 (right)), and seven for *P. halepensis* (six when considering ranked populations from East to West) (Table 3.6; Figure 3.7). In *P. halepensis*, two SNPs were in common between the two environmental variable coding methods: 16710_279 and 10770_281 (Table 3.6).

SNP	***K***
Abies alba	
12088_141	38.2
3585_5	36.9
689_631	27.1
3379_5	24.4
8714_329	20.2
Cedrus atlantica	
2222_191	86.394
Pinus halepensis	
16710_279	129.54
1929_404	55.045
5216_581	50.93
2580_1062	48.314
15421_24	43.565
12407_251	41.699
2204_33	35.97
12529_389	25.521
10770_281	25.238
18603_67	23.63
10028_469	22.638
Pinus halepensis ranked populations	
14854_103	368.26
10770_281	50.623
16710_279	46.432
13836_251	44.19
8335_110	32.394
19308_42	23.199
20017_155	22.957

Table 3.6: Results of the Bayesian linear analysis showing Bayes factors $K \geq 20$

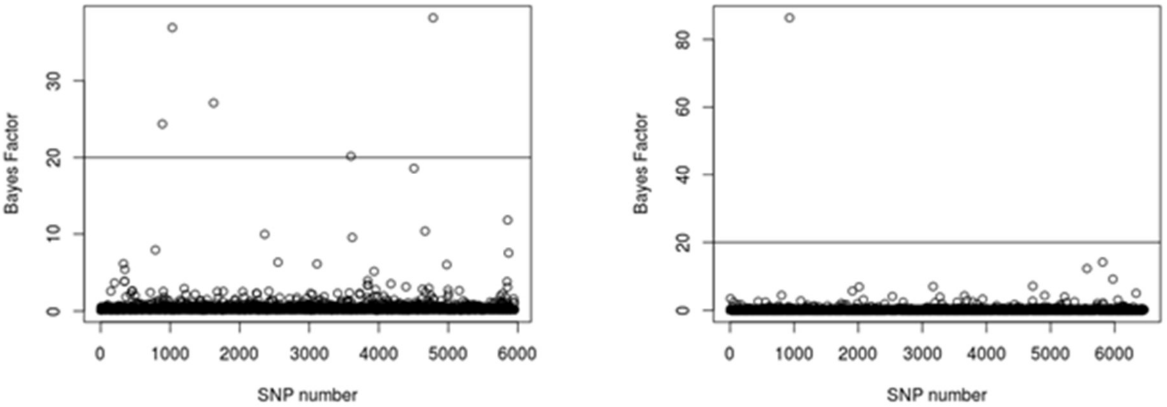


Figure 3.6: Graphs showing the distribution of Bayes factors for *Abies alba* (left) and *Cedrus atlantica* (right) the line represents the threshold at which evidence of a data point being an outlier is strong with a *K* ≥20.

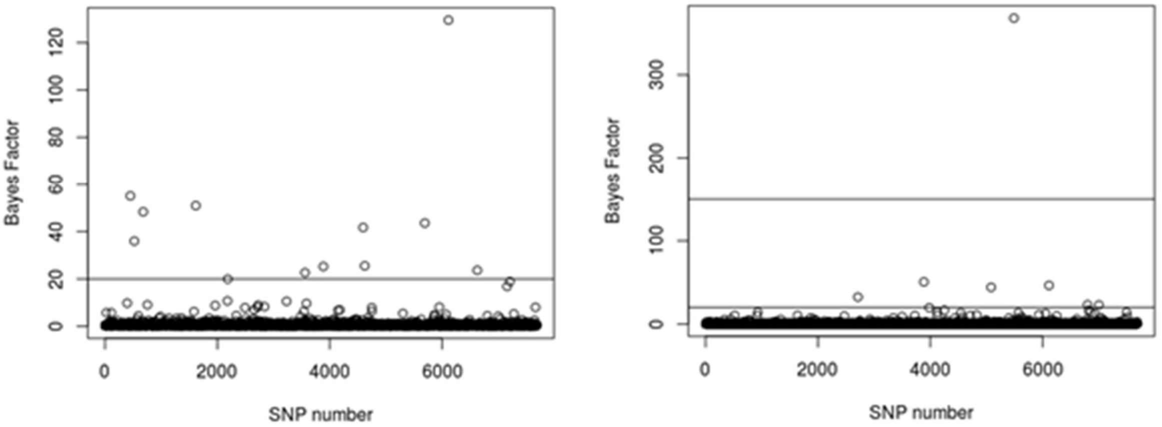


Figure 3.7: Bayes factors for *P. halepensis* for humidity (left) and geography (right). The line represents the threshold at which evidence of a data point being an outlier is strong with a K ≥20 and very strong K ≥150 in the case for the second line (right).

It is important to bear in mind that only two environmental states were used to characterise population pairs for each species, and although the populations were selected based on contrasted environmental conditions, it is possible that other environmental variables may be the drivers for selection in these species. In addition, it would be necessary to check the involvement of these outlier SNPs in the control of an adaptive trait, so that we would have the evidence that the locus acts on adaptive processes through the expression of the trait. Connecting genotypes to phenotypes constitutes a real challenge and is beyond the scope of the present study.

A further limitation in the present study, is due to the limited genomic resources available for conifers, none of the locus corresponding to SNPs detected under selection gave a match with annotated species in Genbank. It would be interesting, not only to compare shared sets of loci across species (via refined annotations or/and blast of contigs), but also to see whether the SNPs found under selection in the different studies in *Pinus halepensis* belong to the same gene families, and therefore if selection is acting similarly or not at different spatial scales. In the present work, we focus on the local scale, where gene flow can counteract or facilitate the action of selection, but since habitat selection is a scale- and temporal-sensitive process, we do not expect selection to act similarly at different spatial-temporal scales. Analysing multiple scales can help unravel the complete picture of how selection is acting.

This work is part of a European consortium, and the final dataset will allow comparative studies for footprints of selection both across methods and species, the aim being to measure and model the amount of divergence among populations under divergent ecological conditions, to be able to predict how populations may respond to future local and global environmental changes.

3.10 References

Gunther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.

Hoban S, Kelley JL, Lotterhos KE *et al.* (2016) Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *The American Naturalist*, **188**, 379–397.

Kass RRE, Raftery AEA (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357-U54.

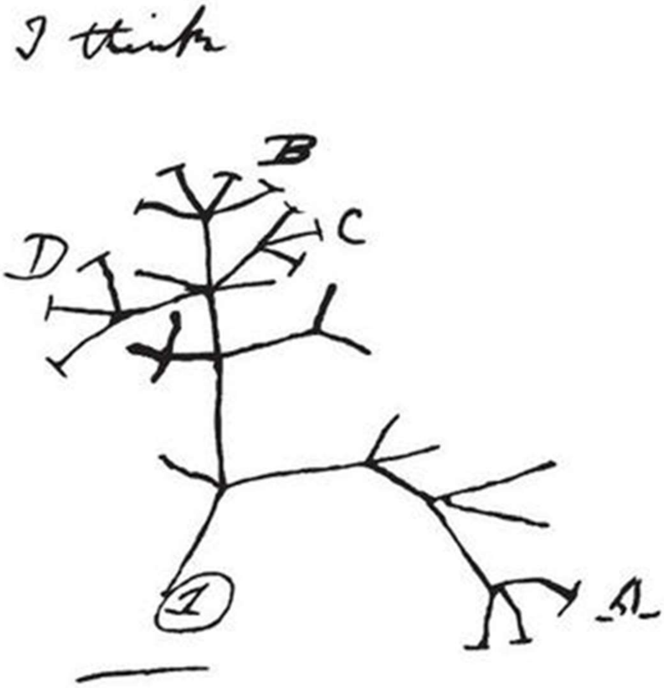
Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–299.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

Zerbino DR (2010) Using the Velvet de novo Assembler for Short-Read Sequencing Technologies. In: *Current Protocols in Bioinformatics*, p. Unit 11.5. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.



Charles Darwin. Notebook sketch for his 'Tree of Life'. 1837.

Chapter 4 **Inference of microRNA orthology in Brassicaceae**

Rose Ruiz Daniels[1*], Richard S. Taylor[2*] & Philip Donoghue[2]

*These authors contributed equally to this work.

[1] Department of Forest Ecology and Genetics, Forest Research Centre, INIA, Carretera A Coruña km 7.5, 28040 Madrid, Spain.

[2] School of Earth Sciences, University of Bristol, Wills Memorial Building, Queen’s Road, Bristol BS8 1RJ, UK

4.1 Abstract

The previous chapters in this thesis have focused on traditional population genetic methods to investigate adaptation and population structure of *Pinus halepensis* in the post-glacial colonisation of the Mediterranean basin, and this has produced impressive results. The aim of this chapter is to focus on an area of research that has been explored less in the context of adaptation, microRNAs (miRNAs), but which may in the future yield important results. There have been few studies so far looking at the role of miRNAs in conifers (Yakolev *et al.* 2010; Quinn. *et al.* 2014; Qui *et al.* 2016) and the bulk of research has been conducted in angiosperms, largely due to the far greater availability of genomic resources in this group of plants. In contrast to conifers, numerous studies in angiosperms have highlighted the importance of miRNAs in a range of processes, and the evolutionary role of miRNAs in this group. In light of this, the aim of this chapter is to attempt to lay down an initial framework of miRNA nomenclature that will both readily allow for the transfer of our understanding of the role of miRNAs in angiosperms to other plant divisions such as conifers, and to improve our understanding of miRNAs generally.

The objectives of this chapter are:

- To validate existing annotations of miRNAs in angiosperms, using the model species *Arabidopsis thaliana*.
- To use this validated dataset to establish methodology for accurately identifying orthologues in non-model species, and demonstrate the usefulness of this for the transfer of knowledge to non-model species.
-

We find that our automated validation pipeline for *A. thaliana* miRNAs is efficient in discriminating between miRNAs and other types of small RNA that have erroneously been annotated as miRNAs. Using this dataset we show that reconstruction of the gene trees of the major plant miRNA families fails to allow for any inference of orthology between members of the families in different species, due to the large time scales involved. However, using BLAST to infer homology on smaller timescales is found to yield better results, and we demonstrate the manner in which this may be used within the Brassicaceae to transfer knowledge from the model species *A. thaliana* to other important crop species.

4.2 Introduction

MicroRNAs (miRNAs) are a class of non-coding regulatory small RNAs that occur in the genomes of several eukaryotic lineages and operate at the post-transcriptional level by mediating gene silencing. Since their discovery in 1990 in the nematode *Caenorhabdis elegans* (Lee *et al.* 1993; Wightman *et al.* 1993) it has been learnt that they are ubiquitous elements in plant transcriptomes, with a diverse range of regulatory roles in development (German *et al.* 2008; Rubio-Somoza & Weigel 2011), disease resistance, abiotic and biotic stress response (Kruszka *et al.* 2012; May *et al.* 2013; Shriram *et al.* 2016) and reproduction. This diverse range of roles make them potential leads when looking at mechanisms of local adaptation, particularly as miRNAs have been shown to play important roles in stress tolerance in plants (Rajwanshi *et al.* 2014), by countering environmental stresses by altering gene expression. Differential expression of miRNAs have been observed in conifers across a range of tissues e.g. the needles, stems, and roots of *Pinus densata* (Qiu *et al.* 2016). In *Pinus taeda* miRNAs have been shown to be differentially expressed in two stages of the male gametophyte of *P. taeda*, the mature (ungerminated) and germinated pollen (Quinn *et al.* 2014), possibly indicating that they play a role in development and potentially adaptation (Yakovlev *et al.* 2010). Additionally, the capacity of microRNAs to regulate large numbers of genes on the post-transcriptional level through the suppression of protein translation has led them to be identified as important potential targets for bioengineering with the aim of improving global food security (Franco-Zorrilla *et al.* 2007; Shriram *et al.* 2016), a pressing issue due to the demands of a rapidly growing global population and concerns over the impact of climate change (May *et al.* 2013). miRNAs have already been studied and transgenically altered (Franco-Zorrilla *et al.* 2007) in ways which demonstrate that their role as crucial regulators of large numbers of plant genes may be exploited to alter important developmental traits or increase yield, not least because of the observation that expression levels react to a wide range of abiotic stresses in tissue dependent ways (Karlsson *et al.* 2015).

The vast majority of work done on miRNAs is in species that have sequenced genomes, particularly model species, and there has been relatively little work done on non-model species, particularly non-angiosperm species such as conifers. This stymies the study of evolutionary topics related to miRNAs within the conifers, and suggests that initial investigations should draw heavily on work that has already been done in other model species. Such an approach underlies many bioengineering strategies, to transfer functional knowledge gained in model species to non-model species (Johnson and Phillips. 1995) and exploit our deeper understanding of model systems. It has long been recognised that a common language for annotation is essential to allow for the identification of functional conservation (Ashburner *et al*. 2000). With miRNAs, this has been already attempted with some success in animals (Fromm. *et al* 2015), where orthologues from all the major miRNA families have been reliably identified and annotated with the intention of aiding evolutionary studies. Such an evolutionary perspective is important since the most powerful ontology is one that reflects the evolutionary history of the gene family; for example, a gene name that reveals a common origin for a gene present in several lineages allows for the ready identification of functional conservation or divergence in those lineages.

Currently, miRNAs are assigned to gene families according to an inferred common evolutionary origin from a single ancestor gene i.e. all miRNA genes in a family are paralogues. The names of individual miRNAs contain this family name along with a lower case alphabetic suffix to signify the paralogue identity within the miRNA family e.g. MIR156a, MIR156b are two different paralogues within the MIR156 family. This scheme of homology works well for identifying which miRNA genes belong to which families. However, the suffix is assigned in the order that each miRNA is identified in each species, meaning that this component of the gene name does not provide an adequate scheme of orthology for genes across species i.e. orthologues genes in two different species may have different names. So, for example, MIR156a in Species 1 may be the orthologue of MIR156g in a different species, meaning any functional knowledge gained in species 1 cannot be easily transferred to species 2 for this miRNA. miRNAs function on the level of the gene and not family, since small sequence differences in the mature sequence of paralogues within a family can lead to different genes being targeted, meaning that paralogues within a family can have radically different functions.

Indeed, many studies have demonstrated differential expression between paralogues within families, and that individual paralogues have specific roles distinct to other members of the same family (e.g. Thatcher *et al.* 2015). However, it is difficult for this information to be transferred to other species without a nomenclature that readily allows for the identification of orthologous.

It is unclear if it is possible to accurately identify orthologues miRNAs between different plant species at large phylogenetic distances. The primary issue is that the most highly conserved section of sequence in a miRNA precursor is the mature sequence (the part of the miRNA gene that facilitates targeting of mRNA), but this has a length of only around 21 nucleotides, and are often identical or close to identical between paralogues within a family. Using this sequence alone it is impossible to distinguish between precursors containing similar or identical mature sequences. Other than the miRNA* strand (the largely complementary sequence on the opposite arm of the precursor to the mature sequence), the rest of the precursor sequence demonstrates only weak levels of purifying selection and over large timescales rapidly demonstrates rapid sequence divergence, again making inference of homology difficult.

A further significant issue that obstructs the functional comparative analyses of miRNAs are widespread inconsistencies and errors in annotation (Ambros *et al.* 2003). The results of this are that large numbers of incorrectly identified miRNAs may erroneously be identified as targets for bioengineering, evolutionary studies are hampered and the impact of "species specific" miRNAs is exaggerated. The deficiencies in the quality and consistency of miRNA annotations are now well documented across animals, plants and phaeophyta (Taylor *et al.* 2014, 2017; Meng et al. 2012; Fromm et al. 2015). As discussed previously, the reason that so many loci are incorrectly identified as miRNAs is due to a failure to adhere to existing annotation criteria. Specifically, the presence of a miRNA* strand that is indicative of dicer processing, an RNA folding structure that is incompatible with dicer processing, and a lack of discrimination between heterogeneously processed siRNAs and miRNAs.

This chapter has two main objectives; the validation of existing annotation of miRNAs in the main model plant *A. thaliana,* and then inference of orthologues miRNAs previously identified in other lineages. Two different approaches are taken to establish orthology. First, known *A. thaliana* miRNAs from the major miRNA families are aligned and the gene trees of these families are inferred, with the intention of adding miRNAs from phylogenetically distant lineages such as conifers, to infer orthologues relationships by virtue of the structure of these gene trees. Second, on a smaller timescale, we establish a consistent scheme of gene level orthology in species from the Brassicaceae family, by inferring homology through the use of BLAST. The rationale of choosing Brassicaceae as the system of study is the presence of the intensively studied model species *Arabidopsis thaliana* along with many crop species such as *Camelina sativa* (an oilseed crop), *Brassica Rapa* (turnip and cabbage), and *Brassica juncea* (mustard) that have had their genomes sequenced. Inference of orthology between the well-studied miRNAome of *A. thaliana* and crop species within Brassicaceae will allow for the practical transference of functional knowledge of miRNAs in *A. thaliana* to many crop species. Ultimately this approach can be extended across the plant kingdom to encompass conifers (the main group of study in this thesis), but as a starting point in this chapter we are exploring methodologies of doing so, and the potential power of doing so, and restrict outselves to a relatively small taxonomic group.

Additionally, the power of adopting an evolutionary perspective for identifying candidate genes for bioengineering is emphasised here. The miRNAome of *Capsella grandiflora* is identified by homology with *A. thaliana* miRNAs and based on this the miRNAome of *Capsella rubella* is predicted based on its phylogenetic position; specifically, *C. rubella* should contain most or all the miRNAs present in *C. grandiflora*, and should share a proportion of the losses observed in *C. grandiflora*. This is later verified by independently identifying the miRNAome in *C. grandiflora*.

4.3 Methods

4.3.1 Validation of known miRNAs in *A. thaliana*

A significant issue that obstructs the functional comparative analyses of miRNAs are widespread inconsistencies and errors in annotation. The result of this are that large numbers of incorrectly identified miRNAs may erroneously be identified as targets for bioengineering, evolutionary studies are hampered and the impact of "species specific" miRNAs is exaggerated. The deficiencies in the quality and consistency of miRNAs annotations are now well documented across animals, plants and phaeophyta (Meng *et al.* 2012: Taylor *et al.* 2014, 2017). Therefore, prior to identifying orthologues of *A. thaliana* miRNAs in other species, it is necessary to verify the set of previously annotated miRNAs in *A. thaliana*, and remove any that do not fulfil the miRNA annotation criteria. The reasons that so many loci are incorrectly identified as miRNAs is due to a failure to adhere to existing annotation criteria. Specifically, the following conditions must be true for a miRNA to be confidently annotated (for a visual representation of this criteria see **Figure 4.1**: A visual representation of the criteria for miRNA validation. Taken from Taylor *et al.* (2014).

- the presence of a miRNA* strand that is indicative of dicer processing,
- an RNA folding structure that is incompatible with dicer processing i.e. a "hairpin" shape
- Most or all of the sRNA sequences mapping to the precursor sequence must map to the position of the "mature" (i.e. functional) or "star sequence.

A previous reanalysis (Taylor *et al.* 2014) of miRNAomes in the plant kingdom adopted the labour-intensive approach of checking the validity of each miRNA locus manually, largely as a response to the failings of many existing tools to effectively discriminate between miRNAs and other types of small RNAs. Our method is to implement the same pipeline bioinformatically, and to use publicly available small RNA data to validate all previously annotated *A. thaliana* miRNAs.

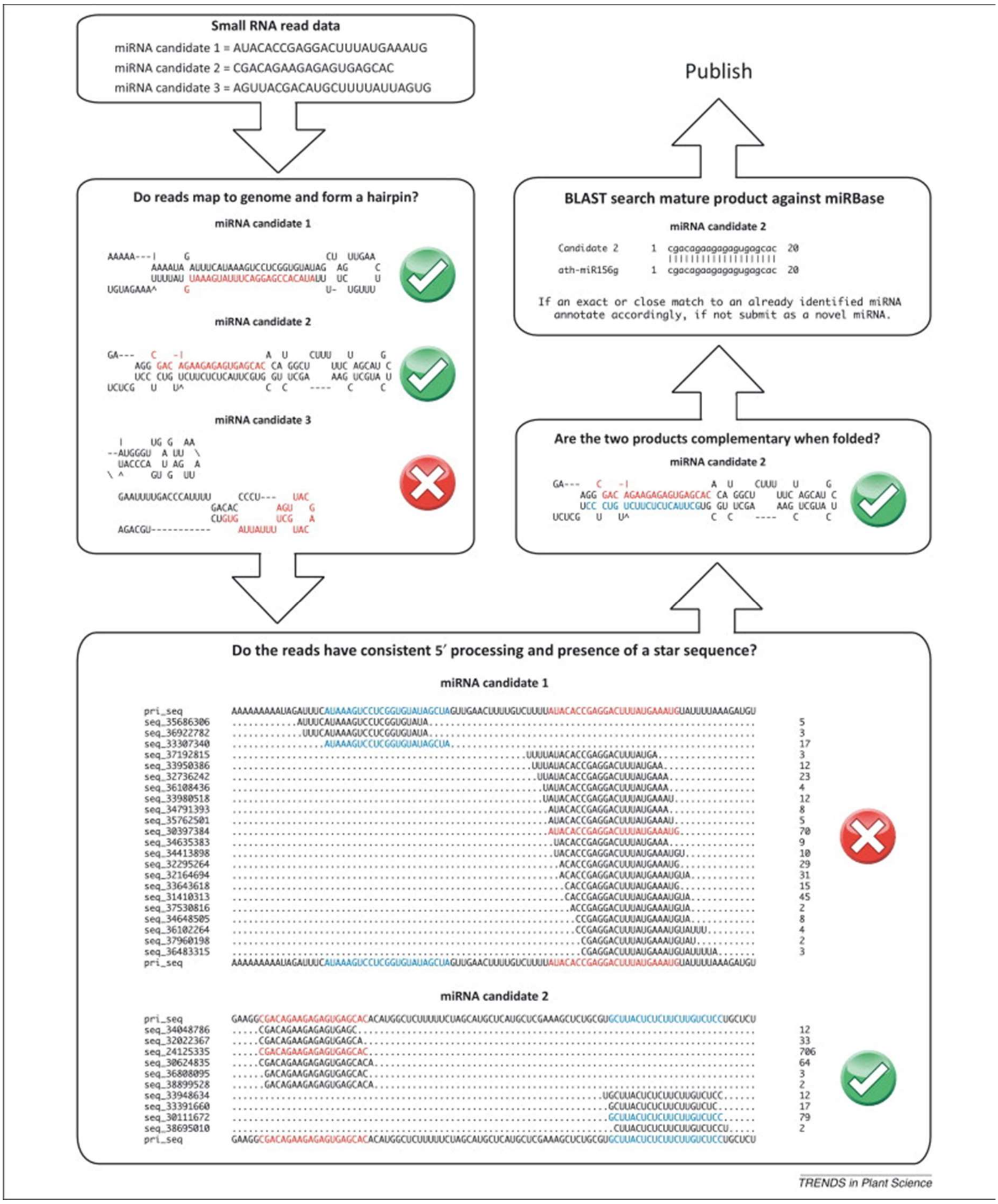

Small RNA read data
miRNA candidate 1 = AUACACCGAGGACUUUAUGAAAUG
miRNA candidate 2 = CGACAGAAGAGAGUGAGCAC
miRNA candidate 3 = AGUUACGACAUGCUUUUAUUAGUG
Do reads map to genome and form a hairpin?
miRNA candidate 1
AAAAA---| G CU UUGAA
AAAAUA AUUUCAUAAAGUCCUCGGUGUAUAG AG C
UUUUAU UAAAGUAUUUCAGGAGCCACAUAUU UC U
UGUAGAAA^ G U- UGUUU
miRNA candidate 2
GA--- C -| A U CUUU U G
AGG GAC AGAAGAGAGUGAGCAC CA GGCU UUC AGCAU C
UCC CUG UCUUCUCUCAUUCGUG GU UCGA AAG UCGUA U
UCUCG U U^ C C ---- C C
miRNA candidate 3
| UG G AA
--AUGGGU A UU \
UACCCA U AG A
\ ^ GU G UU
GAAUUUGACCCAUUUU CCCU--- UAC
GACAC AGU G
CUGUG UCG A
AGACGU---------- AUUAUUU UAC
Publish
BLAST search mature product against miRBase
miRNA candidate 2
Candidate 2 1 cgacagaagagagugagcac 20
ath-miR156g 1 cgacagaagagagugagcac 20
If an exact or close match to an already identified miRNA annotate accordingly, if not submit as a novel miRNA.
Are the two products complementary when folded?
miRNA candidate 2
GA--- C -| A U CUUU U G
AGG GAC AGAAGAGAGUGAGCAC CA GGCU UUC AGCAU C
UCC CUG UCUUCUCUCAUUCGUG GU UCGA AAG UCGUA U
UCUCG U U^ C C ---- C C
Do the reads have consistent 5′ processing and presence of a star sequence?
miRNA candidate 1
pri_seq AAAAAAAAAUAGAUUUCAUAAAGUCCUCGGUGUAUAGCUAGUUGAACUUUUGUCUUUUAUACACCGAGGACUUUAUGAAAUGUAUUUUAAAGAUGU
seq_35686306 AUUUCAUAAAGUCCUCGGUGUAUA 5
seq_36922782 UUUCAUAAAGUCCUCGGUGUAUA 3
seq_33307340 AUAAAGUCCUCGGUGUAUAGCUA 17
seq_37192815 UUUUAUACACCGAGGACUUUAUGA 3
seq_33950386 UUUAUACACCGAGGACUUUAUGAA 12
seq_32736242 UUAUACACCGAGGACUUUAUGAAA 23
seq_36108436 UAUACACCGAGGACUUUAUGAAA 4
seq_33980518 UAUACACCGAGGACUUUAUGAAAU 12
seq_34791393 AUACACCGAGGACUUUAUGAAA 8
seq_35762501 AUACACCGAGGACUUUAUGAAAU 5
seq_30397384 AUACACCGAGGACUUUAUGAAAUG 70
seq_34635383 UACACCGAGGACUUUAUGAAA 9
seq_34413898 UACACCGAGGACUUUAUGAAAUGU 10
seq_32295264 ACACCGAGGACUUUAUGAAAUG 29
seq_32164694 ACACCGAGGACUUUAUGAAAUGUA 31
seq_33643618 CACCGAGGACUUUAUGAAAUG 15
seq_31410313 CACCGAGGACUUUAUGAAAUGUA 45
seq_37530816 ACCGAGGACUUUAUGAAAUGUA 2
seq_34648505 CCGAGGACUUUAUGAAAUGUA 8
seq_36102264 CCGAGGACUUUAUGAAAUGUAUUU 4
seq_37960198 CGAGGACUUUAUGAAAUGUAU 2
seq_36483315 CGAGGACUUUAUGAAAUGUAUUUUA 3
pri_seq AAAAAAAAAUAGAUUUCAUAAAGUCCUCGGUGUAUAGCUAGUUGAACUUUUGUCUUUUAUACACCGAGGACUUUAUGAAAUGUAUUUUAAAGAUGU
miRNA candidate 2
pri_seq GAAGGCGACAGAAGAGAGUGAGCACACAUGGCUCUUUUUCUAGCAUGCUCAUGCUCGAAAGCUCUGCGUGCUUACUCUCUUCUUGUCUCCUGCUCU
seq_34048786 CGACAGAAGAGAGUGAGC 12
seq_32022367 CGACAGAAGAGAGUGAGCA 33
seq_24125335 CGACAGAAGAGAGUGAGCAC 706
seq_30624835 CGACAGAAGAGAGUGAGCACA 64
seq_36808095 GACAGAAGAGAGUGAGCAC 3
seq_38899528 GACAGAAGAGAGUGAGCACA 2
seq_33948634 UGCUUACUCUCUUCUUGUCUCC 12
seq_33391660 GCUUACUCUCUUCUUGUCUC 17
seq_30111672 GCUUACUCUCUUCUUGUCUCC 79
seq_38695010 CUUACUCUCUUCUUGUCUCCU 2
pri_seq GAAGGCGACAGAAGAGAGUGAGCACACAUGGCUCUUUUUCUAGCAUGCUCAUGCUCGAAAGCUCUGCGUGCUUACUCUCUUCUUGUCUCCUGCUCU
TRENDS in Plant Science

Figure 4.1: A visual representation of the criteria for miRNA validation. Taken from Taylor *et al.* (2014).

To compile a dataset of sRNA reads in *A. thaliana* to validate the existing *A. thaliana* miRNAs annotations, the Gene Expression Omnibus (**www.ncbi.nlm.nih.gov**) was used to collate all studies which involved non-coding RNA profiling by high throughput sequencing. Galaxy (Afgan *et al.* 2016) was then used to process these small RNA data: all separate sRNA datasets were concatenated; all reads smaller than 18 nucleotides or longer than 25 nucelotides were removed; all duplicate reads were counted and collapsed to a single read with the header containing information on the number of duplicates. The datasets used include small microRNA libraries of Arabidopsis in all developmental stages, exposed to different physiological conditions as well as from all different tissues to account for as many scenarios as possible that MicroRNA could be expressed. To obtain the sequences of known A. thaliana miRNA, all annotations present on miRBase v.20 (Kozomara & Griffiths-Jones 2011, 2014) were downloaded.

The method for assessing the validity of miRNA loci was as follows: The folding structure of all precursor sequence was estimated using Vienna RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi). All sRNA reads in our dataset were mapped to these miRBase precursor sequences using Bowtie (Langmead *et al.* 2009). The most common read that mapped to each precursor was designated the "mature" sequence, and based on the folding structure given by RNAfold, the position of the "star" sequence on the opposing arm was calculated. The sRNA dataset was checked for the presence of the "star" sequence. To calculate the processing precision of each miRNA, the number of reads that map to within a 2nt of the mature and star sequence was calculated. Nowhere in the literature is a precise number given for how precise the processing of a miRNA need be to be classified as a miRNA, largely because any such number would be arbitrary and ignores the substantial "grey area" that can exist between siRNAs and miRNAs. The analysis here chooses a figure of 85% of reads mapping to either the mature product or the star product of a miRNA, which is an objective if arbitrary condition. Therefore, there is some degree of ambiguity when classifying miRNAs, and those close to this precision level cannot be confidently excluded as microRNAs.

The presence in the sRNA library of an exact match for the putative star sequence is also checked. The identification of a star sequence is required for confident annotation of a novel miRNA family, but if the family is already known to be genuine this condition is relaxed. To check for this, the mature sequence is blasted against the given database (a verified version of miRBase v.20 here) to check if the family is known and valid. If it is known then the presence of a star sequence is unnecessary. If the family is previously unidentified then the requirement for the presence of a star sequence is required. The minimum number of nucleotides required to be bound to the opposite arm is set at 15 to qualify as a genuine miRNA. The automated pipeline utilises the additional packages Bowtie, Vienna RNAfold and BLAST and was scripted in Python. A database in fasta format is required to check for conservation of predicted miRNAs against previous annotations (Mirbase v.20 was used here). As input the package requires a file in fasta format containing precursor sequences to be checked and a file containing small RNA data in fasta format. The output is a text file showing the mapping of reads to each precursor sequence and summary statistics for processing precision and structure, and a .csv file containing the validity of each precursor sequence. The final output includes all the above information in text format and abbreviated .csv format, along with a visual mapping of the sRNA sequence to the precursor and predicted locations of the mature and star sequences.

4.3.2 Inference of miRNA orthology between species by constructing gene trees

Once the miRNA candidates were established as genuine in *Arabidopsis thaliana* they were categorised into families of paralogues and their genome positions were taken from Mirbase and used to extract the flanking region of 1000 base pairs at either side of the mature microRNA sequences. Each family was grouped with their respected 1000bp flanking regions and MAFFT (Katoh & Standley 2013) was used to align them in a phylyp format. Previously unidentified paralogues were searched for by blasting the known mature miRNA sequences to the Arabidopsis genome, and significant hits (<1E-8) along with their respective 1000bp flanking regions were added to the know miRNAs. Once aligned, Phylobayes (Lartillot *et al.* 2009, 2013), which makes use of a Bayesian Monte Carlo Markov Chain (MCMC) sampler for phylogenetic reconstruction, was used to establish the phylogenetic intra-family relationships. The aligned microRNA families with their respective 1000bp either side were independently run in Phylobayes using a CAT_GTR model that is an infinite mixture model accounting for site-specific amino-acid or nucleotide preferences. Two separate chains were run to be able to detect convergence. In order to root the paralogue trees, the most evolutionarily distant genotyped species that contained a miRNA paralogue of that family was used, so it represented the linage of the earliest known occurrence of the paralogue of interest. The paralogues of interest were then taken from Taylor *et al*, (2014) and blasted against the genome to extract the 1000 base pairs flanking regions on either side of the mature miRNA sequence. For the families 156, 160, 166, 171 and 319 the ancestral paralogues were taken from *Physcomitrella patens*. The family 396 was taken from *Selaginella moellendorffi*. Families of 164, 167, 169, 398 and 399 were taken from *Pinus taeda*. 158 were taken from the genome of *Arabidopsis lyrata*. And 395 was taken from *Picea abies* because position of many of the 395 families in the contigs of the *Pinus taeda* genome did not allow for the extraction of the full 1000 base pairs on each side of the mature sequence. Furthermore, 167 was taken from *P. taeda* due to the fact that it was found not to be present in the genome of *P. patens* as previously reported. Trees were visualised using Trex (Alix *et al.* 2012).

4.3.3 Identification of miRNA orthologues in Brassicaceae species

Four taxa within the Brassicaceae were chosen to perform a homology search for validated *A. thaliana* miRNAs: *Arabidopsis lyrata*, *Capsella grandiflora*, *Capsella rubella*, and *Camelina sativa*. The published genomes sequences were used for each and sRNA data downloaded from the NCBI Gene Expression Omnibus Datasets were downloaded. The accession numbers for these data were: SRR1736511, SRR1736512, SRR1736513, SRR1736515, GSM518389-

GSM518396, GSM518429-GSM518432. To infer the presence of sequences orthologues to *A. thaliana* miRNAs in these other Brassicaceae, BLAST was used to search for the all validated precursor sequence in each of the genomes of the species. All hits with an e-value of less than 1E-4 where taken as valid hits, the sequences excised from the genomes and were then subjected to the validation checks outlined in section 4.3.1. miRNAs that demonstrated sufficient evidence of orthology through the BLAST search and pass the validation checks were deemed to be orthologues miRNAs to the respective *A. thaliana* miRNA. miRNAs that demonstrated evidence of orthology through the BLAST search but did not pass the validation checks were deemed to be equivocally orthologues.

4.4 Results

4.4.1 Results of miRNA validation

The validation of the miRBase v.21 *A. thaliana* miRNAs reveals 157 out of 325 *A. thaliana* miRNAs pass all the criterion necessary to be confident of their validity (table S1). All loci were present in the *A. thaliana* genome and folded into a hairpin like structure. The primary reason for failure was that fewer than 85% of reads mapped to the mature sequence or star sequence locations on the precursor, suggesting a classification of siRNA is more likely for these loci. 17 loci had processing precision of between 75% and 85% so fall into borderline cases where validity is equivocal. 7 miRNAs lacked only evidence of expression of the miRNA* strand so the validity is unclear (table S1). To benchmark these results, comparisons were made to the "high confidence" set of *A. thaliana* miRNAs on miRBase, and to the manual analysis of Taylor et al. (2014) (table S1). There was a high level of concordance between the three analyses with 70 miRNAs being classed as genuine in each rating system, which corresponds to 91% of the loci given high confidence status in miRBase. 19 loci that were deemed to be valid in chapter 2 analysis were revealed to lack the necessary processing precision here, 13% of the total deemed valid. A further 30 loci were found to satisfy all the conditions that were not deemed valid in either the miRBase assessment or that of Taylor *et al.*

2014, either due to additional information on processing here or evidence of the miRNA strand expression is present here.

4.4.2 Development of a miRNA ontology within the Brassicaceae

The homology search of *A. thaliana* miRNAs in the genomes of *A. lyrata, C. sativa* and *C. rubella* resulted in 128 orthologues of 158 *A. thaliana* miRNAs being found in all four lineages, and 138 of the 158 miRNAs evolved prior to the divergence of the *Capsella/Camelina* lineage from the *Arabidopsis* lineage (**Figure 4.2**). Two miRNAs were inferred to be lost in *A. lyrata*, seven miRNAs were lost in *C. grandiflora* and none in *C. sativa*. Only 6 of the 158 *A. thaliana* miRNAs were found to be specific to *A. thaliana*. After mapping sRNA data onto the precursor sequences in each lineage, it was found that 40 miRNAs displayed processing inconsistent with annotation as miRNAs in at least one lineage. 24 of these were in *C. sativa*, 5 of these were in *C. grandiflora* and 11 of these were in *A. lyrata*. A consistent miRNA ontology requires the use of a nomenclature that accurately reflects orthology between the different species. For the purposes of this chapter, the orthologues of *A. thaliana* miRNAs identified in the other lineages are assigned the name previously assigned to the orthologues miRNA in *A. thaliana* loci on miRBase v. 21. In the case of miRNAs identified in *A. lyrata*, this involves renaming miRNA previously allocated a different name on miRBase (that did not reflect orthology). No miRNAs have been previously identified in *C. sativa* and *C. rubella* so each of the names assigned here are novel. The justification for this system of miRNA naming is primarily that because *A. thaliana* is the most studied plant species with perhaps the most completely annotated miRNA repertoire, and this chapter focused on demonstrating the potential for the transfer of knowledge of miRNA genes in *A. thaliana* to the other lineages. Should this work be extended beyond the Brassicaceae to the plant kingdom more generally, a new nomenclature would probably be necessary to avoid confusion with work done in other studies using official miRBase names. A phylogenetically informed prediction of the microRNAome of *C. rubella* was made to corroborate the validity of the homology search; specifically that the microRNAome of *C. rubella* should consist of most or all the miRNAs identified in sister lineage *C. grandiflora*, and should share a proportion of the losses observed in this lineage. This prediction was confirmed as correct, with the homology search of *A. thaliana* miRNAs in *C. rubella* revealing that all miRNAs present in *C. grandiflora* are present in *C. rubella* and 6 out of the 7 secondary losses also being shared, suggesting these losses occurred in the common ancestor of both these lineages (Figure 4.2)

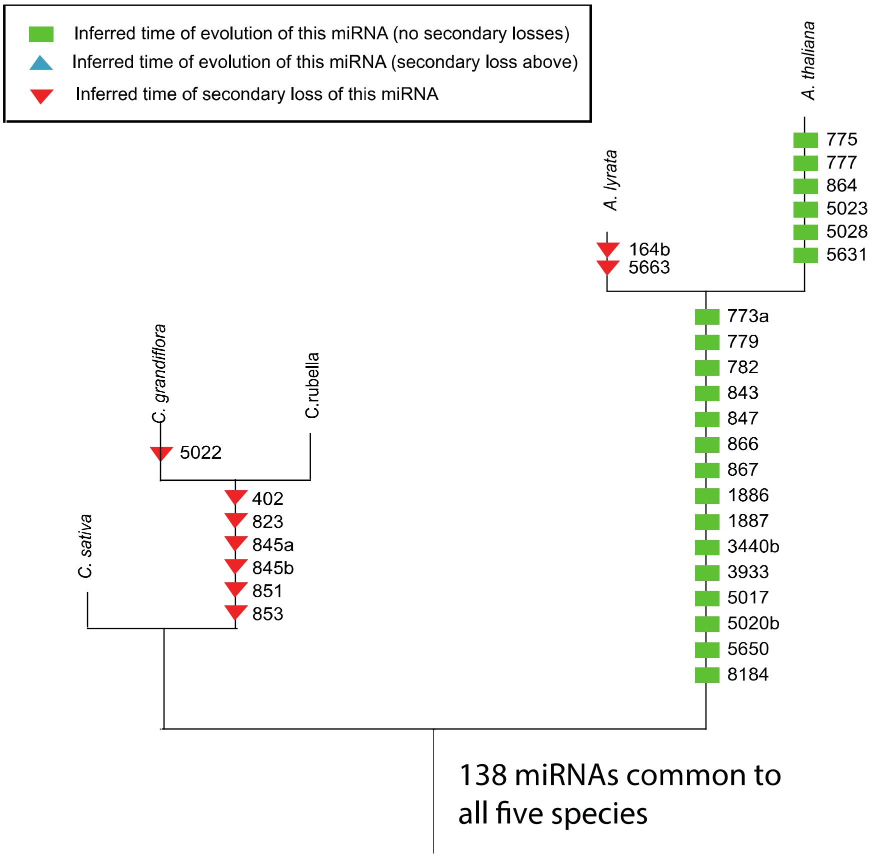

🟩 Inferred time of evolution of this miRNA (no secondary losses)
▲ Inferred time of evolution of this miRNA (secondary loss above)
▼ Inferred time of secondary loss of this miRNA
A. thaliana
775
777
864
5023
5028
5631
A. lyrata
164b
5663
773a
779
782
843
847
866
867
1886
1887
3440b
3933
5017
5020b
5650
8184
C. grandiflora
C.rubella
5022
402
823
845a
845b
851
853
C. sativa
138 miRNAs common to all five species

Figure 4.2: The miRNAome of C. rubella was predicted to share most or all of the miRNAs present in *C. grandiflora* and share a proportion of the lineage specific gains or losses. After conducting a homology search of *A. thaliana* miRNAs in the *C. rubella* genome this is shown to be the case. The numbers correspond to the family names on Mirbase.

4.4.3 Gene tree analysis of miRNA families

The phylogenetic analysis done in Phylobayes (Lartillot *et al.* 2013) show that there is not enough confidence in the results to be able to establish a reliable homology. For the miRNA families analysed in PPhylobayes, (those with three or more members) 156, 158, 160, 164,166, 167,169,171,319,396, 398 and 399 convergence was reached, with the largest discrepancy (max driff) threshold below 0.1. However, all the trees contained low support values, with very few branch support values above 60% certainty (e.g. Figure 4.3 and Figure 4.4 show two typical example for families MIR156 and MIR169). The fact that the convergence was good but the branch confidence was low for the Phylobayes analysis, indicated that this methodology has been done correctly but yielded insufficiently clear results to establish a confident orthology. So, the information conferred in these trees although interesting in some families is not sufficient to establish a naming system of orthologues, because adding in potentially orthologues sequences from other species is unlikely to result in stable trees given the analysis in *A. thaliana* did not produce stable trees. This negative result convinced the authors that to proceed Blasting the known microRNAs within the Brassicaceae (section 4.4.2) would be the best course of action in an attempt to demonstrate the potential uses for establishing orthology.

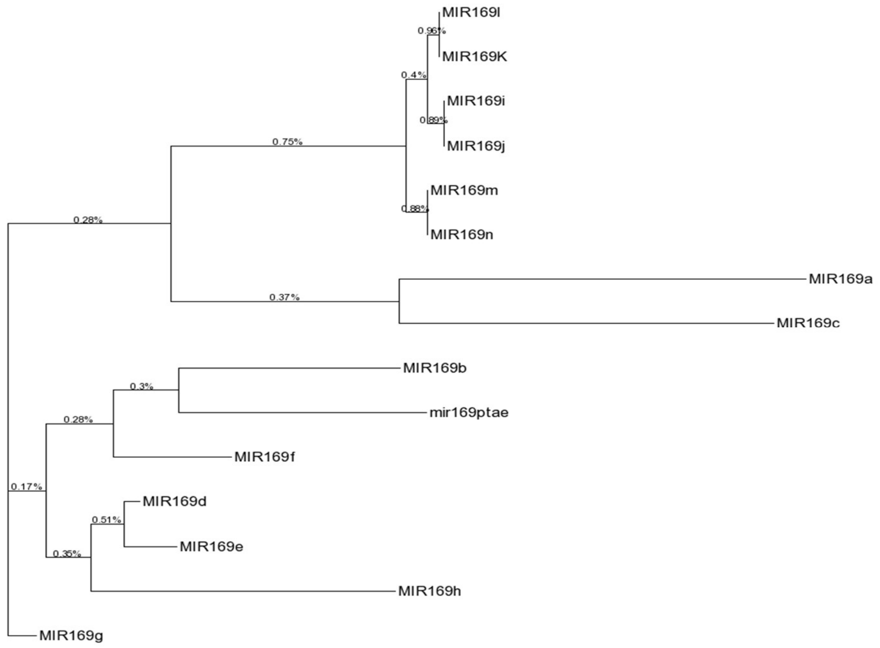


Figure 4.3: The phylogenetic analysis of family 169 from *Arabidopsis thaliana*, routed using the most distant relative available Mir169 of *Pinus taeda*.

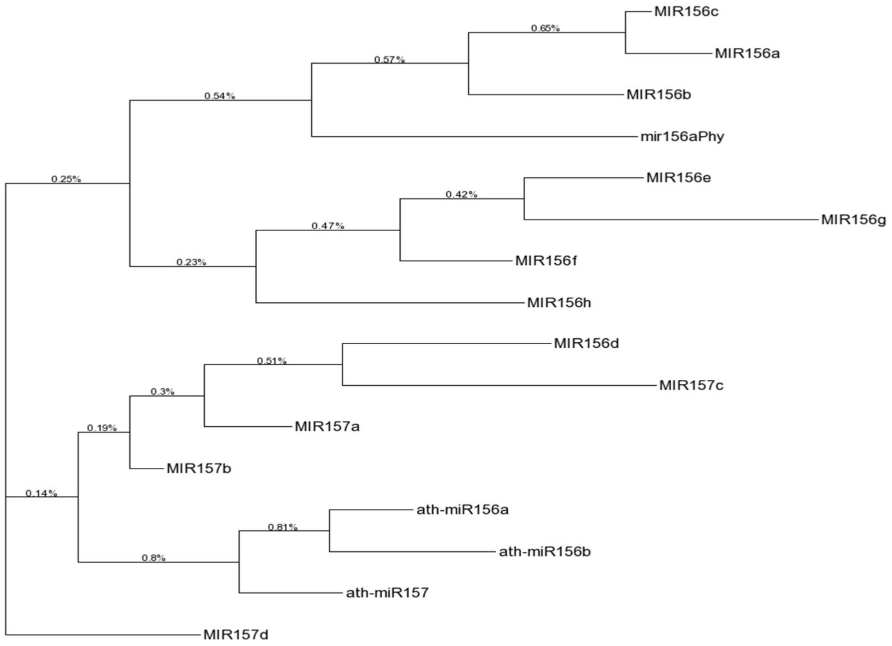


Figure 4.4: The phylogenetic analysis of family 156 (Some are erroneously annotated as 157 in Mirbase), rooted using MIR156 from *Physcomitrella patens*.

4.5 Discussion

4.5.1 An effective and objective method for the validation of miRNAs

It is clear that there are significant issues concerning the accuracy of miRNA annotation (Ming *et al.* 2012; Taylor. *et al* 2014, 2017). Many of these issues arise from the lack of discrimination of siRNAs from miRNAs. Standard experimental validation of miRNAs such as Northern blots and real-time reverse transcription polymerase chain reaction (RT-PCR) are commonly performed and presented as evidence supporting annotation of miRNAs but these do not confer any additional information about processing precision. The only practical methodology for determining processing precision is through NGS data mapped to the putative precursor sequence, but many bioinformatic packages do not give sufficient emphasis to this aspect of annotation. Likewise, presence/absence of the miRNA* strand is often ignored by bioinformatic software, despite this being a necessary criterion for the confident identification of novel miRNA families. As the quantity and quality of data used increases, so does our confidence in the annotation of miRNAs. All available data on the sRNA section of the NCBI Gene Expression Omnibus was used to assess the validity of *A. thaliana* miRNAs and, combined with the objective nature of the analysis, this gives the most comprehensive list of genuine *A. thaliana* miRNAs currently available. This updates the knowledge from the miRBase "high confidence" set of miRNAs, which by its nature is conservative and refrains from passing judgement on the invalidity of annotated miRNAs, and the manual analysis in Taylor *et al*. (2014), both of which use substantially smaller datasets. Over time such re-analyses can be performed on other model species as additional data is generated, and we can enter a virtuous cycle of reassessment and refinement of our knowledge. It is crucial that such a high-quality list of miRNAs exists, so that functional studies focus on the most promising candidates for bioengineering.

4.5.2 Reconstruction of gene trees for each miRNA family

Although the direct phylogenetic approaches used at the start of this study stated that there are some interesting results in vast kingdom wide phylogenetic approaches, it is best to start at a family level and progress by blasting microRNA in closely related species. The considerable predictive power of adopting an evolutionary perspective when describing miRNAomes is rarely used. The high level of conservation of miRNAs (once integrated into the gene regulatory network they are rarely lost) allows surprisingly accurate *a priori* predictions of miRNAomes, identifying miRNAs that have potentially been missed by annotation software or, at least as importantly, have been secondarily lost in a particularly lineage. The prediction and later verification of the miRNAome of *C. rubella,* based on the miRNAome of the nearest sister species, *C. grandiflora,* demonstrates both that the identification of homologous genes in *C. grandiflora* was accurate and that adopting an evolutionary perspective can allow for the accurate prediction of those microRNAs that have been lost in a lineage - *C. rubella* shared six of the seven losses observed in *C. grandiflora.* The identification of such losses is only possible by adopting an evolutionary perspective, and any losses identified in this manner may potentially be reinserted into the genomes should the function knowledge in model species suggest this may be fruitful.

4.5.3 Can an orthology be established across a taxonomic family?

Using a standard BLAST search to detect for orthologues miRNAs, followed by validation of these genes as miRNAs using sRNA data was an effective method for establishing homology. There are two competing hypotheses surrounding the results we should expect from such a homology search. The first conventional belief, is that most of miRNAs are species specific and undergoing a rapid birth death process (Cuperus *et al.* 2011), and therefore should not be found in sister lineages. Alternatively, it can be argued that because miRNAs are highly conserved and rarely lost we should therefore expect that the clear majority of miRNA loci in *A. thaliana* are to be found in closely related sister species. It is clear that in this case the latter is true as only six miRNAs are specific to *A. thaliana* and the vast majority (87%) evolved prior to the common ancestor of *Arabidopsis, Camelina* and *Capsella*.

It is unclear the extent to which this method of inferring homology can be expanded beyond the taxonomic family level. This is due to only small sections of miRNA precursor sequences being highly conserved and there are significant levels of sequence divergence at greater phylogenetic distances, along with complex duplication and loss histories, this explains somewhat why our kingdom-wide phylogenetic approach did not work. However, within the family level developing such a scheme of homology is both possible and productive.

Interestingly, the presence of an orthologue sequence in another genome did not guarantee that this potential miRNA would display canonical miRNA processing precision in the lineage it was identified in. A total of 40 miRNAs that are precisely processed in *A. thaliana* were found not to exhibit the same degree of processing precision in other taxa. There are two possible explanations for this. First, the miRNAs maybe relatively young and are not yet fully integrated into the gene regulatory network, therefore are more susceptible to loss, and are examples of young miRNAs undergoing the rapid birth/death process mentioned earlier. Alternatively, these miRNAs may have specific roles in responding to particular biotic or abiotic conditions and such conditions were not present when the sRNA sequencing was conducted.

4.6 Conclusions

4.6.1 Potential Application to bioengineering

The fundamental motivation for pursuing a scheme of gene level orthology is that functional knowledge is then easily transferable from miRNAs in model species to crop species. *C. sativa* is grown for oilseed and is an emerging biofuel crop and as such there is likely to be substantial interest in future bioengineering of this crop. The evolutionary framework used to convey homology (Figure 4.2) provides a basis for selecting which miRNAs to focus bioengineering studies on. While no miRNAs are seen to be lost in the *C. sativa* lineage (Figure 4.2) miRNA exhibit variable processing precision relative to *A. thaliana*. This must be either due to these miRNAs being under little selective pressure due to not being part of the gene regulatory network, or because they are involved in reactions to biotic or abiotic stresses which were not present at the time of sequencing. In either case they represent good targets for bioengineering. Of these 24 miRNAs, at least 13 have been observed to respond to abiotic stress in *A. thaliana* variety of salinity, copper, temperature and drought stresses (Khraiwesh *et al.* 2012; May *et al.* 2013; Shriram *et al.* 2016). This high proportion of genes, which both exhibit poor processing precision in *C. sativa* and respond to abiotic stresses, strongly suggest that selecting miRNAs for intensive study for bioengineering based on the function of orthologues genes in to *A. thaliana*, either to improve the processing precision or up regulate/down regulate their expression, may prove fruitful.

4.6.2 The power of phylogenetic inference

The considerable predictive power of adopting an evolutionary perspective when describing miRNAomes is rarely used. The high level of conservation of miRNAs allows surprisingly accurate a priori predictions of miRNAomes, identifying miRNAs that have potentially been missed by annotation software or, at least as importantly, have been secondarily lost in a lineage. The prediction and later verification of the miRNAome of *C. rubella*, based on the miRNAome of the nearest sister species, *C. grandiflora*, demonstrates both that the identification of homologous genes in *C. grandiflora* was accurate and that adopting an evolutionary perspective can allow for the accurate prediction of those miRNAs that have been lost in a lineage - *C. rubella* shared six of the seven losses observed in *C. grandiflora*. The identification of such losses is only possible by adopting an evolutionary perspective, and any losses identified in this manner may potentially be reinserted into the genomes should the function knowledge in model species suggest this may be fruitful.

miRNAs act on the level of individual genes, not on the level of the family, since different miRNA within a family may target different protein coding genes. Consequently, it is of great importance to establish a consistent scheme of homology that allows for transfer of knowledge from model species to crop species. Additionally, the evolution of miRNA families has been extensively studied, but such studies rarely attempt to track the evolution on the gene level. The scheme of homology developed here within the Brassicaceae allows for the immediate transfer for knowledge from the most intensively studied model plant species *A. thaliana* to a host of crop species. It is unknown whether this can be expanded across a broader evolutionary range but at the very least, this is an effective way to study the evolution of miRNA genes on the level of orders or families.

4.6.3 Potential for future work in conifers

As more conifer genomes are sequenced (Nystedt *et al.* 2013; Birol *et al.* 2013; Wegrzyn *et al.* 2014), and the quality of these genome annotations continue to improve (De La Torre *et al.* 2014), there will be a wider scope for the study of miRNAs in conifers and the role these play in adaptation. So far, the establishment of a system of proof checking genuine microRNAs has given us a further tool to analyse the newly emerging small RNA libraries from these conifers, such as has been done with the Norway spruce (Nystedt *et al.* 2013). The establishment of gene orthology with model species can only help further applications of the knowledge gained so far in other species in the study of microRNAs in conifers.

4.7 References

Afgan E, Baker D, van den Beek M *et al.* (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, **44**, W3–W10.

Alix B, Boubacar DA, Vladimir M (2012) T-REX: A web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Research*, **40**, W573–W579.

Ambros V, Bartel B, Bartel DP *et al.* (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277–279.

Ashburner M, Ball CA, Blake JA *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Birol I, Raymond A, Jackman SD *et al.* (2013) Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics*, **29**, 1492–1497.

Cuperus JT, Fahlgren N, Carrington JC (2011) Evolution and functional diversification of miRNA Genes. *The Plant Cell*, **23**, 431–442.

De La Torre AR, Birol I, Bousquet J *et al.* (2014) Insights into conifer giga-genomes. *Plant physiology*, **166**, 1724–32.

Franco-Zorrilla JM, Valli A, Todesco M et al. (2007) Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics*, **39**, 1033–1037.

Fromm B, Billipp T, Peck LE *et al.* (2015) A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annual Review Genetics*, **49**, 213–242.

German MA, Pillay M, Jeong DH *et al.* (2008) Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nature biotechnology*, **26**, 941–946.

Johnson AT, Phillips WM (1995) Philosophical foundations of biological engineering. *Journal of Engineering Education*, 311–320.

Karlsson P, Christie MD, Seymour DK *et al.* (2015) KH domain protein RCF3 is a tissue-biased regulator of the plant miRNA biogenesis cofactor HYL1. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 14096–14101.

Katoh K, Standley DM (2013) MAFFT Multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Khraiwesh B, Zhu JK, Zhu JH (2012) Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. *Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms*, **1819**, 137–148.

Kozomara A, Griffiths-Jones S (2011) MiRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, **39**, D152-7.

Kozomara A, Griffiths-Jones S (2014) MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, **42**, D68-73.

Kruszka K, Pieczynski M, Windels D *et al.* (2012) Role of microRNAs and other sRNAs of plants in their changing environments. *Journal of Plant Physiology*, **169**, 1664–1672.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics Applicatons Note* , **25**, 2286–228810.

Lartillot N, Rodrigue N, Stubbs D, Richer J (2013) PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology*, **62**, 611–615.

Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.

May P, Liao W, Wu YJ *et al.* (2013) The effects of carbon dioxide and temperature on microRNA expression in *Arabidopsis* development. *Nature Communications*, **4**.

Meng Y, Shao C, Wang H, Chen M (2012) Are all the miRBase-registered microRNAs true? *RNA Biology*, **9**, 249–253.

Nystedt B, Street NR, Wetterbom A *et al.* (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**, 579–84.

Qiu Z-B, Yuan M-M, Hai B-Z, Wang L, Zhang L (2016) Characterization and expression analysis of conserved miRNAs and their targets in *Pinus densata*. *Biologia Plantarum*, **60**, 427–434.

Quinn CR, Iriyama R, Fernando DD (2014) Expression patterns of conserved microRNAs in the male gametophyte of loblolly pine (*Pinus taeda*). *Plant Reproduction*, **27**, 69–78.

Rajwanshi R, Chakraborty S, Jayanandi K, Deb B, Lightfoot DA (2014) Orthologous plant microRNAs: microregulators with great potential for improving stress tolerance in plants. *Theoretical and Applied Genetics*, **127**, 2525–2543.

Rubio-Somoza I, Weigel D (2011) MicroRNA networks and developmental plasticity in plants. *Trends in Plant Science*, **16**, 258–264.

Shriram V, Kumar V, Devarumath RM, Khare TS, Wani SH (2016) MicroRNAs as potential targets for abiotic stress tolerance in plants. *Frontiers in Plant Science*, **7**.

Taylor RS, Tarver JE, Foroozani A, Donoghue PCJ (2017) MicroRNA annotation of plant genomes − Do it right or not at all. *BioEssays*, **39**, 1600113.

Taylor RS, Tarver JE, Hiscock SJ, Donoghue PCJ (2014) Evolutionary history of plant microRNAs. *Trends in Plant Science*, **19**, 175–182.

Thatcher SR, Burd S, Wright C, Lers A, Green PJ (2015) Differential expression of miRNAs and their target genes in senescing leaves and siliques: insights from deep sequencing of small RNAs and cleaved target RNAs. *Plant Cell and Environment*, **38**, 188–200.

Wegrzyn JL, Liechty JD, Stevens KA *et al.* (2014) Unique features of the Loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics*, **196**.

Wightman B, Ha I, Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. *Cell*, **75**, 855–862.

Yakovlev IA, Fossdal CG, Johnsen Ø (2010) MicroRNAs, the epigenetic memory and climatic adaptation in Norway spruce. *New Phytologist*, **187**, 1154–1169.

Conclusions

- We propose a framework to detect molecular footprint of selection in instances of long range colonisation taking into account important sources of false positives: shared history and gene flow, as well as gene surfing. Although the first source of false positive is included routinely in different approaches to detect markers under selection, the impact of gene surfing (that may leave the same molecular footprint as selection) during long range colonisation has been studied theoretically but more rarely empirically, and even more rarely in a scenario of adaptation.
- Our results based on genetic-environment association methods on the full distribution range of Aleppo pine and controlling for gene surfing, lead us to identify 7 SNPs potentially under selection, most of these related to drought, an environmental driver of crucial importance for the adaptation of this conifer inhabiting the Mediterranean Basin.
- At the local scale, the use of a paired sampling technique combined with the selection of various statistical methods that aid data exploration as well as adding confidence to our results, lead to the identification of 3 outlier SNPs in Aleppo pine. Altogether our results point to one SNP as a very good candidate for selection and indicate that precipitation is a potential predominant driver of selection at two distinct spatial scales.
- Our comparative studies point to the importance of taking into account several factors when interpreting the outputs of outlier tests: the evolutionary history of the species, as well as the sampling size in terms of individual per population as well as population repetitions.

- When completed, the outputs from the comparative analyses across methods and species will allow assessing the amount of divergence among populations under divergent ecological conditions, a pivotal information for the prediction of how populations may respond to future local and global environmental changes.
- MicroRNAs are becoming a key area of research when exploring adaptation, and it is of great importance to establish a consistent scheme of homology that allows transferring knowledge from model species to non-model species such as pines. The scheme of homology developed here within the Brassicaceae allows for the immediate transfer of knowledge from the most intensively studied model plant species *A. thaliana* to crop species. It is unknown for now (this work is in progress) whether this can be expanded across a broader evolutionary range, but at the very least, this is an effective way to study the evolution of miRNA genes on the level of orders or families.

Conclusiones

- Proponemos una metodología para detectar la huellas de selección molecular en instancias de colonización a larga distancia, teniendo en cuenta potenciales causantes de falsos positivos: historia genética compartida, flujo genético y el impacto del surfing genético, la primera causa de falso positivos está corregida por diferentes métodos, el impacto del surfing genético (que puede dejar la misma huella molecular que la selección) durante colonizaciones ha sido estudiado teóricamente pero raramente empíricamente, y mucho menos mirando a su efecto en adaptación.
- Nuestros resultados se basan en asociaciones genéticas-medioambientales, en la completa distribución natural del pino Carrasco, controlando por el efecto del surfing genético, nos ha llevado a encontrar 7 SNPs potencialmente bajo selección, la mayoría relacionados con sequía, un factor medio ambiental crucial para una conífera mediterránea.
- A escala local, el uso de muestreo en pares combinado con la selección de varios métodos estadísticos, nos ayudaron en explorar los datos, como también confirmar nuestros resultados, y nos llevaron a la identificación de 3 SNPs potencialmente bajo selección en el pino Carrasco. Los resultados nos indicaron un SNP en especial como muy buen candidato de estar bajo selección, y también señalan a la precipitación como uno de las presiones selectivas más importante a dos escalas espaciales.

- Una vez completados, los resultados de los análisis comparativos en varios métodos y especies, permitirán el asesoramiento de la divergencia entre poblaciones procedentes de diferentes ecomorfos, información esencial para predecir como poblaciones responderán a futuros cambios climáticos tanto locales como globales.
- Nuestros análisis comparativos nos señalan la importancia de tener en cuenta varios factores al interpretar los resultados de los tests de selección: la historia evolutiva de las especies, como también el muestreo en términos de individuos por población y repeticiones de estas.
- Los MicroRNAs es un área de investigación clave para explorar la adaptación. Es necesario desarrollar un sistema consistente para la clasificación de los mismos, para poder transferir conocimientos de especies modelos a especies no tan estudiadas como las coníferas. El sistema de clasificación basado en la homología de estas moléculas, realizado en las Brassicaceae, permite transmitir conocimientos de la especie modelo más estudiada *A.thaliana* a especies de importancia agraria. Por ahora no está claro (ya que estos trabajos siguen en proceso) si esto puede expandirse a un contexto más amplio, pero por lo menos es una manera efectiva de estudiar microRNA en el nivel de órdenes y familias.

Appendix A Supplementary Information

A.1 Chapter 2

A.1.1 Data collection.

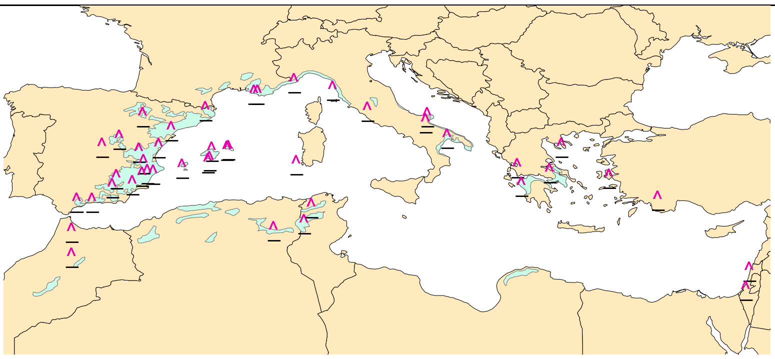


Figure S1: location of population sampled of SSRs in the Mediterranean basin. In the background highlighted the natural range of *Pinus halepensis*.

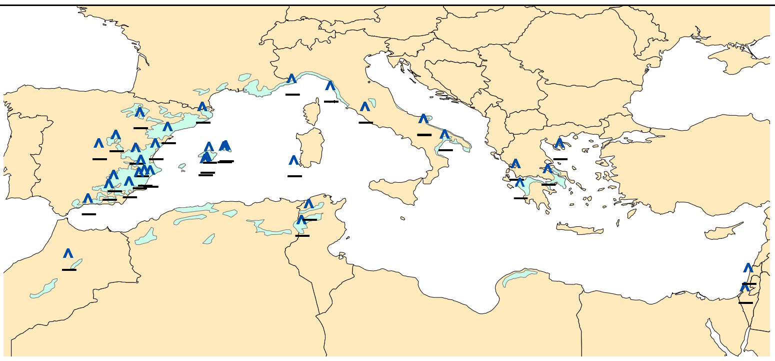


Figure S2: location of population sampled of SNP in the Mediterranean basin. In the background highlighted the natural range of *Pinus halepensis*.

Table S1: Sampling details of the 44 populations genotyped for SSR. The populations that have been genotyped for both markers are shown in Bold.

Population code	**population**	**country**	**latitude**	**longitude**
11	**Cabanellas**	Spain	42.248294°	2.783798°
21	**Tivissa**	Spain	41.059193°	0.760224°
61	**Zuera**	Spain	41.918800°	-0.921611°
83	**Alcantud**	Spain	40.564133°	-2.313436°
84	**Colmenar de Oreja**	Spain	40.090858°	-3.333642°
92	**Tuéjar**	Spain	39.819100°	-1.159188°
101	**Tibi**	Spain	38.519440°	-0.648611°
105	**Bicorp**	Spain	39.103986°	-0.858193°
111	**Benicàssim**	Spain	40.077655°	0.025914°
131	**Villajoyosa**	Spain	38.496100°	-0.303656°
142	**Monovar**	Spain	38.385360°	-0.957389°
152	**Benamaurel**	Spain	37.702100°	-2.738858°
154	**Santiago de la Espada**	Spain	38.227130°	-2.467588°
157	**Alhama de Murcia**	Spain	37.864996°	-1.534291°
173	**Frigiliana**	Spain	36.818198°	-3.920522°
182	**Palma de Mallorca**	Spain	39.149725°	2.941020°
183	**Santanyi**	Spain	39.284035°	3.047450°
184	**Alcudia**	Spain	39.872892°	3.170338°
186	**Alcotx**	Spain	39.971779°	4.168438°
187	**Atalix**	Spain	39.915380°	4.053586°
172	Carratraca	Spain	36.842800°	-4.833715°
231	**Litorale Tarantino**	Italy	40.619829°	17.116000°
232	**Gargano Marzini**	Italy	41.902422°	15.941800°
233	**Gargano Monte Pucci**	Italy	41.547383°	15.857200°
IT.IMP	**Imperia**	Italy	43.900000°	8.050000°
IT.QUARC	**Quercianella**	Italy	43.490000°	10.340000°
IT.OTRIC	**Otricoli**	Italy	42.240000°	12.380000°
IT. Carlofort	**Carlo Forte**	Italy	39.080000°	8.180000°
212	**Amfilohia**	Greece	38.883652°	21.283507°
214	**Kassandra**	Greece	40.091078°	23.881487°
GR. ELE	**Elea**	Greece	37.766667°	21.533333°
GR.EUBO	**North Eubea**	Greece	38.580000°	23.180000°
241	**Thala**	Tunisia	35.567200°	8.650593°
242	**Tabarka**	Tunisia	36.505600°	9.075704°
Ifrane	**Zaouia Ifrane**	Morocco	33.570000°	-5.140000°

ALG.AUR	**Aures Beni Melloul**	Algeria	35.166667°	6.833333°
ISR NAT	**Nat (Mont Carmel)**	Israel	32.720000°	35.030000°
ISR A6	**A6 (Shaharia)**	Israel	31.600000°	34.830000°
FBn	Font_Blanche	France	43.240000°	5.680000°
Bouc	Ouardane Bouksane	Morocco	35.050000°	-5.130000°
StM	Saint Mître	France	43.240000°	5.890000°
H1	Izmir-Urla	Turkey	38.261944°	26.713889°
H2	Muğla-Fethiye	Turkey	36.950000°	29.600000°
192	Ses Salines	Spain	38.841242°	1.398844°

Table S2: Sampling details of the 49 populations genotyped for SNP array. The populations that have been genotyped for both markers are shown in bold.

Population code	**population**	**Country**	**latitude**	**longitude**
VO_IV_011_4	**Cabanellas**	Spain	42.248294°	2.783798°
VO_IV_021_4	**Tivissa**	Spain	41.059193°	0.760224°
VO_IV_061_4	**Zuera**	Spain	41.918800°	-0.921611°
VO_IV_083_4	**Alcantud**	Spain	40.564133°	-2.313436°
VO_IV_084_4	**Colmenar de Oreja**	Spain	40.090858°	-3.333642°
VO_IV_092_4	**Tuéjar**	Spain	39.819100°	-1.159188°
VO_IV_101_4	**Tibi**	Spain	38.519440°	-0.648611°
VO_IV_105_4	**Bicorp**	Spain	39.103986°	-0.858193°
VO_IV_111_4	**Benicassim**	Spain	40.077655°	0.025914°
VO_IV_131_4	**Villajoyosa**	Spain	38.496100°	-0.303656°
VO_IV_142_4	**Monovar**	Spain	38.385360°	-0.957389°
VO_IV_152_4	**Benamaurel**	Spain	37.702100°	-2.738858°
VO_IV_154_4	**Santiago de la Espada**	Spain	38.227130°	-2.467588°
VO_IV_157_3	**Alhama de Murcia**	Spain	37.864996°	-1.534291°
VO_IV_172_4	**Carratraca**	Spain	36.842800°	-4.833715°
VO_IV_173_3	**Frigiliana**	Spain	36.818198°	-3.920522°
VO_III_182_2	**Palma de Mallor**	Spain	39.149725°	2.941020°
VO_IV_183_4	**Santanyi**	Spain	39.284035°	3.047450°
VO_III_184_4	**Alcudia**	Spain	39.872892°	3.170338°
VO_IV_186_4	**Alcotx**	Spain	39.971779°	4.168438°
VO_IV_187_4	**Atalix**	Spain	39.915380°	4.053586°
VO_IV_212_4	**Amfilohia**	Greece	38.883652°	21.283507°
VO_IV_214_3	**Kassandra**	Greece	40.091078°	23.881487°
VO_IV_231_4	**Litorale Tarantino**	Italy	40.619829°	17.116000°
VO_IV_232_4	**Gargano Marzini**	Italy	41.902422°	15.941800°

VO_IV_233_4	**Garzano Monte Pucci**	Italy	41.547383°	15.857200°
VO_IV_241_4	**Thala**	Tunisia	35.567200°	8.650593°
VO_IV_242_4	**Tabarka**	Tunisia	36.505600°	9.075704°
4R.ELE-2	**Elea**	Greece	37.766667°	21.533333°
4R.EUBO-7	**North Eubea**	Greece	38.580000°	23.180000°
ISRA6_24	**Shaharia**	Israel	31.600000°	34.830000°
ISRNAT_22	**Nat**	Israel	32.720000°	35.030000°
IT.Carlofort_24	**Carlo Forte**	Italy	39.080000°	8.180000°
IT.IMP_31	**Imperia**	Italy	43.900000°	8.050000°
IT.OTRIC_48	**Otricoli**	Italy	42.240000°	12.380000°
IT.QUARC_27	**Quercianella**	Italy	43.490000°	10.340000°
AL4.AUR-2	**Aures Beni Melloul**	Algeria	35.166667°	6.833333°
Ifrane_24	**Zaouia Ifrane**	Morocco	33.570000°	-5.140000°
Ph1V317	Calderona	Spain	39.740000°	-0,480000°
Ph2V318	Sinarcas	Spain	39.80000°	-1.20000°
Ph5A67	Eslida	Spain	39.870000°	-0.290000°
1292	Cabanes	Spain	40.100000°	0.040000°
1040	Titaguas	Spain	39.890000°	-1,300000°
1140	Serra d'Irta	Spain	40.350000°	0.320000°
1630	Montan	Spain	40.050000°	-0.590000°
1602	Alzira	Spain	39.120000°	-0,390000°
34A	Serra d'IrtaB	Spain	40.350000°	0.320000°
33B	CabanesB	Spain	40.10000°	0.040000°
33C	CalderonaB	Spain	39.740000°	-0,480000°

Environmental data collection

We characterized each SNP-genotyped population with 27 environmental variables in order to assess SNP-environment interactions. These variables were representative of the period 1950-2000 and included:

- Nineteen bioclimatic variables (BIO1-BIO19; see Table S1) available in WORLDCLIM (Hijmans *et al.* 2005);
- Four variables created following Zimmermann *et al.,* 2007 to characterize water availability better: summer and spring potential evapotranspiration (ETPTsummer ; ETPTspring), and summer and spring moisture index (MINDsummer; MINDspring). Potential evapotranspiration estimates water loss as a ratio depending on average temperature and solar radiation (obtained from Kumar *et al,.* 2007) following Turc's empirical equation(Turc 1963). Moisture index is defined as the difference between precipitation (water source) and potential evapotranspiration (water loss), so values below zero indicate drought, while positive scores indicate that precipitation exceeds potential evapotranspiration.
- Four aridity indexes (one per quarter of the year - aridityQ1-Q4). We calculated aridity as a ratio between monthly precipitation and potential evapotranspiration and then estimated mean values for each quarter, as indicated in Eckert *et al.,* 2010.

Table S3: List of bioclimatic variables BIO1-BIO19 available in WORDLCLIM (top), and list of the bioclimatic variables used in running the Bayesian linear model.

Name	**Description**
BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) (* 100)
BIO4	Temperature Seasonality (standard deviation *100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

Table S4:

Procedencia	bio1	bio2	bio3	bio4	bio5	bio6	bio7	bio8	bio9	bio10	bio11	bio12	bio13	bio14	bio15	bio16	bio17	bio18	bio19	etpt_summer	etpt_spring	mind_summer	mind_spring
Cabanellas	147	77	32	5507	274	38	236	157	219	219	79	711	96	35	24	220	147	147	151	95.66	77.33	-243.33	-171.67
Tivissa	144	92	36	5687	282	27	255	149	219	220	77	597	74	18	32	198	104	124	129	94.33	75	-253.67	-173.67
Zuera	119	103	37	5937	274	-1	275	141	196	197	45	476	62	24	24	151	94	101	106	95.33	76.33	-257.67	-185.33
Alcantud	107	116	38	6444	282	-22	304	92	193	194	30	505	65	22	26	161	87	95	119	95.33	75.33	-259.7	-176.33
Colmenar de Oreja	137	113	36	6735	318	9	309	93	227	227	56	433	52	10	36	147	49	49	124	96.66	78.66	-279	-197.67
Tu駛ar	135	104	38	6099	286	14	272	143	62	216	62	412	50	17	30	131	78	80	78	96.33	79.33	-267.67	-201
Tibi	124	108	39	5886	279	3	276	95	203	205	55	573	73	18	31	189	84	94	136	96	76.66	-265.33	-178.67
Bicorp	143	107	39	5770	291	22	269	154	219	221	74	467	65	14	34	162	70	82	103	97	79.33	-273	-198.33
Benicassim	148	87	36	5373	276	38	238	157	218	221	85	516	72	17	34	190	83	108	111	97	80.33	-268.67	-201.33
Villajoyosa	179	104	40	5237	313	59	254	156	246	249	115	418	72	6	51	175	39	65	109	98.33	82.66	-287.33	-220.67
Monovar	142	114	40	5844	297	16	281	152	219	220	72	457	62	12	35	153	61	71	106	94.67	75.67	-269	-185.33

Benamaurel	142	115	37	6371	318	15	303	95	228	229	67	452	57	8	43	155	38	44	140	97.34	80.34	-284.67	-193.67
Santiago de la Espada	136	119	38	6430	316	6	310	116	223	223	60	453	59	9	38	157	49	51	125	97.67	81.34	-282	-196.33
Alhama de Murcia	142	115	39	5959	303	15	288	110	221	222	71	422	54	9	37	139	48	54	109	96.34	79.67	-278.33	-197.67
Carratraca	159	102	39	5371	305	47	258	104	230	232	95	675	105	2	66	306	20	27	289	96	79.34	-286.67	-183.33
Frigiliana	156	107	39	5650	311	38	273	97	232	232	89	457	67	4	58	194	20	20	181	97.34	81.67	-290.67	-207.33
Santanyi	169	79	36	5028	287	69	218	148	233	237	110	561	91	5	52	238	40	78	176	97	79.67	-283	-202
Alcotx	165	71	33	5005	281	71	210	145	229	233	107	616	100	5	52	263	47	91	190	96.67	79.34	-279.67	-198.33
Atalix	166	72	34	5077	283	72	211	146	231	235	107	622	100	5	52	265	47	91	193	97	79.69	-280.67	-198.67
Amfilohia	142	106	36	6419	305	15	290	77	225	225	62	975	157	17	57	426	69	69	405	96.67	78.34	-272.33	-170
Kassandra	159	95	32	6902	317	25	292	90	244	249	73	461	63	15	37	173	60	63	158	97	78	-275.33	-200.33
Litorale Tarantino	150	89	34	5874	293	37	256	122	227	227	79	561	69	22	33	197	78	78	175	97	78.67	-270.33	-196
Gargano Marzini	161	94	35	5999	307	45	262	133	239	239	88	472	58	21	29	167	76	76	140	96.34	77.67	-269	-203.67
Garzano Monte Pucci	135	64	28	5895	261	35	226	109	212	212	65	549	67	30	27	188	95	95	162	95.34	74.67	-259.67	-189
Thala	156	130	38	6977	353	15	338	137	247	247	70	426	47	12	28	133	59	59	121	97.34	79.34	-277.67	-198

Tabarka	178	127	40	6598	361	45	316	97	264	265	97	559	82	6	52	230	37	51	230	98.34	82.34	-288	-203
Elea	166	94	37	5678	307	54	253	113	238	239	97	814	151	6	74	409	25	46	369	98	82.34	-291	-201.33
North Eubea	158	98	34	6604	318	33	285	93	245	245	77	607	97	11	55	254	47	47	248	97.34	80	-281.67	-196.67
Shaharia	199	114	47	4836	320	78	242	135	254	256	135	384	97	0	107	246	0	0	246	98.67	86.34	-301.33	-240.67
Nat	193	92	41	5008	305	83	222	129	249	253	128	646	174	0	112	434	0	2	361	98.67	86	-301.33	-229.33
Imperia	152	83	35	5250	277	46	231	128	220	222	89	808	108	18	37	289	94	120	233	95.67	76.67	-261	-164.67
Otricoli	134	96	34	6393	293	14	279	100	218	218	56	837	106	36	26	283	149	149	218	96	76.67	-243.67	-170.33
Quercianella	144	84	34	5621	278	36	242	115	217	217	75	830	110	23	35	306	111	144	230	94.34	74.34	-251.33	-162.33
Aures Beni Melloul	145	129	40	6456	324	6	318	129	230	230	64	325	45	9	33	109	43	43	82	97.67	81.34	-284	-212.33
Zaouia Ifrane	123	142	41	6396	319	-26	345	88	209	210	46	802	115	7	55	302	44	47	291	97	79	-281.67	-158
Serra Calderona	133	94	37	5689	272	21	251	143	208	210	66	512	67	20	30	176	89	104	107	95.67	76.34	-262.67	-186.33
Sinarcas	122	105	37	6173	278	0	278	139	48	205	48	459	57	21	28	149	87	93	87	96.34	77.67	-263.33	-189.67
Eslida	138	90	36	5531	271	27	244	148	210	213	72	514	69	19	32	182	87	106	109	93.67	74	-257.33	-180.67

Cabanes	145	87	36	5415	273	34	239	153	215	218	81	526	72	18	34	191	86	110	113	95	76	-261.67	-186.67
Titaguas	115	107	37	6281	275	-7	282	133	41	200	41	475	61	22	28	158	90	100	90	95	74.67	-257	-178.33
Serra d'Irta	152	88	36	5350	279	40	239	158	223	225	91	546	77	16	36	201	88	115	117	95.34	76.67	-262	-187.33
Montan	125	94	36	5852	266	10	256	133	55	203	55	486	58	22	29	159	92	105	92	95.34	76	-259	-186
Alzira	173	97	40	5197	301	60	241	187	239	243	109	466	81	9	48	191	50	80	120	97.34	80.67	-280.67	-211.33
Serra d'Irta	152	88	36	5350	279	40	239	158	223	225	91	546	77	16	36	201	88	115	117	95.34	76.67	-262	-187.33
Cabanes	145	87	36	5415	273	34	239	153	215	218	81	526	72	18	34	191	86	110	113	95	76	-261.67	-186.67
Serra Calderona	133	94	37	5689	272	21	251	143	208	210	66	512	67	20	30	176	89	104	107	95.67	76.34	-262.67	-186.33

A.1.2 Population genetics and Statistics

Fis and Fst and all tests of hardy Weinberg equilibrium (P values) where done using Genepop (version 4.3) (Raymond & Rousset 1995; Rousset 2008) Markov chain parameters for all tests Dememorization: 10000, Batches: 20 Iterations per batch: 5000. Different genetic diversity parameters were estimated across all populations using SPAGeDI (Hardy & Vekemans 2002) there included; number of alleles (*NA*), effective number of alleles (*NAe*; Nielsen *et al.* 2003), allelic richness (*Ae*), expected heterozygosity or gene diversity corrected for sample size (*He*; Nei 1978) and observed heterozygosity (Ho).

Table S4. SSRs statistics of each loci; Fis, Fst, percentage of missing genotypes, p-values for Hardy Weinberg equilibrium. A: Number of alleles. NAe: effective number of alleles (Nielsen *et al.* 2003). He: Gene diversity corrected for sample size (Nei 1978). AR (k=2): Allelic richness expected number of alleles among 2 gene copies. He; expected numbers of heterozygotes and HO: observed number of heterozygotes.

Locus	Fis	Fst	Missing genotypes (%)	P-val	NA	NAe:	AR(k=2):	He	Ho	Fi
epi3	0.0863	0.1265	8 (0.8%)	0,0002	8	2.06	1.52	0.5156	0.413	0.200
FRPP94	0.2449	0.1147	6 (0.6%)	0	11	1.73	1.42	0.4221	0.283	0.330
ITPH4516	0.1035	0.2140	22 (2.1%)	0	20	3.67	1.73	0.7273	0.515	0.292
pEST2669	0.1082	0.3212	11 (1.1%)	0,0001	8	1.86	1.46	0.4630	0.282	0.390
B4F08	0.0418	0.1338	14 (1.3%)	0,0371	14	2.65	1.62	0.6220	0.518	0.167
pEST8	0.1613	0.1678	19 (1.8%)	0	9	2.47	1.60	0.5953	0.417	0.299
PtTX3030	0.1199	0.1424	23 (2.2%)	0	9	2.31	1.57	0.5672	0.430	0.243
PtTX3116	0.0716	0.1135	92 (8.8%)	0,0002	6	1.93	1.48	0.4806	0.397	0.175

Table S5: SSRs statistics per population.
Sample size; number of individuals per population, percentage of missing genotypes per population. P-value, for hardy Weinberg equilibrium. NA: Number of alleles. NAe: effective number of alleles. AR(k=2): Allelic richness (expected number of alleles among 2 gene copies) He: Gene diversity corrected for sample size. Fi: inbreeding coefficient. **Average He per pop=0,4654**

population code	Sample size	Missing genotypes (%)	P-val HW	NA:	NAe:	AR(k=2)	He	Fi
VO_IV_11_4	27	0.1 (0.4%)	0.1291	3	1,55	1,32	0,3204	-0,06
VO_IV_21_4	26	1.1 (4.3%)	0.5662	3,33	1,85	1,41	0,4087	-0,007
VO_IV_61_4	25	0.6 (2.2%)	0.3593	3,11	1,82	1,4	0,4025	-0,046
VO_IV_83_4	27	0.1 (0.4%)	0.0345	3,44	1,87	1,43	0,4314	0,032
VO_IV_84_4	27	0.1 (0.4%)	0.0001	3,44	1,78	1,42	0,4181	-0,016
VO_IV_92_4	25	0.6 (2.2%)	0.0598	3,44	1,83	1,43	0,4341	0,068
VO_IV_101_4	26	1.2 (4.7%)	0.0080	3,44	1,81	1,42	0,4201	0,041
VO_IV_105_4	27	0.3 (1.2%)	0.0774	3,33	1,75	1,41	0,4098	0,014
VO_IV_111_4	25	0.1 (0.4%)	0.3073	3	1,69	1,37	0,3705	-0,037
VO_IV_131_4	25	0.3 (1.3%)	0.0421	3,22	1,66	1,36	0,355	-0,037
VO_IV_142_4	27	0.3 (1.2%)	0.0003	3,56	1,91	1,45	0,4532	0,121
VO_IV_152_4	26	0.1 (0.4%)	0.0061	3,33	1,97	1,47	0,4725	0,048
VO_IV_154_4	27	0.3 (1.2%)	0.0000	3,44	1,99	1,47	0,4699	0,099
VO_IV_157_4	26	0.3 (1.3%)	0.0005	3,67	1,95	1,47	0,4737	0,123
VO_IV_173_4	23	0.2 (1.0%)	0.0114	2,78	1,94	1,46	0,4572	0,139
VO_IV_182_4	25	0.0 (0.0%)	0.0356	2,44	1,63	1,34	0,3395	0,005
VO_IV_183_4	26	0.1 (0.4%)	0.0915	2,67	1,65	1,35	0,3542	-0,089
CU_III_184_2	21	0.1 (0.5%)	0.0009	2,67	1,84	1,4	0,3987	-0,029
VO_IV_186_4	25	0.2 (0.9%)	0.2891	2,67	1,65	1,37	0,3663	-0,119
VO_IV_187_4	26	0.2 (0.9%)	0.2904	2,89	1,73	1,37	0,369	-0,14
VO_IV_172_4	27	0.4 (1.6%)	0.1046	2,78	1,71	1,39	0,3874	0,013
VO_IV_231_4	22	0.2 (1.0%)	0.0003	4,22	2,88	1,59	0,5945	0,154
VO_IV_232_4	24	0.3 (1.4%)	0.0000	4,22	2,9	1,58	0,578	0,229
VO_IV_233_1	20	0.7 (3.3%)	0.0000	4,56	2,94	1,57	0,5738	0,267
Imp_9	27	0.4 (1.6%)	0.6566	3	1,65	1,34	0,342	-0,221
Quer_9	28	0.2 (0.8%)	0.0000	4,11	2,94	1,64	0,6415	0,414
Otric_9	27	0.2 (0.8%)	0.0355	3,11	1,83	1,43	0,4254	-0,02
Carlofort_24	24	4.3 (18.1%)	0.0020	2,56	1,78	1,42	0,4225	0,026
VO_IV_212_4	28	0.1 (0.4%)	0.0438	5,33	2,88	1,62	0,6152	-0,01
VO_IV_214_3	24	0.1 (0.5%)	0.0001	4,67	3	1,61	0,6104	0,062
Gr.Ele_2	2	0.0 (0.0%)	0.0843	2,11	2,13	1,52	0,5185	0,333
Gr.Eubo_7	7	0.6 (7.9%)	0.0059	3,33	2,88	1,57	0,5746	0,188
VO_III_241_4	20	1.4 (7.2%)	0.0209	3,78	2,73	1,53	0,5337	0,048
VO_III_242_4	17	1.6 (9.2%)	0.3358	3,44	2,7	1,5	0,5038	-0,066
Ifrane_9	24	0.1 (0.5%)	0.2292	3,56	2,16	1,45	0,4501	0,053
ALG_Aur_2	2	0.1 (5.6%)	0.4395	2,44	2,81	1,67	0,6667	0

Isr.Nat_9	24	0.4 (1.9%)	0.0000	4,11	2,55	1,57	0,5678	0,211
Isr.A6_9	24	3.1 (13.0%)	0.6182	3,56	2,06	1,48	0,4812	-0,121
TD_417	30	0.0 (0.0%)	0.0010	3,56	1,93	1,44	0,4419	0,113
Bou_30_0_2	8	1.0 (12.5%)	0.0000	3,67	3,16	1,52	0,5201	0,396
StM_9	24	0.3 (1.4%)	0.3042	3,56	2,06	1,44	0,4435	-0,031
H1_H03	24	0.2 (0.9%)	0.0002	6,11	3,33	1,63	0,6334	0,15
H2_H06	24	0.2 (0.9%)	0.0044	3,56	2,64	1,54	0,536	0,007
192_IV_4	21	0.7 (3.2%)	0.1133	2,56	1,51	1,29	0,2919	-0,206

Table S6: SNP statistics per population
Sample size; number of individuals per population, percentage of missing genotypes per population. P-value, for Hardy Weinberg equilibrium. NA: Number of alleles. NAe: effective number of alleles. AR (k=2): Allelic richness (expected number of alleles among 2 gene copies) He: Gene diversity corrected for sample size. Fi: inbreeding coefficient. **Average He per pop= 0,2341**

population code	Sample size	Missing genotypes (%)	P-val HW	NA:	NAe:	AR(k=2)	He	Fi
VO_IV_011_4	27	0.8%	0.4428	1,71	1,37	1,22	0,219	0,009
VO_IV_021_4	26	1.3%	0.0236	1,75	1,39	1,23	0,2304	0,015
VO_IV_061_4	25	0.7%	0.7585	1,74	1,38	1,23	0,2273	-0,016
VO_IV_083_4	27	0.7%	0.0006	1,73	1,4	1,24	0,2387	0,05
VO_IV_084_4	27	1.0%	0.7601	1,75	1,39	1,23	0,2291	-0,021
VO_IV_092_4	25	1.0%	0.0000	1,75	1,42	1,24	0,2441	0,069
VO_IV_101_4	26	1.2%	0.0623	1,75	1,43	1,25	0,2454	0,017
VO_IV_105_4	27	2.0%	0.9753	1,76	1,43	1,25	0,2483	-0,041
VO_IV_111_4	24	1.0%	0.0132	1,73	1,36	1,21	0,2118	0,049
VO_IV_131_4	23	0.9%	0.0000	1,75	1,41	1,24	0,2419	0,058
VO_IV_142_4	27	0.8%	0.0000	1,78	1,43	1,25	0,2491	0,096
VO_IV_152_4	26	1.0%	0.0001	1,74	1,44	1,25	0,2515	0,047
VO_IV_154_4	26	1.1%	0.0000	1,76	1,42	1,24	0,2443	0,13
VO_IV_157_3	24	1.0%	0.0000	1,73	1,43	1,25	0,2488	0,08
VO_IV_172_4	27	2.0%	0.0000	1,67	1,36	1,21	0,2118	0,046
VO_IV_173_3	21	1.0%	0.0052	1,7	1,38	1,23	0,2268	0,032
VO_III_182_2	20	0.5%	0.0000	1,64	1,33	1,19	0,1948	0,277
VO_IV_183_4	25	2.0%	0.0049	1,68	1,38	1,22	0,221	0,031
VO_III_184_4	15	1.5%	0.0000	1,66	1,38	1,22	0,2215	0,165
VO_IV_186_4	22	0.9%	0.2335	1,67	1,36	1,21	0,2103	0,013
VO_IV_187_4	23	1.0%	0.0000	1,68	1,37	1,22	0,2162	0,166

VO_IV_212_4	26	2.7%	0.0043	1,68	1,32	1,19	0,1902	0,049
VO_IV_214_3	24	2.1%	0.0000	1,76	1,44	1,26	0,2551	0,132
VO_IV_231_4	21	1.5%	0.0000	1,8	1,45	1,26	0,2596	0,082
VO_IV_232_4	25	1.3%	0.0000	1,79	1,44	1,26	0,2563	0,107
VO_IV_233_4	25	1.8%	0.0000	1,78	1,45	1,26	0,256	0,166
VO_IV_241_4	26	1.1%	0.0000	1,79	1,45	1,26	0,2634	0,072
VO_IV_242_4	26	0.8%	0.0000	1,78	1,43	1,25	0,2536	0,066
4R.ELE-2	2	2.0%	0.2129					
4R.EUBO-7	7	2.5%	0.0000					
ISRA6_24	22	3.2%	10.000	1,7	1,38	1,22	0,2192	-0,103
ISRNAT_22	20	2.5%	0.1543	1,72	1,37	1,22	0,2167	0,026
IT.Carlofort_24	21	5.7%	0.0000	1,65	1,38	1,22	0,2174	0,099
IT.IMP_31	16	1.3%	0.0000	1,62	1,29	1,18	0,1767	0,168
IT.OTRIC_48	33	1.7%	0.0000	1,82	1,44	1,26	0,2571	0,142
IT.QUARC_27	23	8.7%	0.0000	1,71	1,43	1,25	0,2487	0,487
AL4.AUR-2	2	0.7%	0.5137	1,47	1,42	1,26	0,2647	0,007
Ifrane_24	24	4.5%	10.000	1,65	1,36	1,21	0,2102	-0,066
Ph1V317	63	0.8%	0.0000	1,84	1,44	1,26	0,2592	0,072
Ph2V318	66	2.0%	0.0093	1,81	1,42	1,25	0,2497	0,02
Ph5A67	67	2.8%	0.0000					
1292	31	0.5%	0.0000	1,73	1,35	1,21	0,211	0,087
1040	33	0.3%	0.0000	1,76	1,41	1,24	0,242	0,057
1140	39	0.6%	0.0128	1,78	1,42	1,24	0,2427	0,035
1630	31	0.6%	0.0000	1,74	1,41	1,24	0,2425	0,064
1602	39	0.4%	0.0342	1,77	1,43	1,25	0,2508	0,026
34A	34	1.0%	0.0039	1,73	1,37	1,22	0,2175	0,043
33B	33	0.2%	0.0008	1,75	1,38	1,23	0,2283	0,038
33C	33	0.5%	0.4774	1,76	1,42	1,25	0,2487	-0,001

A.1.3 Population structure

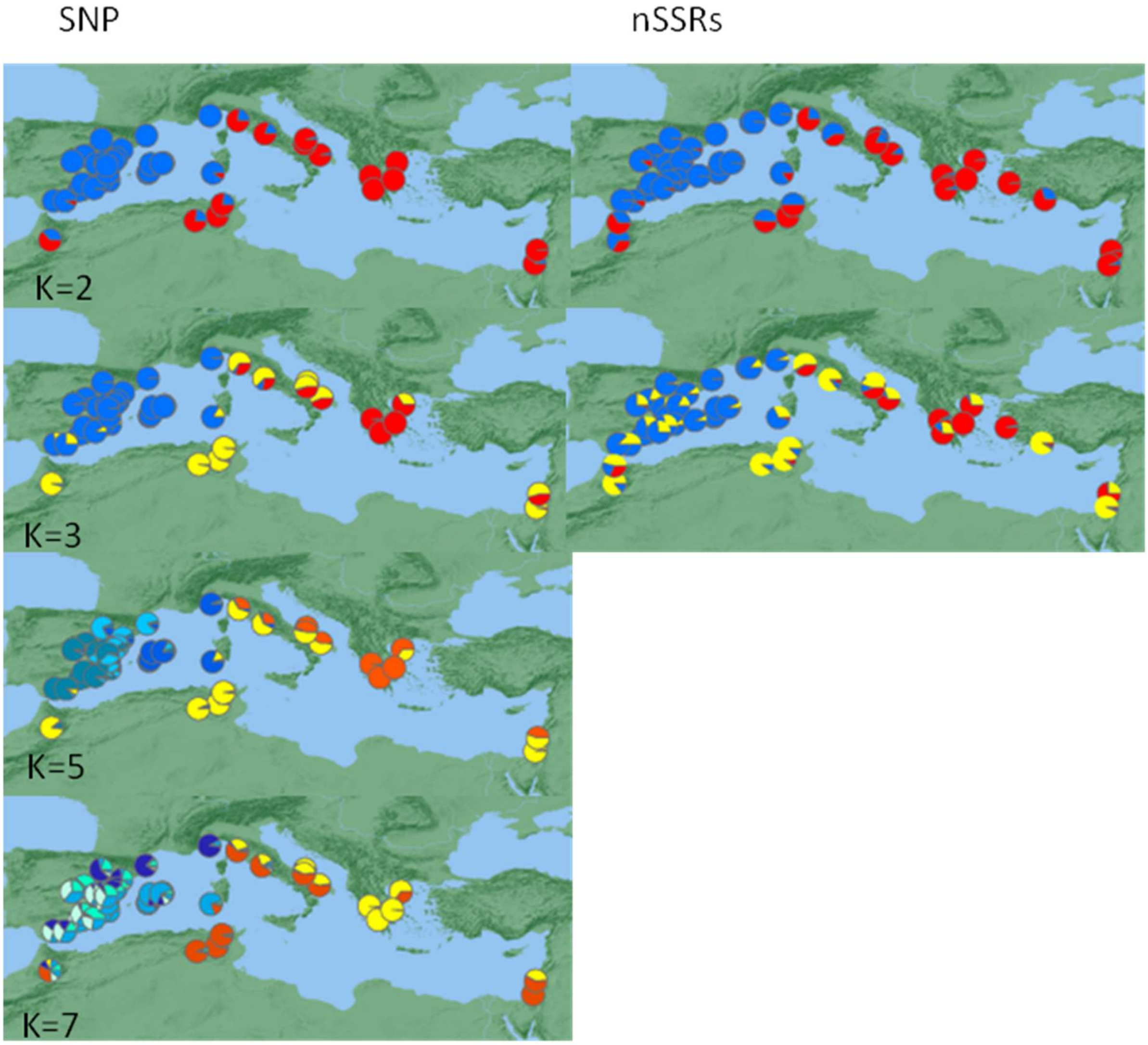


Figure S3: These figures represent Bayesian clustering of populations across most of the natural range of *pinus halepensis*. It's possible to see that the Bayesian clustering and therefore demography coincides in both sets of markers up to k=3 after with the nSSRs shows very noisy clustering. However with the SNP a further clustering emerges in the western part of the range that can be visualised up to k=7.

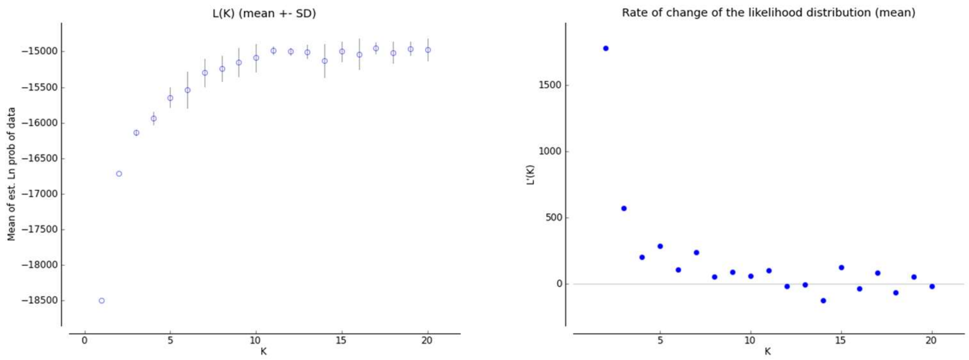


Figure S4 :both the delta k and the rate of change o the likelihood (Evanno *et al.* 2005) for the nSSR data set, that indicate a possible k=2. Data taken from structure and processed in structure harvester.

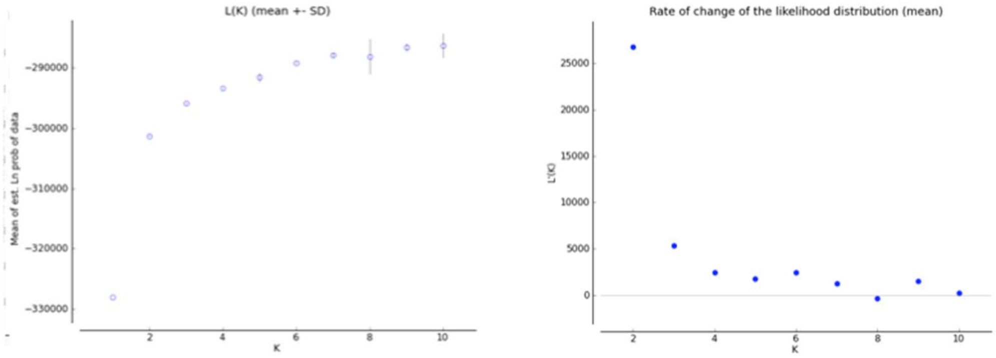


Fig S5 :both the delta k and the rate of change o the likelihood (Evanno *et al.* 2005) for the SNP data set, that indicate a possible k=2. Data taken from structure and processed in structure harvester.

A.1.4 Selection

)

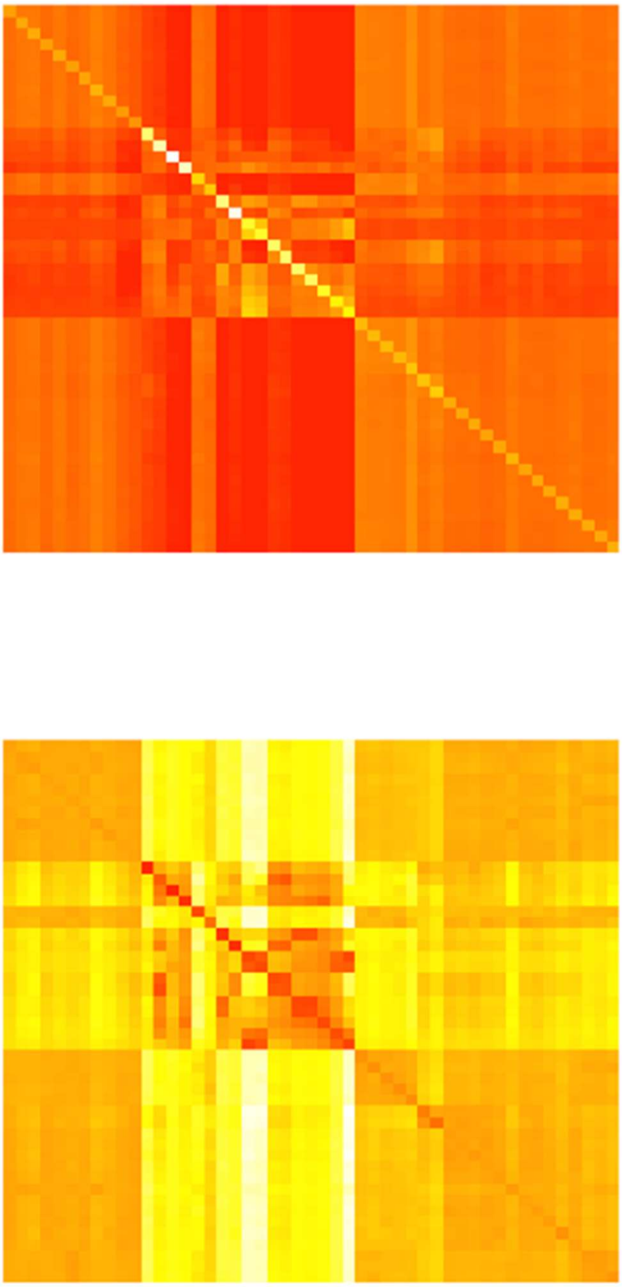


Figure S6: heatmap of pairwise Fst (left covariance matrix (left) produced using Bayenv2, these visually show that there are similar in representing the genetic variation amongst the populations in the SNP data.

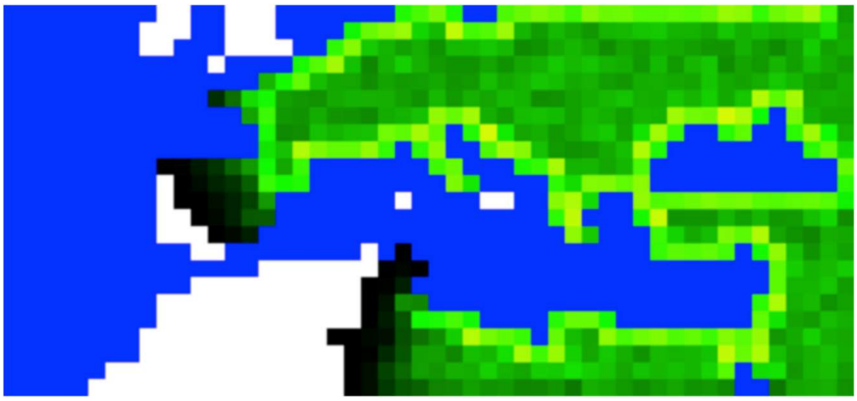


Figure S7: A map showing the simulated data set produced in Splatche2 colonising the Mediterranean basin.

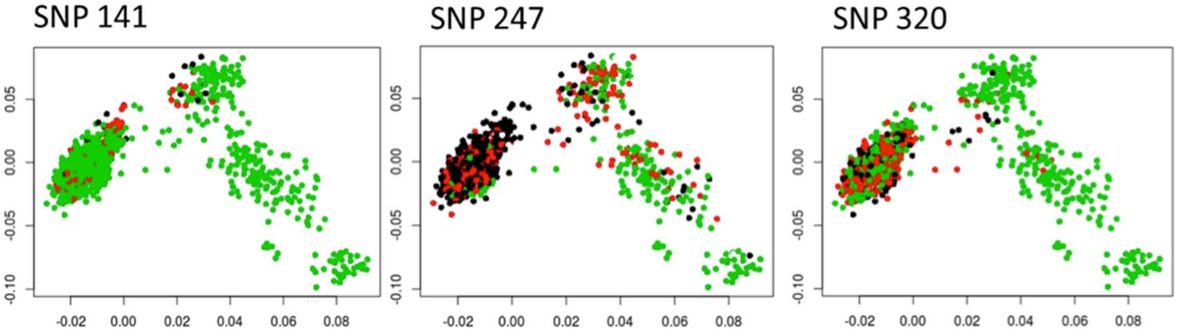


Figure S8: Showing the distribution of SNP 141, 247 and 320, here you see the distribution of allele1 in green, allele 2 in red and heterozygote in black, over a backdrop of a PCA of all alleles in the SNP dataset. There alleles are candidate for genesufing with Bayes factors between 3 and 4.

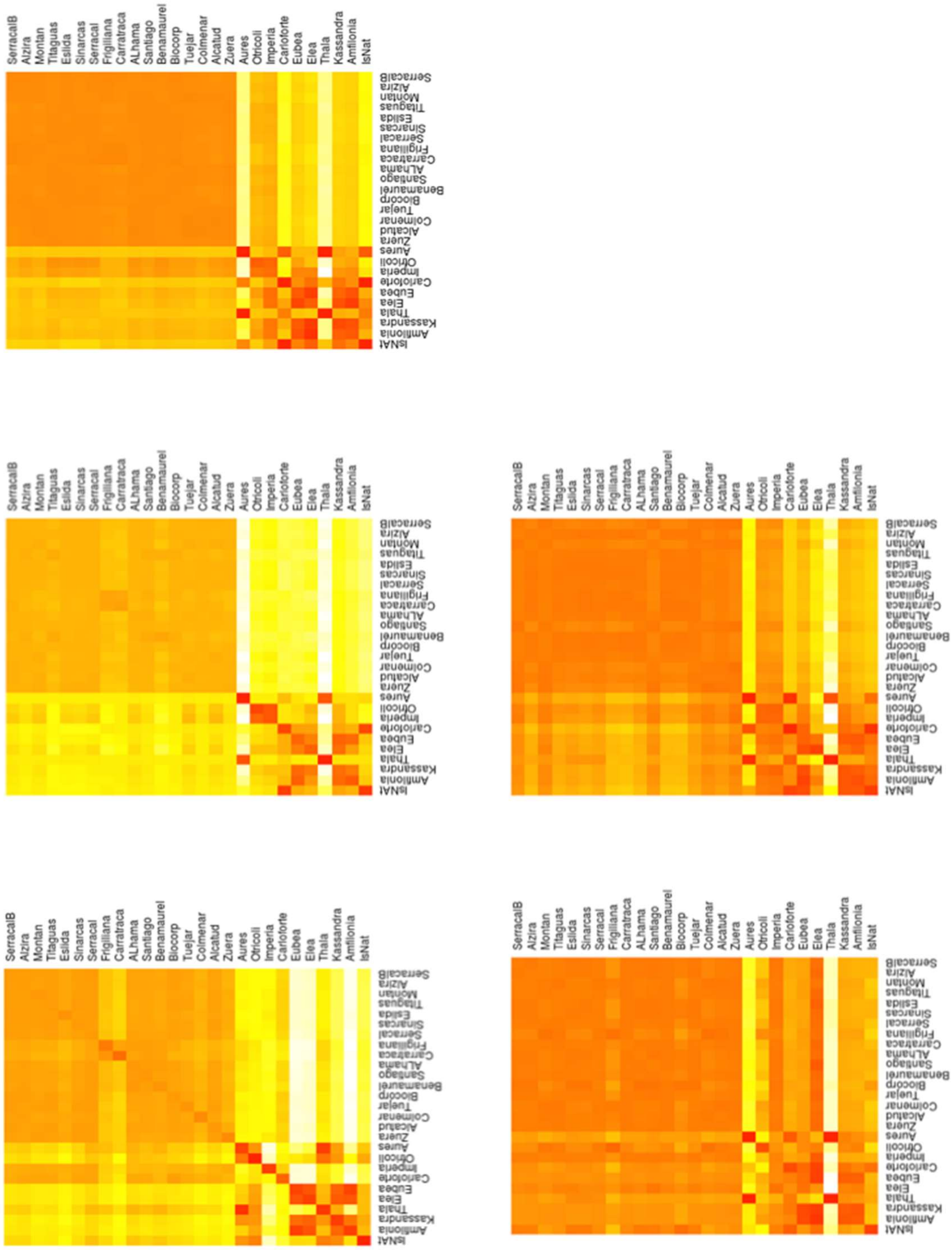


Figure S9: heatmaps showing the pairwise fst of the real(top right) data , compared to pairwise fst of its analogous simulated data sets from right to left with migration rates of 0.1, 0.18, 0.26 and 0.34

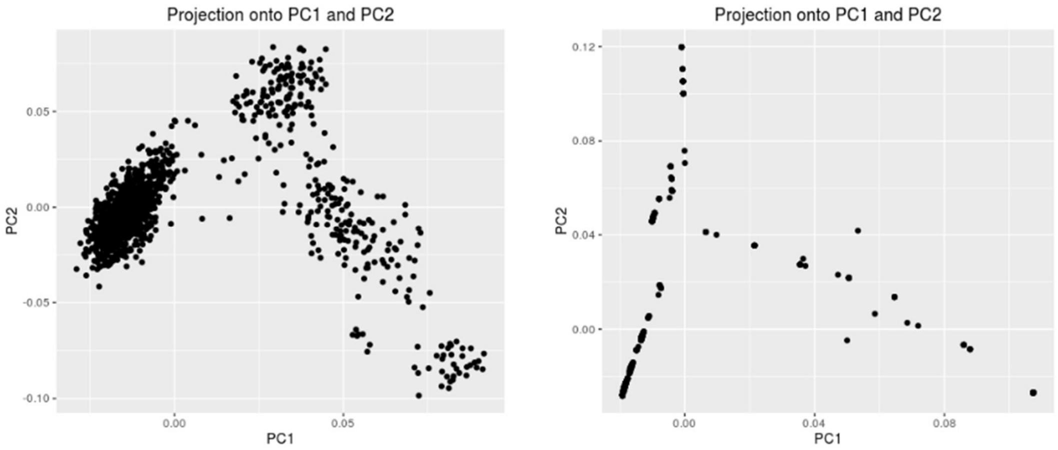


Figure S10: The pcas from the real (a) and simulates data mig 0.18(b) these help select the number of pcas for the analysis.

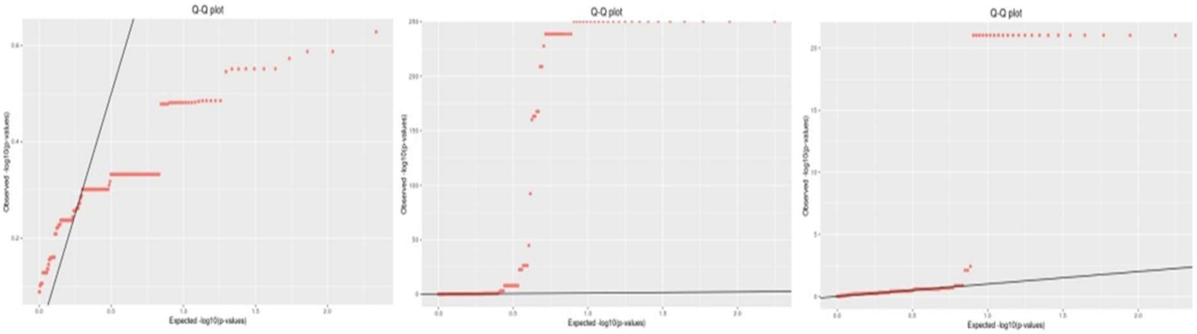


Figure S11: Quantile-quantile plots showing outliers from the *PCAdapt* analysis for simulated data for migration rate of 0.18(a) 0.26 (b) and 0.34 (C).

A.1.5 References

Eckert AJ, Van Heerwaarden J, Wegrzyn JL *et al.* (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (Pinus taeda L., Pinaceae). *Genetics*, **185**, 969–982.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Hardy OJ, Vekemans X (2002) spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Kumar L, Skidmore AK, Knowles E (2007) Modelling topographic variation in solar radiation in a GIS environment. *International Journal of Geographical Information Science*, **11**, 475–497.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583–90.

Nielsen R, Tarpy DR, Reeve HK (2003) Estimating effective paternity number in social insects and the effective number of alleles in a population. *Molecular Ecology*, **12**, 3157–3164.

Raymond M, Rousset F (1995) GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism. *J. Hered.*, **86**, 248–249.

Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Turc L (1963) Evaluation des besoins en eau d'irrigation, évapotranspiration potentielle, formulation simplifié et mise à jour. *Annales agronomiques*, **12**, 13–49.

Zimmermann NE, Edwards TC, Moisen GG, Frescino TS, Blackard JA (2007) Remote sensing-based predictors improve distribution models of rare, early successional and broadleaf tree species in Utah. *The Journal of applied ecology*, **44**, 1057–1067.

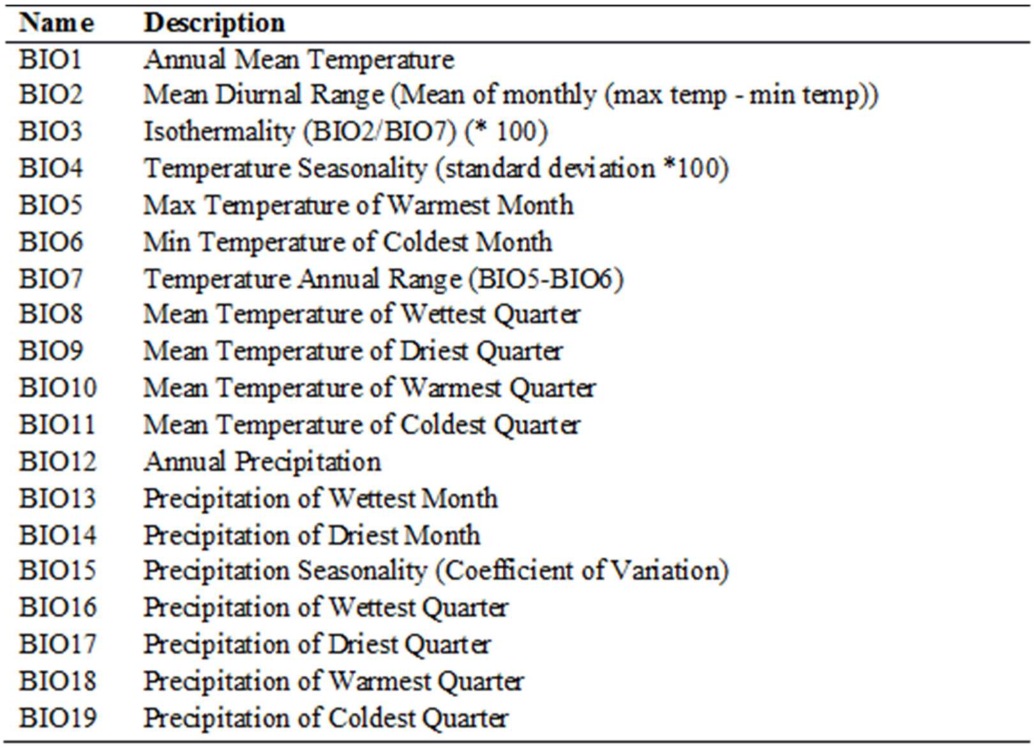
A.2 Chapter 3 (Part I)

A.2.1 Supplementary Tables

Table S1: details of the seven sites sampled.

Country	Altitude	Plot	Coordinates		Sampling size
			latitude	longitude	
France	Low	Saint Mitre les Remparts	43.4518140	5.0419330	50
France	Intermediate	Font Blanche	43.2407610	5.6791030	256
France	Wet	Sioux-Blanc	43.2375560	5.8875000	50
Italy	Low	Mattinata	41.6949892	16.0589908	25
Italy	High	Monte San Angelo	41.6949104	16.0216220	25
Spain	Low	Alzira	39.1223287	-0.3892886	39
Spain	High	Montan	40.0472534	-0.5925626	31

Table S2: list of bioclimatic variables BIO1-BIO19 available in WORDLCLIM.



Name	Description
BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) (* 100)
BIO4	Temperature Seasonality (standard deviation *100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

Table S3: environmental variables retrieved from WORDLCLIM

Population	**alt**	**bio_1**	**bio_2**	**bio_3**	**bio_4**	**bio_5**	**bio_6**	**bio_7**	**bio_8**	**bio_9**	**bio_10**	**bio_11**	**bio_12**	**bio_13**	**bio_14**	**bio_15**	**bio_16**	**bio_17**	**bio_18**	**bio_19**
FontBlanche	408	126	83	34	5505	257	17	240	97	198	198	58	711	90	17	33	253	94	94	206
Sioux Blanc	614	116	80	34	5419	245	10	235	89	187	187	50	777	94	20	32	273	107	107	226
St Mitre les Remparts	138	134	96	36	5885	279	13	266	142	210	210	59	614	85	18	33	220	89	89	164
Monte San Angelo	462	135	69	29	5958	265	33	232	108	214	214	65	535	63	29	25	180	93	93	156
Mattinata	79	160	76	32	5783	289	55	234	135	235	235	90	465	60	20	35	173	69	69	143
Montan	848	123	94	36	5878	265	10	255	132	54	202	54	490	59	23	29	158	93	106	93
Alzira	112	173	97	40	5197	301	60	241	187	239	243	109	466	81	9	48	191	50	80	120

Table S3: environmental variables retrieved from WORDLCLIM

Table S4: pairwise F_{ST} between the seven populations from France (FR), Italy (IT) and Spain (SP).

		FR	**FR**	**FR**	**IT**	**IT**	**SP**	**SP**
		Font Blanche	**Sioux-Blanc**	**Saint Mirtre**	**Monte San Angelo**	**Mattinata**	**Montan**	**Alzira**
FR	**Font Blanche**	0						
FR	**Sioux-Blanc**	0.0121	0					
FR	**Saint Mirtre**	0.0049	0.0138	0				
IT	**Monte San Angelo**	0.3082	0.3124	0.3254	0			
IT	**Mattinata**	0.3073	0.3115	0.3232	0.0122	0		
SP	**Montan**	0.0853	0.1056	0.0861	0.2978	0.2877	0	
SP	**Alzira**	0.0999	0.1210	0.0990	0.2760	0.2663	0.0151	0

Table S5: summary of the results of the Bayesian linear model performed in Bayenv2 with different bioclimatic variables, as well as distance from East to West and altitude.

SNP	Contig	altitude	BIO9	
169	seq-0_10162_01-244	20.903	41.971	
		BIO2		
312	seq-UMN_3408_01-293	20.853		
		BIO2	BIO19	
316	seq-10373-2483	20.505	53.894	
		BIO12	BIO16	BIO19
378	seq-2_3941_01-381	47.378	64.627	71.261

Table S6: summary of the results of the Bayesian linear model performed in Baypass with different bioclimatic variables.

SNP	Contig	Env	M-pearson	SD_Pearson	eBPis
2	seq-6890-2409	BIO3	-0,5733702	0,23280148	3,11399102
4	seq-9882-801	BIO12	-0,5251087	0,1238397	5,37596943
7	seq-7270-1484	BIO2	-0,8051462	0,11801382	4,14633727
7	seq-7270-1484	BIO7	-0,7214707	0,1499661	3,06285797
7	seq-7270-1484	BIO12	0,76639455	0,08623712	4,46058182
7	seq-7270-1484	BIO16	0,72976022	0,13401115	3,03150769
7	seq-7270-1484	BIO19	0,7887976	0,11145945	3,75236008
30	seq-9243-371	BIO12	0,64362547	0,14199203	5,24558596
30	seq-9243-371	BIO16	0,59596188	0,1386626	3,71530188
30	seq-9243-371	BIO19	0,58877171	0,14257403	3,97485605
148	seq-0_12216_02-537	BIO12	-0,7620282	0,15270792	4,15829304
148	seq-0_12216_02-537	BIO19	-0,7102334	0,16325765	3,07555576
151	seq-0_8992_01-119	BIO12	0,65511587	0,19022025	3,07242975
169	seq-0_10162_01-244	Altitude	0,76153998	0,13068642	3,76859495
169	seq-0_10162_01-244	BIO1	-0,6542224	0,14417419	3,09435782
169	seq-0_10162_01-244	BIO9	-0,8379422	0,08956155	5,48337617
169	seq-0_10162_01-244	BIO11	-0,6239025	0,14530298	3,00001402
169	seq-0_10162_01-244	BIO13	-0,7137226	0,12101459	5,19887839
169	seq-0_10162_01-244	BIO15	-0,6434817	0,15061643	3,08844092
182	seq-0_16860_01-314	Dry/Wet	0,8022037	0,14918111	3,00958542

205	seq-CL708CONTIG1_02-173	BIO12	-0,762647	0,17323487	3,24830529
258	seq-9882-2209	Altitude	0,47978421	0,14401977	5,0545209
258	seq-9882-2209	BIO12	-0,5801209	0,11965805	6,33377503
258	seq-9882-2209	BIO19	-0,4719654	0,14022928	3,01268994
269	seq-16094-1379	BIO12	0,6201313	0,13851844	4,85579828
269	seq-16094-1379	BIO16	0,63737081	0,13333415	3,77450901
269	seq-16094-1379	BIO19	0,67220926	0,12747789	4,72441191
281	seq-16094-410	Altitude	-0,447868	0,16543734	3,2116434
281	seq-16094-410	BIO12	0,58038238	0,135169	5,20973055
281	seq-16094-410	BIO12	0,57368845	0,13367159	3,02821708
281	seq-16094-410	BIO16	0,64456484	0,12215464	4,64057127
281	seq-16094-410	BIO19	0,63426007	0,12530935	4,91602902
316	seq-10373-2483	Altitude	-0,5209439	0,13402176	4,36836379
316	seq-10373-2483	BIO12	-0,525043	0,13970482	3,71427816
316	seq-10373-2483	BIO12	0,67126778	0,10078796	6,71104438
316	seq-10373-2483	BIO16	0,58991637	0,1342952	3,57113029
316	seq-10373-2483	BIO19	0,67505682	0,1205309	5,17564163
325	seq-8188-285	Altitude	-0,5498621	0,2263624	3,03248511
325	seq-8188-285	BIO12	0,75402557	0,16680693	3,51109846
337	seq-36858-735	BIO10	0,56181138	0,15827186	3,08795106
337	seq-36858-735	BIO12	-0,6011001	0,17039471	3,3155499
364	seq-2_2937_01-309	BIO2	-0,8005791	0,12499065	4,36356318

364	seq-2_2937_01-309	BIO7	-0,6979643	0,15040826	3,26117794
364	seq-2_2937_01-309	BIO12	0,76465655	0,0946094	4,85192671
364	seq-2_2937_01-309	BIO16	0,70987328	0,14077677	3,14422469
364	seq-2_2937_01-309	BIO19	0,76967958	0,11920607	4,04043984
378	seq-2_3941_01-381	BIO12	0,51160742	0,15811359	3,65435041

A.2.2 Supplementary Figures

In the following figures the number corresponds to the following populations; 1: Font Blanche (France), 2: Sioux-Blanc (France), 3: Saint Mitre (France), 4: Monte San Angelo (Italy), 5: Mattianata (Italy), 6: Montan (Spain) and 7: Alzira (Spain)

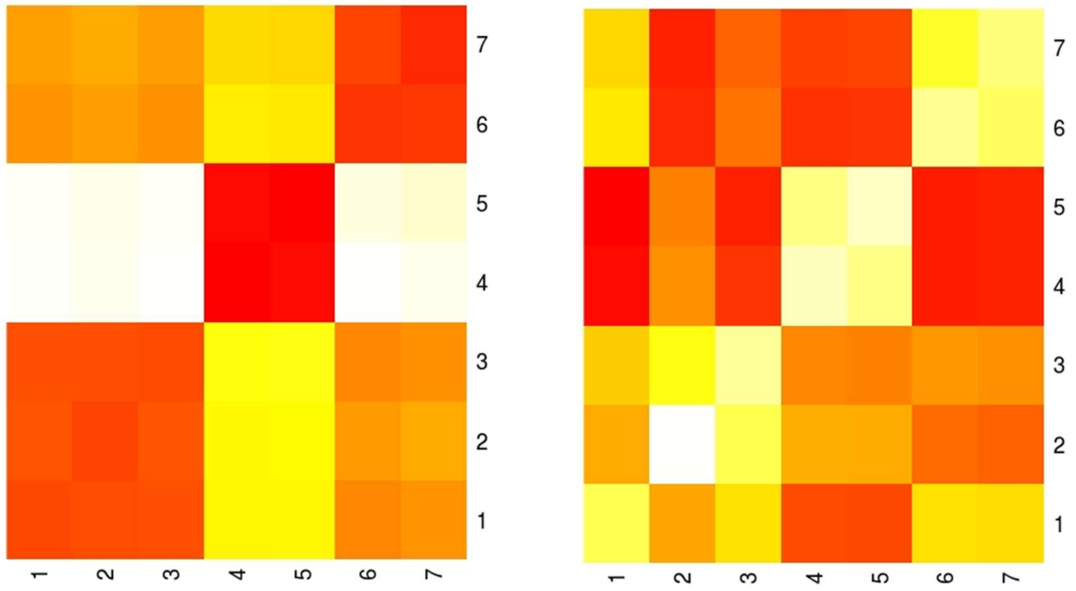


Figure S1: Heatmaps of the pairwise F_{ST} distance (left) and the covariance matrix calculated in bayenv2 (right).

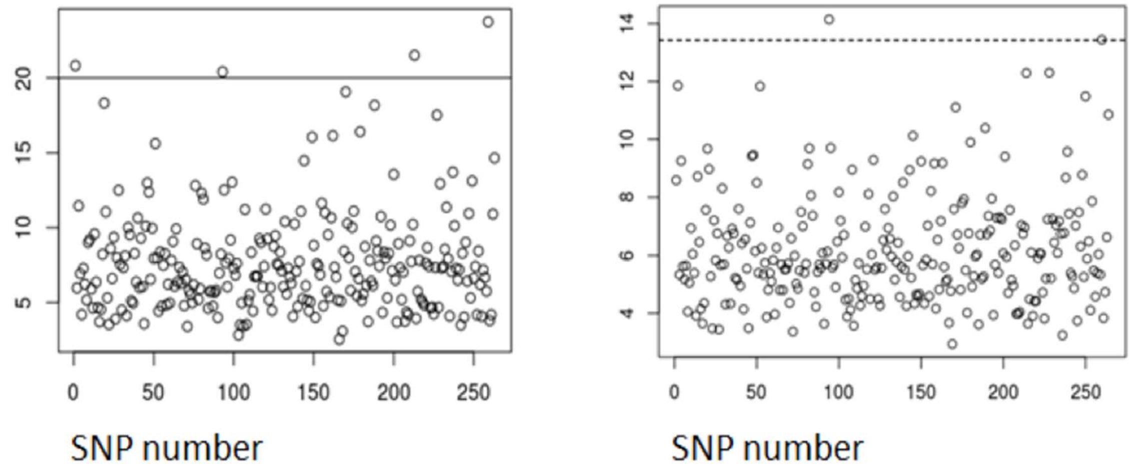


Figure S2: The *XtX*s estimated from Bayenv2 (left) with a threshold estimated based on a cut off from the distribution (horizontal dark line) and from Baypass (right) showing the 0.01 per cent cut off threshold (horizontal dotted line) computed using the POD data.

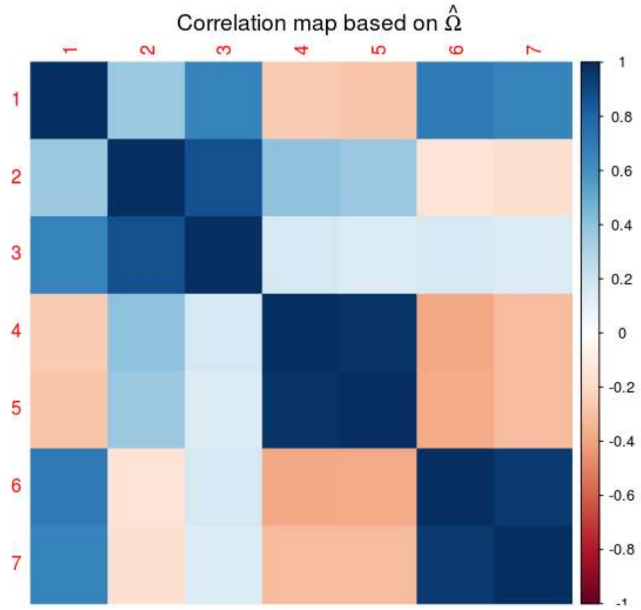


Figure S3: A correlation matrix comparing the $\widehat{\Omega}$ values amongst the populations of the SNP data set computed in Baypass.

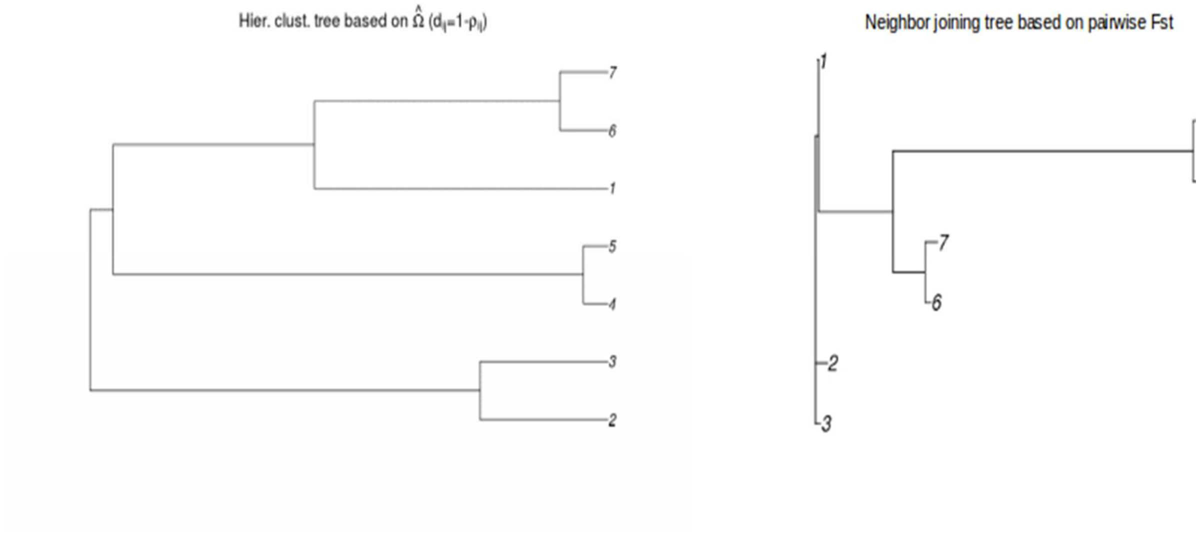

Hier. clust. tree based on $\hat{\Omega}$ (d_{ij}=1-ρ_{ij})
7
6
1
5
4
3
2
Neighbor joining tree based on pairwise Fst
1
7
6
2
3

Figure S4: The correlation matrix visualised as a hierarchical cluster tree in order to where the relationships between populations can be appreciated (left) compared to a neighbour joining (right) tree of pairwise *F_{ST}*.

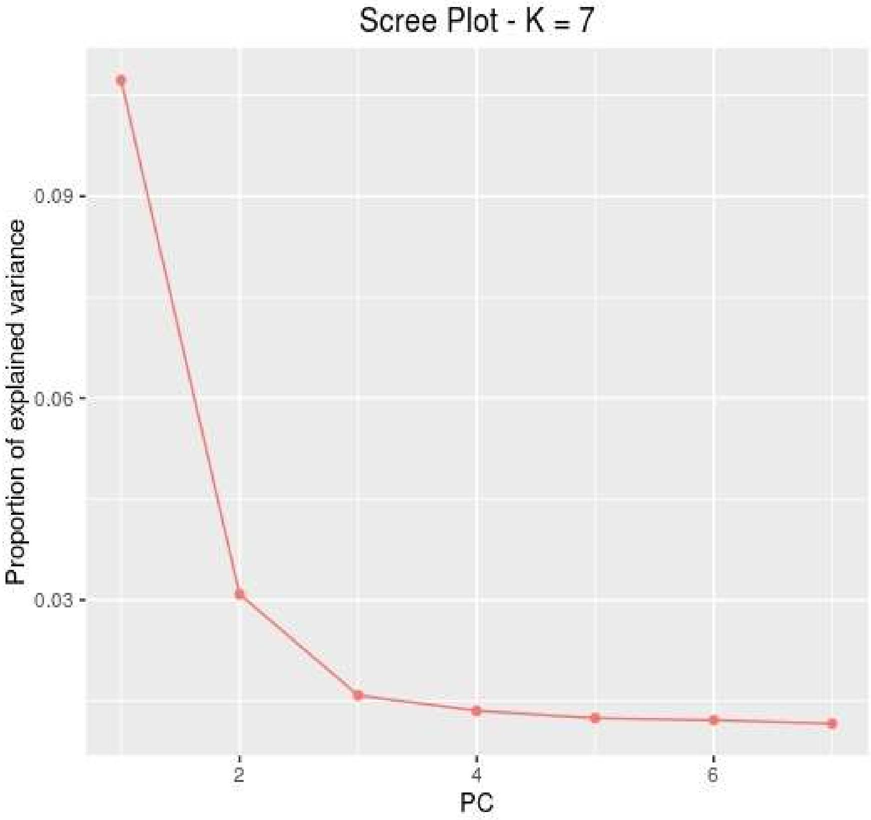


Figure S5: Scree plot that displays in decreasing order the percentage of variance explained by each PC. These correspond to the eigenvalues in decreasing order.

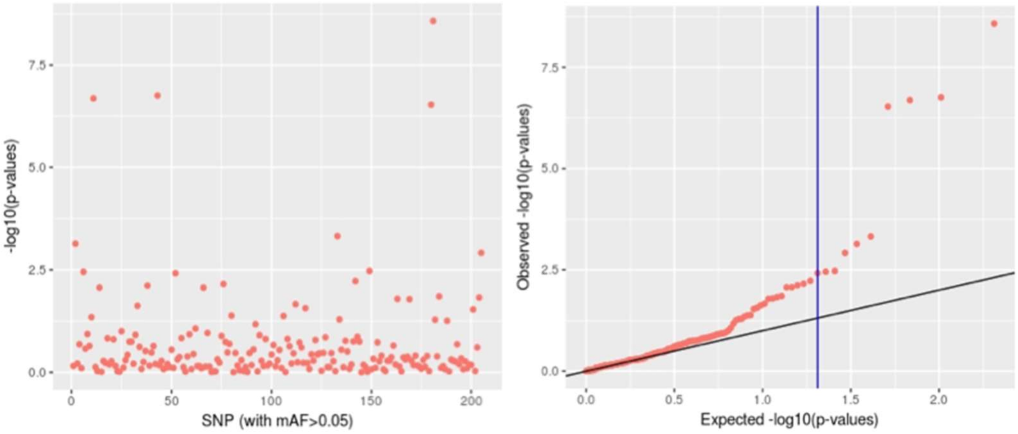


Figure S6: Distribution of the empirical p-values obtained in PCAdapt visualized through a Manhattan plot (left) and a QQ-plot (right) showing the cut off of 0.1% (vertical blue line).

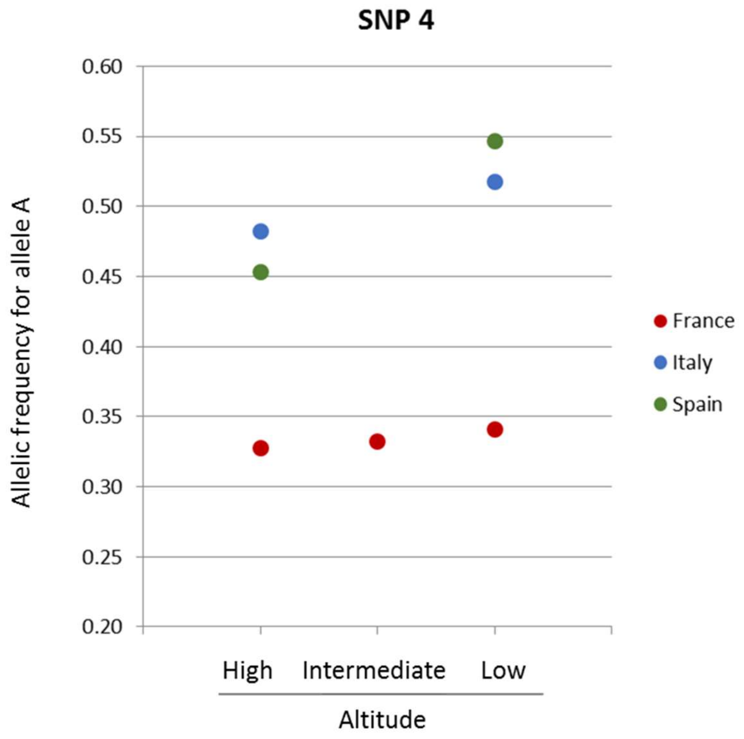


Figure S7: Distribution of the allelic frequency of allele A for SNP 4 detected under selection with PCAdapt, Bayenv2 and Baypass.

A.3 Chapter 4

Table S1: The validation of the miRBase v.21 A. thaliana miRNAs

miRNA	% reads mapping to star or mature	Presence of miRNA* sequence	Complementarity of mature	Family	Automated	miRBase	Taylor *et al.*2014
ath-MIR156a_MI0000178	99,99769214	Yes	20	miR156	1	1	1
ath-MIR156b_MI0000179	99,99496169	Yes	20	miR156	1	1	1
ath-MIR156c_MI0000180	99,9974817	Yes	20	miR156	1	1	1
ath-MIR156d_MI0000181	99,9985222	Yes	20	miR156	1	1	1
ath-MIR156e_MI0000182	99,99980973	Yes	20	miR156	1	1	1
ath-MIR156f_MI0000183	99,9994585	Yes	19	miR156	1	1	1
ath-MIR156g_MI0001082	99,9941232	Yes	19	miR156	1	0	1
ath-MIR156h_MI0001083	99,99749499	Yes	19	miR156	1	1	1
ath-MIR156i_MI0019232	85,08583691	No	14	miR156	1	0	1
ath-MIR156j_MI0019234	99,8327934	No	16	miR156	1	0	1
ath-MIR157a_MI0000184	99,98727787	Yes	19	miR157	1	1	1
ath-MIR157b_MI0000185	99,99025119	Yes	19	miR157	1	1	1
ath-MIR157c_MI0000186	99,7075539	No	13	miR156	0	1	1
ath-MIR157d_MI0000187	99,9815959	Yes	20	miR156	1	1	1
ath-MIR158a_MI0000188	99,83923445	Yes	17	miR158	1	1	1
ath-MIR158b_MI0001084	99,8507874	Yes	17	miR158	1	1	1
ath-MIR159a_MI0000189	99,95746902	Yes	19	miR159	1	0	1
ath-MIR159b_MI0000218	99,8984352	Yes	18	miR159	1	1	1
ath-MIR159c_MI0001085	99,99019556	No	17	miR159	1	0	1
ath-MIR160a_MI0000190	99,71605247	Yes	19	miR160	1	1	1
ath-MIR160b_MI0000191	99,97894445	Yes	19	miR160	1	1	1
ath-MIR160c_MI0000192	99,25265837	Yes	19	miR160	1	1	1
ath-MIR161_MI0000193	87,37174984	Yes	19	miR161	1	0	1
ath-MIR162a_MI0000194	99,88245099	Yes	18	miR162	1	1	1
ath-MIR162b_MI0000195	99,78765967	Yes	18	miR162	1	1	1
ath-MIR163_MI0000196	96,00293958	Yes	21	miR163	1	1	1
ath-MIR164a_MI0000197	99,28241706	Yes	18	miR164	1	1	1
ath-MIR164b_MI0000198	99,87888799	Yes	16	miR164	1	1	1

ath-MIR164c_MI0001087	99,95795786	Yes	19	miR164	1	1	1
ath-MIR165a_MI0000199	99,98619363	Yes	17	miR166	1	1	1
ath-MIR165b_MI0000200	99,99857221	Yes	17	miR166	1	1	1
ath-MIR166a_MI0000201	99,99357807	Yes	17	miR166	1	1	1
ath-MIR166b_MI0000202	99,99402585	Yes	17	miR166	1	1	1
ath-MIR166c_MI0000203	99,99975415	Yes	16	miR166	1	0	1
ath-MIR166d_MI0000204	99,99972735	Yes	16	miR166	1	0	1
ath-MIR166e_MI0000205	99,99676908	Yes	17	miR166	1	1	1
ath-MIR166f_MI0000206	99,999751	Yes	18	miR166	1	1	1
ath-MIR166g_MI0000207	99,99972745	Yes	17	miR166	1	1	1
ath-MIR167a_MI0000208	99,99535784	Yes	20	miR167	1	1	1
ath-MIR167b_MI0000209	99,99733109	Yes	19	miR167	1	0	1
ath-MIR167c_MI0001088	94,9166902	Yes	20	miR167	1	1	1
ath-MIR167d_MI0000975	99,98653093	Yes	20	miR167	1	0	1
ath-MIR168a_MI0000210	99,89026446	Yes	18	miR168	1	1	1
ath-MIR168b_MI0000211	99,97832055	Yes	18	miR168	1	1	1
ath-MIR169a_MI0000212	99,54020399	Yes	20	miR169	1	1	1
ath-MIR169b_MI0000976	97,71837602	Yes	20	miR169	1	0	1
ath-MIR169c_MI0000977	99,97822431	No	19	miR169	1	0	1
ath-MIR169d_MI0000978	99,86415699	Yes	18	miR169	1	1	1
ath-MIR169e_MI0000979	99,94715029	Yes	18	miR169	1	0	1
ath-MIR169f_MI0000980	99,79536729	Yes	18	miR169	1	1	1
ath-MIR169g_MI0000981	99,86550167	Yes	18	miR169	1	0	1
ath-MIR169h_MI0000982	99,70928112	Yes	19	miR169	1	1	1
ath-MIR169i_MI0000983	99,23480521	Yes	19	miR169	1	1	1
ath-MIR169j_MI0000984	99,74021173	Yes	19	miR169	1	1	1
ath-MIR169k_MI0000985	99,83444568	Yes	19	miR169	1	1	1
ath-MIR169l_MI0000986	99,77943693	Yes	19	miR169	1	0	1
ath-MIR169m_MI0000987	99,6514562	Yes	19	miR169	1	1	1
ath-MIR169n_MI0000988	99,74189523	Yes	19	miR169	1	1	1
ath-MIR170_MI0000213	99,96633512	Yes	19	miR171	1	1	1
ath-MIR171a_MI0000214	98,97857222	Yes	19	miR171	1	1	1
ath-MIR171b_MI0000989	99,5969427	Yes	19	miR171	1	1	1
ath-MIR171c_MI0000990	97,98714824	Yes	20	miR171	1	1	1

ath-MIR172a_MI0000215	99,93400803	Yes	19	miR172	1	1	1
ath-MIR172b_MI0000216	99,9934712	Yes	18	miR172	1	1	1
ath-MIR172c_MI0000991	99,96112999	Yes	19	miR172	1	0	1
ath-MIR172d_MI0000992	99,8607567	Yes	18	miR172	1	1	1
ath-MIR172e_MI0001089	99,90625232	Yes	19	miR172	1	1	1
ath-MIR173_MI0000217	97,21842376	Yes	20	miR173	1	1	1
ath-MIR319a_MI0000544	77,71876274	Yes	17	miR319	Borderline	0	1
ath-MIR319b_MI0000545	89,8192695	Yes	17	miR319	1	1	1
ath-MIR319c_MI0001086	95,30894387	Yes	19	miR319	1	0	1
ath-MIR390a_MI0001000	99,67888119	Yes	20	miR390	1	1	1
ath-MIR390b_MI0001001	99,92747877	Yes	19	miR390	1	1	1
ath-MIR391_MI0001002	98,70642022	Yes	17	miR391	1	1	1
ath-MIR393a_MI0001003	99,73438555	Yes	20	miR393	1	1	1
ath-MIR393b_MI0001004	99,96051652	Yes	21	miR393	1	1	1
ath-MIR394a_MI0001005	99,87347992	Yes	17	miR394	1	1	1
ath-MIR394b_MI0001006	99,48883515	Yes	17	miR394	1	1	1
ath-MIR395a_MI0001007	99,87231612	Yes	20	miR395	1	0	1
ath-MIR395b_MI0001008	99,53484642	Yes	19	miR395	1	0	1
ath-MIR395c_MI0001009	99,54282231	Yes	19	miR395	1	0	1
ath-MIR395d_MI0001010	99,93642945	Yes	20	miR395	1	0	1
ath-MIR395e_MI0001011	99,76392027	Yes	20	miR395	1	0	1
ath-MIR395f_MI0001012	99,84589153	Yes	20	miR395	1	0	1
ath-MIR396a_MI0001013	98,03881817	Yes	19	miR396	1	1	1
ath-MIR396b_MI0001014	99,92817988	Yes	20	miR396	1	1	1
ath-MIR397a_MI0001015	90,07006788	No	19	miR397	1	0	1
ath-MIR397b_MI0001016	99,95992386	Yes	19	miR397	1	0	1
ath-MIR398a_MI0001017	99,67294826	Yes	18	miR398	1		1
ath-MIR398b_MI0001018	99,21945989	Yes	19	miR398	1	1	1
ath-MIR398c_MI0001019	99,30918932	Yes	18	miR398	1	1	1
ath-MIR399a_MI0001020	99,90543736	Yes	20	miR399	1	0	1
ath-MIR399b_MI0001021	98,54517819	Yes	19	miR399	1	0	1
ath-MIR399c_MI0001022	99,67695175	Yes	19	miR399	1	1	1
ath-MIR399d_MI0001023	99,84101749	Yes	20	miR399	1	0	1
ath-MIR399e_MI0001024	97,82608695	No	17	miR399	1	0	1
ath-MIR399f_MI0001025	99,88296357	Yes	18	miR399	1	0	1
ath-MIR400_MI0001069	98,49273962	Yes	20	miR400	1	1	1

ath-MIR401_MI0001070	15,823153	No	18	UNKNOWN	0	0	0
ath-MIR402_MI0001071	98,31123517	Yes	20	miR402	1	0	1
ath-MIR403_MI0001072	92,53%	7,46%	99,98689248	Yes	1	1	1
ath-MIR404_MI0001073	18,97%	0,15%	19,11913601	No	0	0	0
ath-MIR405a_MI0001074	51,76%	0,28%	52,03781644	No	0	0	0
ath-MIR405b_MI0001075	38,14%	0,49%	38,62228069	No	0	0	0
ath-MIR405d_MI0001077	33,37%	2,50%	35,87061688	No	0	0	0
ath-MIR406_MI0001078	39,13%	14,20%	53,32359387	No	0	0	0
ath-MIR407_MI0001079	24,81%	5,86%	30,67023232	Yes	0	0	0
ath-MIR408_MI0001080	99,35595141	Yes	17	miR408	1	1	1
ath-MIR413_MI0001424	33,33333333	No	18	miR413	0	0	0
ath-MIR414_MI0001425	65,99496222	No	7	UNKNOWN	0	0	0
ath-MIR415_MI0001426	55	No	12	miR415	0	0	0
ath-MIR416_MI0001427	25,63090128	No	16	UNKNOWN	0	0	0
ath-MIR417_MI0001428	50	No	16	UNKNOWN	0	0	0
ath-MIR418_MI0001429	100	No	16	UNKNOWN	Lacks evidence	0	0
ath-MIR419_MI0001430	40	No	18	miR419	0	0	0
ath-MIR420_MI0001431	80	No	Mature occurs on a loop	UNKNOWN	0	0	0
ath-MIR426_MI0001441	60,30927835	No	13	UNKNOWN	0	0	0
ath-MIR447a_MI0002407	51,07553909	Yes	17	miR447	0	0	0
ath-MIR447b_MI0002408	51,04098722	Yes	17	miR447	0	0	0
ath-MIR447c_MI0002409	35,62913907	No	16	miR447	0	0	0
ath-MIR472_MI0005102	29,83507178	Yes	20	UNKNOWN	0	0	1
ath-MIR771_MI0005101	96,80727928	Yes	19	miR771	1	0	1
ath-MIR773a_MI0005103	96,32257693	Yes	19	miR773	1	0	1
ath-MIR773b_MI0014666	40,50046339	Yes	20	UNKNOWN	0	0	1
ath-MIR774a_MI0005104	79,96474735	Yes	20	miR774	Borderline	0	1
ath-MIR774b_MI0014668	36,84210526	No	25	miR774	0	0	1
ath-MIR775_MI0005105	98,07410509	Yes	16	miR775	1	0	1
ath-MIR776_MI0005106	89,34756644	Yes	18	miR776	1	0	1
ath-MIR777_MI0005107	98,9693446	Yes	20	miR777	1	0	1
ath-MIR778_MI0005108	58,68818169	Yes	20	miR778	0	0	0
ath-MIR779_MI0005109	93,38801374	Yes	19	miR779	1	0	0

ath-MIR780_MI0005110	76,98030208	Yes	19	miR780	Borderline	0	0
ath-MIR781a_MI0005111	97,31198735	Yes	21	miR781	1	0	1
ath-MIR781b_MI0019231	99,66317816	Yes	20	miR781	1	0	0
ath-MIR782_MI0005112	87,94326241	Yes	18	miR782	1	0	0
ath-MIR822_MI0005379	79,29804439	Yes	20	miR822	Borderline	1	0
ath-MIR823_MI0005380	96,80512889	Yes	19	miR823	1	0	1
ath-MIR824_MI0005409	99,87685807	Yes	20	miR824	1	1	1
ath-MIR825_MI0005381	94,59355052	Yes	17	miR825	1	0	0
ath-MIR826a_MI0005382	93,42342342	No	12	miR826	0	0	1
ath-MIR826b_MI0029231	57,5	Yes	20	miR826	0	0	1
ath-MIR827_MI0005383	99,89973178	Yes	17	miR827	1	0	1
ath-MIR828_MI0005384	94,31455898	No	20	miR828	1	0	0
ath-MIR829_MI0005385	93,18657299	Yes	20	miR829	1	0	1
ath-MIR830_MI0005386	89,43877551	Yes	18	miR830	1	0	0
ath-MIR831_MI0005387	61,30625686	Yes	19	miR831	0	0	0
ath-MIR832_MI0005388	72,86356822	Yes	19	miR832	0	0	0
ath-MIR833a_MI0005389	83,46859413	Yes	21	miR833	Borderline	0	0
ath-MIR833b_MI0019230	79,06394453	Yes	20	miR833	Borderline	0	1
ath-MIR834_MI0005390	92,37012987	Yes	17	miR834	1	0	0
ath-MIR835_MI0005391	61,15245182	Yes	18	UNKNOWN	0	0	0
ath-MIR836_MI0005392	90,41916168	Yes	18	UNKNOWN	1	0	1
ath-MIR837_MI0005393	88,87771141	Yes	20	miR837	1	1	0
ath-MIR838_MI0005394	76,07904316	Yes	19	miR838	Borderline	0	0
ath-MIR839_MI0005395	37,1753394	Yes	21	miR839	0	0	1
ath-MIR840_MI0005396	93,69906962	Yes	18	miR840	1	1	0
ath-MIR841a_MI0005397	69,40802983	Yes	18	miR841	0	0	0
ath-MIR841b_MI0014669	74,13112365	Yes	19	miR841	0	0	1
ath-MIR842_MI0005398	91,17850125	Yes	19	miR842	1	0	0
ath-MIR843_MI0005399	99,7876203	Yes	20	miR843	1	0	0
ath-MIR844_MI0005400	64,50396367	Yes	20	miR844	0	1	1
ath-MIR845a_MI0005401	99,63612614	Yes	19	miR845	1	0	0
ath-MIR845b_MI0005444	99,79253112	Yes	19	miR845	1	0	1
ath-MIR846_MI0005402	98,03772589	Yes	18	miR846	1	1	1
ath-MIR847_MI0005410	88,56437984	Yes	17	miR847	1	0	0
ath-MIR848_MI0005403	78,08961103	Yes	22	miR848	Borderline	0	1
ath-MIR849_MI0005404	50,02307337	Yes	20	miR849	0	0	0

ath-MIR850_MI0005405	46,55764428	Yes	18	miR850	0	0	0
ath-MIR851_MI0005406	86,54203293	Yes	17	miR851	1	1	1
ath-MIR852_MI0005407	98,72370796	Yes	21	miR852	1	0	1
ath-MIR853_MI0005408	95,5109088	Yes	18	miR853	1	0	0
ath-MIR854a_MI0005412	13,51857704	No	13	UNKNOWN	0	0	0
ath-MIR854b_MI0005413	14,0494742	No	16	UNKNOWN	0	0	0
ath-MIR854c_MI0005414	17,45860331	No	6	UNKNOWN	0	0	0
ath-MIR854d_MI0005415	14,0494742	No	16	UNKNOWN	0	0	0
ath-MIR854e_MI0015956	14,0494742	No	16	UNKNOWN	0	0	0
ath-MIR855_MI0005411	31,74331551	No	Mature occurs on a loop	UNKNOWN	0	0	1
ath-MIR856_MI0005433	34,70542835	Yes	17	miR856	0	0	1
ath-MIR857_MI0005434	98,83973895	Yes	17	miR857	1	0	1
ath-MIR858a_MI0005435	99,12124603	No	19	miR858	1	0	1
ath-MIR858b_MI0019228	99,28079666	No	15	miR858	1	0	1
ath-MIR859_MI0005436	81,24649073	Yes	20	miR859	Borderline	0	0
ath-MIR860_MI0005437	61,44406706	Yes	21	miR860	0	0	1
ath-MIR861_MI0005438	81,4497944	Yes	19	miR861	Borderline	0	0
ath-MIR862_MI0005439	64,28385755	Yes	17	miR862	0	0	0
ath-MIR863_MI0005440	77,63570922	Yes	19	miR863	Borderline	0	1
ath-MIR864_MI0005441	98,10253388	Yes	19	miR864	1	1	0
ath-MIR865_MI0005442	69,91937726	Yes	16	UNKNOWN	0	0	0
ath-MIR866_MI0005443	98,55263158	Yes	21	miR866	1	0	1
ath-MIR867_MI0005445	98,92376682	Yes	18	miR867	1	0	1
ath-MIR868_MI0005446	60,58637966	No	17	miR868	0	0	0
ath-MIR869_MI0005447	48,95486936	Yes	18	miR869	0	0	0
ath-MIR870_MI0005448	57,75234131	No	17	UNKNOWN	0	0	0
ath-MIR1886_MI0008303	92,062856	Yes	21	miR1886	1	0	0
ath-MIR1887_MI0008304	99,77836879	Yes	21	miR1887	1	0	0
ath-MIR1888a_MI0008305	70,88056255	No	19	miR1888	0	0	0
ath-MIR1888b_MI0019247	35,36523071	Yes	23	UNKNOWN	0	0	1
ath-MIR2111a_MI0010630	98,66028089	Yes	20	miR2111	1	0	1
ath-MIR2111b_MI0010631	99,43206641	Yes	18	miR2111	1	0	0
ath-MIR2112_MI0010632	46,8124309	No	20	miR2112	0	0	0

ath-MIR2933a_MI0013362	8,52763231	No	Mature occurs on a loop	UNKNOWN	0	0	0
ath-MIR2933b_MI0013363	18,31958493	Yes	17	miR2933	0	0	0
ath-MIR2934_MI0013364	62,27937469	Yes	20	miR2934	0	0	0
ath-MIR2936_MI0013366	34,78618421	Yes	20	miR2936	0	0	0
ath-MIR2937_MI0013367	50,95913261	No	17	miR2937	0	0	0
ath-MIR2938_MI0013368	41,17647059	No	22	miR2938	0	0	1
ath-MIR2939_MI0013369	66,59856997	Yes	19	miR2939	0	0	0
ath-MIR3434_MI0014667	58,5259213	Yes	15	miR3434	0	0	0
ath-MIR3440b_MI0015818	85,21592132	No	19	miR3440	Lacks evidence	0	0
ath-MIR3932a_MI0016587	70,24021353	No	19	UNKNOWN	0	0	0
ath-MIR3932b_MI0016588	59,26428062	Yes	18	UNKNOWN	0	0	1
ath-MIR3933_MI0016589	85,29460998	Yes	19	miR3933	1	0	1
ath-MIR4221_MI0015815	84,92035398	Yes	18	miR4221	Borderline	0	0
ath-MIR4227_MI0015816	92,5	No	Mature occurs on a loop	UNKNOWN	0	0	0
ath-MIR4228_MI0015817	41,77863636	No	11	miR-1357	0	0	0
ath-MIR4239_MI0015819	19,89966555	No	Mature occurs on a loop	UNKNOWN	0	0	0
ath-MIR4240_MI0015820	78,9135514	Yes	19	miR4240	Borderline	0	0
ath-MIR4243_MI0015821	97,98449612	Yes	20	miR4243	1	0	0
ath-MIR4245_MI0020194	85,54741554	Yes	22	miR4245	1	0	1
ath-MIR5012_MI0017880	98,36531627	No	18	miR5012	Lacks evidence	0	0
ath-MIR5013_MI0017881	51,49700599	Yes	20	miR5013	0	0	0
ath-MIR5014a_MI0017882	84,47039973	Yes	20	miR5014	Borderline	0	0
ath-MIR5014b_MI0020195	27,31719332	Yes	22	miR5014	0	0	0
ath-MIR5015_MI0017883	27,90697674	Yes	19	UNKNOWN	0	0	0
ath-MIR5016_MI0017884	53,00387597	Yes	16	UNKNOWN	0	0	1
ath-MIR5017_MI0017885	92,29642648	Yes	18	miR5017	1	0	0
ath-MIR5018_MI0017886	70,87022242	Yes	22	UNKNOWN	0	0	0
ath-MIR5019_MI0017887	95,14563107	Yes	17	miR5019	1	0	0
ath-MIR5020a_MI0017888	74,66318992	No	19	miR5020	0	0	0
ath-MIR5020b_MI0017893	99,06115982	Yes	21	miR5020	1	0	0
ath-MIR5020c_MI0019203	75,31308972	Yes	21	miR5020	Borderline	0	0
ath-MIR5021_MI0017889	76,47058823	No	17	miR5021	0	0	0
ath-MIR5022_MI0017891	94,05940594	Yes	20	miR5022	1	0	0

ath-MIR5023_MI0017892	98,29299007	Yes	20	miR5023	1	0	1
ath-MIR5024_MI0017894	38,96457765	Yes	19	miR5024	0	0	0
ath-MIR5025_MI0017895	65,59139785	Yes	21	miR5025	0	0	1
ath-MIR5026_MI0017896	90,95615058	Yes	17	miR5026	1	0	0
ath-MIR5027_MI0017897	59,58770906	No	18	miR5027	0	0	1
ath-MIR5028_MI0017898	98,83097905	Yes	16	miR5028	1	0	0
ath-MIR5029_MI0017899	59,38136914	Yes	21	miR5029	0	0	0
ath-MIR5595a_MI0020188	97,68941429	Yes	21	miR5995	1	0	0
ath-MIR5628_MI0019199	17,20766524	Yes	20	UNKNOWN	0	0	0
ath-MIR5629_MI0019200	24,16283326	Yes	20	UNKNOWN	0	0	0
ath-MIR5630a_MI0019201	33,02752293	No	19	miR5630	0	0	0
ath-MIR5630b_MI0019211	33,02752293	No	19	miR5630	0	0	0
ath-MIR5631_MI0019202	94,89795918	Yes	20	miR5631	1	0	0
ath-MIR5632_MI0019204	30,03838772	Yes	21	UNKNOWN	0	0	0
ath-MIR5633_MI0019205	61,37879911	Yes	20	miR5633	0	0	0
ath-MIR5634_MI0019206	76,4832794	Yes	20	miR5634	Borderline	0	0
ath-MIR5635a_MI0019207	31,56495016	No	23	UNKNOWN	0	0	0
ath-MIR5635b_MI0019229	38,68130334	Yes	21	UNKNOWN	0	0	0
ath-MIR5635c_MI0019239	18,86137921	No	22	UNKNOWN	0	0	0
ath-MIR5635d_MI0019217	17,38497841	No	22	UNKNOWN	0	0	0
ath-MIR5636_MI0019208	40,02659574	No	17	UNKNOWN	0	0	0
ath-MIR5637_MI0019209	79,3918919	Yes	19	miR5637	Borderline	0	0
ath-MIR5638a_MI0019210	17,95499021	Yes	21	UNKNOWN	0	0	0
ath-MIR5638b_MI0019219	68,91025641	Yes	19	miR5638	0	0	0
ath-MIR5639_MI0019212	70,34930407	Yes	20	miR5639	0	0	0
ath-MIR5640_MI0019213	94,809834	Yes	17	miR5640	1	0	0
ath-MIR5641_MI0019214	43,7447757	No	19	miR5641	0	0	0
ath-MIR5642a_MI0019215	26,9948439	No	Mature occurs on a loop	UNKNOWN	0	0	0
ath-MIR5642b_MI0019246	28,35915838	No	Mature occurs on a loop	UNKNOWN	0	0	0
ath-MIR5643a_MI0019216	38,82237166	No	16	miR5643	0	0	0
ath-MIR5643b_MI0019256	39,37670999	No	13	UNKNOWN	0	0	0
ath-MIR5644_MI0019218	40,54698717	No	19	miR5644	0	0	0
ath-MIR5645a_MI0019220	39,28537147	Yes	23	miR5645	0	0	0

ath-MIR5645b_MI0019221	21,67753961	No	13	UNKNOWN	0	0	0
ath-MIR5645c_MI0019225	18,40106343	No	Mature occurs on a loop	UNKNOWN	0	0	0
ath-MIR5645d_MI0019244	88,5098073	No	3	UNKNOWN	0	0	0
ath-MIR5645e_MI0019257	38,13010182	No	22	miR5645	0	0	0
ath-MIR5645f_MI0019252	46,76674049	No	21	miR5645	0	0	0
ath-MIR5646_MI0019222	62,2706422	No	18	miR5646	0	0	0
ath-MIR5647_MI0019223	27,86822301	Yes	16	miR5647	0	0	0
ath-MIR5648_MI0019224	59,50150257	Yes	18	miR5648	0	0	0
ath-MIR5649a_MI0019226	51,66666667	No	16	UNKNOWN	0	0	0
ath-MIR5649b_MI0019248	35,88621445	No	21	miR5649	0	0	0
ath-MIR5650_MI0019227	89,94668697	Yes	21	miR5650	1	0	0
ath-MIR5651_MI0019233	84,67966574	No	16	miR5651	0	0	0
ath-MIR5652_MI0019235	35,59593975	No	Mature occurs on a loop	UNKNOWN	0	0	0
ath-MIR5653_MI0019236	45,13547827	No	22	miR5653	0	0	0
ath-MIR5654_MI0019237	99,42232309	Yes	21	miR5654	1	0	0
ath-MIR5655_MI0019238	13,29105897	Yes	21	miR5655	0	0	0
ath-MIR5656_MI0019240	51,20113048	No	17	UNKNOWN	0	0	0
ath-MIR5657_MI0019241	38,1803598	No	Mature occurs on a loop	UNKNOWN	0	0	0
ath-MIR5658_MI0019242	74,01129944	No	15	UNKNOWN	0	0	0
ath-MIR5659_MI0019243	66,78405672	No	22	miR5659	0	0	1
ath-MIR5660_MI0019245	21,70542636	No	Mature occurs on a loop	UNKNOWN	0	0	0
ath-MIR5661_MI0019249	91,58770511	No	17	miR5661	Lacks evidence	0	0
ath-MIR5662_MI0019250	42,93123641	Yes	23	miR5662	0	0	0
ath-MIR5663_MI0019251	93,49158732	Yes	21	miR5663	1	1	0
ath-MIR5664_MI0019253	39,98073217	Yes	20	UNKNOWN	0	0	0
ath-MIR5665_MI0019254	31,58573963	Yes	22	UNKNOWN	0	0	0
ath-MIR5666_MI0019255	28,51083883	Yes	20	miR5666	0	0	0
ath-MIR5995b_MI0020189	73,64903138	Yes	24	miR5995	0	0	0
ath-MIR5996_MI0020190	69,6930009	Yes	21	miR5996	0	0	0
ath-MIR5997_MI0020191	63,81984675	Yes	21	miR5997	0	0	0
ath-MIR5998a_MI0020192	60,61929337	Yes	21	miR5998	0	0	0
ath-MIR5998b_MI0020193	59,87725354	Yes	21	miR5998	0	0	0

ath-MIR5999_MI0020196	93,53741497	No	20	miR5999	Lacks evidence	0	0
ath-MIR8121_MI0026055	38,64171374	No	21	UNKNOWN	0	0	0
ath-MIR8165_MI0026793	72,64522031	No	20	UNKNOWN	0	0	0
ath-MIR8166_MI0026794	54,70205285	No	19	UNKNOWN	0	0	0
ath-MIR8167a_MI0026795	32,71095945	No	18	UNKNOWN	0	0	0
ath-MIR8167b_MI0026796	32,71095945	No	18	UNKNOWN	0	0	0
ath-MIR8167c_MI0026797	32,71095945	No	18	UNKNOWN	0	0	0
ath-MIR8167d_MI0031739	32,71095945	No	18	UNKNOWN	0	0	0
ath-MIR8167e_MI0031740	32,71095945	No	18	UNKNOWN	0	0	0
ath-MIR8167f_MI0031741	32,71095945	No	18	UNKNOWN	0	0	0
ath-MIR8168_MI0026798	69,08068783	No	Mature occurs on a loop	UNKNOWN	0	0	0
ath-MIR8169_MI0026799	55,71667596	No	18	UNKNOWN	0	0	0
ath-MIR8170_MI0026800	90,73543457	Yes	17	UNKNOWN	1	0	0
ath-MIR8171_MI0026801	41,71060864	No	16	UNKNOWN	0	0	0
ath-MIR8172_MI0026802	23,26519879	No	15	UNKNOWN	0	0	0
ath-MIR8173_MI0026803	42,17123756	No	20	UNKNOWN	0	0	0
ath-MIR8174_MI0026804	61,36930793	No	17	UNKNOWN	0	0	0
ath-MIR8175_MI0026805	73,55595121	No	17	UNKNOWN	0	0	0
ath-MIR8176_MI0026806	41,92559565	Yes	19	UNKNOWN	0	0	0
ath-MIR8177_MI0026807	33,72755814	No	Mature occurs on a loop	UNKNOWN	0	0	0
ath-MIR8178_MI0026808	19,5026178	No	22	UNKNOWN	0	0	0
ath-MIR8179_MI0026809	97,53032112	Yes	20	UNKNOWN	1	0	0
ath-MIR8180_MI0026810	63,98363489	Yes	20	UNKNOWN	0	0	0
ath-MIR8181_MI0026811	93,96396396	No	16	UNKNOWN	Lacks evidence	0	0
ath-MIR8182_MI0026812	91,1392405	No	18	UNKNOWN	Lacks evidence	0	0
ath-MIR8183_MI0026813	64,81924035	No	18	UNKNOWN	0	0	0
ath-MIR8184_MI0026814	97,91183294	Yes	21	UNKNOWN	1	0	0

Appendix B Other Work Arising from this Thesis

This appendix includes a publication to which I contributed: **Assessing vulnerability of two Mediterranean conifers to support genetic conservation management in the face of climate change** (Serra-Varela, 2017). My contribution was to help perform the genetic analysis, intellectual input, as well as provide support during the write up stage. The publication is included here for completeness.

DOI: 10.1111/ddi.12544

BIODIVERSITY RESEARCH

WILEY Diversity and Distributions A Journal of Conservation Biogeography

Assessing vulnerability of two Mediterranean conifers to support genetic conservation management in the face of climate change

María Jesús Serra-Varela[1,2,3] | Ricardo Alía[2,3] | Rose Ruiz Daniels[3] | Niklaus E. Zimmermann[4,5] | Julián Gonzalo-Jiménez[1,2]* | Delphine Grivet[2,3]*

[1]Department of Plant Production and Forest Resources, University of Valladolid, Palencia, Spain

[2]Sustainable Forest Management Research Institute, INIA—University of Valladolid, Palencia, Spain

[3]Department of Forest Ecology and Genetics, INIA, Forest Research Centre, Madrid, Spain

[4]Landscape Dynamics, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

[5]Department of Environmental Systems Science, Swiss Federal Institute of Technology ETH, Zürich, Switzerland

Correspondence
Julián Gonzalo-Jiménez, Department of Plant Production and Forest Resources, University of Valladolid, Palencia, Spain.
Email: jgonzalo@pvs.uva.es

Funding information
We acknowledge funding from COST Action FP1202 that enabled the collaboration between the different groups involved, which has supported the development of this paper. We also thank funding from the Spanish National Research Plan (AdapCon CGL2011-30182-C02-01 and RTA2013-00048-C03-01), as well as from the ERA-Net BiodivERsA (LinkTree project, EUI2008-03713 and TipTree BiodivERsA2012-15) and Era-NET Foresterra (PCIN-2014-138). M.J.S.V. was supported by a FPU from the Spanish Ministry of Education, Culture and Sport; D.G. was supported by a Ramón y Cajal fellowship from the Spanish Ministry of Science and Innovation; R.R.D. was supported by a FPI from the Spanish Ministry of Economy and Competitiveness; N.E.Z. was supported by an SNSF Grant (No: 31003A_149508/1).

Editor: Lluís Brotons

Abstract

Aim: To integrate two major components of vulnerability to climate change: adaptive capacity (approached by genetic groups) and exposure (approached by risk of habitat loss) illustrated with the maritime (*Pinus pinaster* Ait.) and Aleppo (*Pinus halepensis* Mill.) pines. To integrate such information in the selection of conservation strategies (ex situ vs. in situ) and to evaluate current European efforts in the conservation of forest genetic resources.

Location: Mediterranean Basin and European Atlantic coast.

Methods: With the objective of preserving the overall genetic diversity of our two target species, we individually assess each of their genetic groups. We fit a species distribution model and project it to current climate and 42 different future climatic predictions. We create future suitability maps to assess risk of habitat loss based on the number of future climate projections that predict suitability. According to this assessment on the risk of habitat loss, we propose suitable conservation strategies.

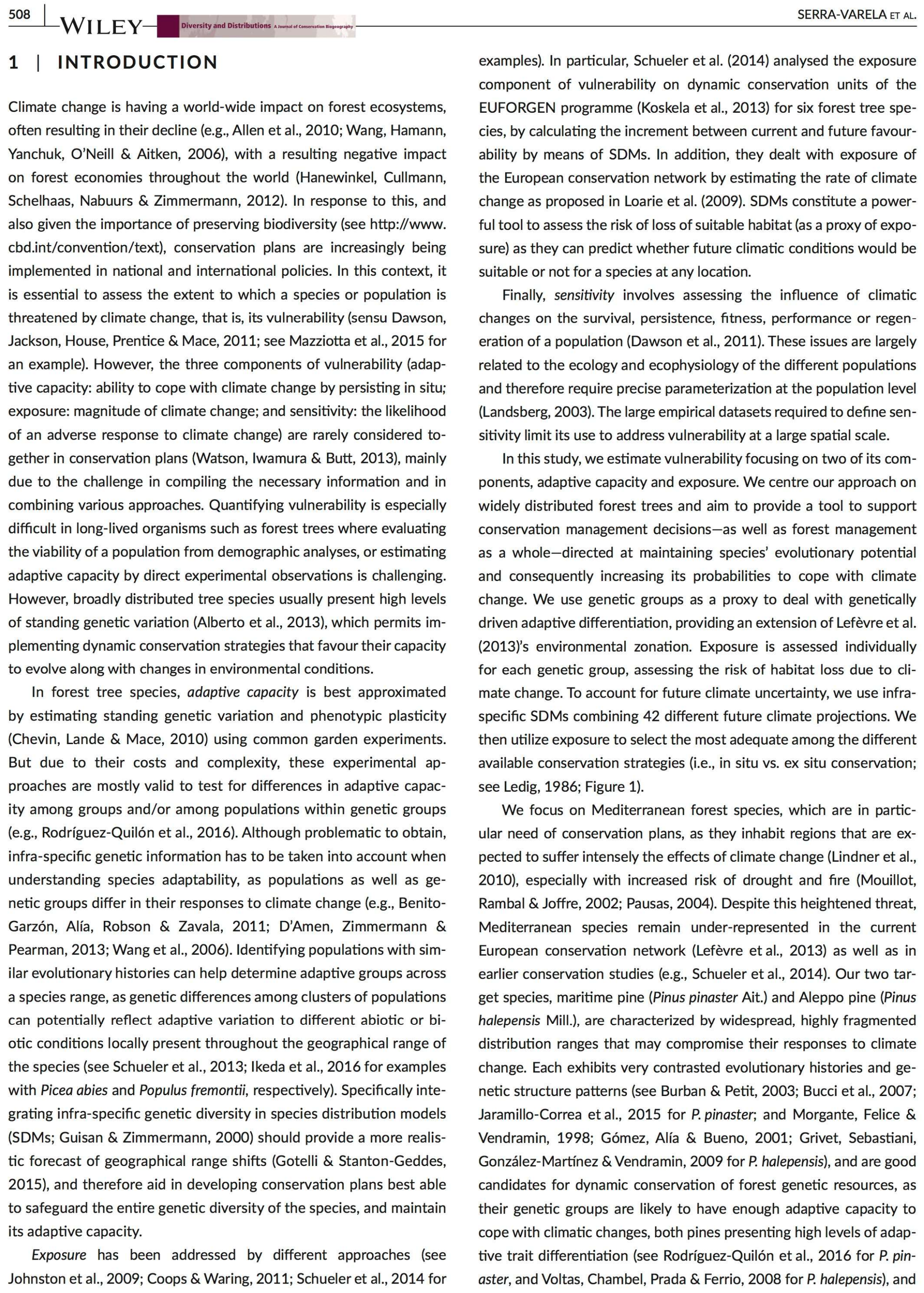
Results: We found areas suitable for in situ conservation for most of the genetic groups, the exception being the central–eastern–southern Iberian genetic groups of maritime pine and the Moroccan genetic group of Aleppo pine which required ex situ conservation. In the current European conservation network, three genetic groups for maritime pine and two for Aleppo pine remain unrepresented, and the representation of the rest is unbalanced.

Main conclusions: We provide a tool to support conservation management of forest trees, an increasingly important task given the negative impact of climate change on forest ecosystems. We also provide a framework to increase the efficiency of the European conservation network: (i) exposure assessment should be considered as a requirement for a population to become a dynamic conservation unit (DCU); and (ii) as illustrated with our two target species, the selection of DCUs should represent all existing genetic groups.

KEYWORDS
Aleppo pine, conservation biology, maritime pine, *Pinus halepensis*, *Pinus pinaster*, species distribution models

*These authors contributed equally to the supervision of the work.

z



1 | INTRODUCTION

Climate change is having a world-wide impact on forest ecosystems, often resulting in their decline (e.g., Allen et al., 2010; Wang, Hamann, Yanchuk, O'Neill & Aitken, 2006), with a resulting negative impact on forest economies throughout the world (Hanewinkel, Cullmann, Schelhaas, Nabuurs & Zimmermann, 2012). In response to this, and also given the importance of preserving biodiversity (see http://www.cbd.int/convention/text), conservation plans are increasingly being implemented in national and international policies. In this context, it is essential to assess the extent to which a species or population is threatened by climate change, that is, its vulnerability (sensu Dawson, Jackson, House, Prentice & Mace, 2011; see Mazziotta et al., 2015 for an example). However, the three components of vulnerability (adaptive capacity: ability to cope with climate change by persisting in situ; exposure: magnitude of climate change; and sensitivity: the likelihood of an adverse response to climate change) are rarely considered together in conservation plans (Watson, Iwamura & Butt, 2013), mainly due to the challenge in compiling the necessary information and in combining various approaches. Quantifying vulnerability is especially difficult in long-lived organisms such as forest trees where evaluating the viability of a population from demographic analyses, or estimating adaptive capacity by direct experimental observations is challenging. However, broadly distributed tree species usually present high levels of standing genetic variation (Alberto et al., 2013), which permits implementing dynamic conservation strategies that favour their capacity to evolve along with changes in environmental conditions.

In forest tree species, *adaptive capacity* is best approximated by estimating standing genetic variation and phenotypic plasticity (Chevin, Lande & Mace, 2010) using common garden experiments. But due to their costs and complexity, these experimental approaches are mostly valid to test for differences in adaptive capacity among groups and/or among populations within genetic groups (e.g., Rodríguez-Quilón et al., 2016). Although problematic to obtain, infra-specific genetic information has to be taken into account when understanding species adaptability, as populations as well as genetic groups differ in their responses to climate change (e.g., Benito-Garzón, Alía, Robson & Zavala, 2011; D'Amen, Zimmermann & Pearman, 2013; Wang et al., 2006). Identifying populations with similar evolutionary histories can help determine adaptive groups across a species range, as genetic differences among clusters of populations can potentially reflect adaptive variation to different abiotic or biotic conditions locally present throughout the geographical range of the species (see Schueler et al., 2013; Ikeda et al., 2016 for examples with *Picea abies* and *Populus fremontii*, respectively). Specifically integrating infra-specific genetic diversity in species distribution models (SDMs; Guisan & Zimmermann, 2000) should provide a more realistic forecast of geographical range shifts (Gotelli & Stanton-Geddes, 2015), and therefore aid in developing conservation plans best able to safeguard the entire genetic diversity of the species, and maintain its adaptive capacity.

Exposure has been addressed by different approaches (see Johnston et al., 2009; Coops & Waring, 2011; Schueler et al., 2014 for examples). In particular, Schueler et al. (2014) analysed the exposure component of vulnerability on dynamic conservation units of the EUFORGEN programme (Koskela et al., 2013) for six forest tree species, by calculating the increment between current and future favourability by means of SDMs. In addition, they dealt with exposure of the European conservation network by estimating the rate of climate change as proposed in Loarie et al. (2009). SDMs constitute a powerful tool to assess the risk of loss of suitable habitat (as a proxy of exposure) as they can predict whether future climatic conditions would be suitable or not for a species at any location.

Finally, *sensitivity* involves assessing the influence of climatic changes on the survival, persistence, fitness, performance or regeneration of a population (Dawson et al., 2011). These issues are largely related to the ecology and ecophysiology of the different populations and therefore require precise parameterization at the population level (Landsberg, 2003). The large empirical datasets required to define sensitivity limit its use to address vulnerability at a large spatial scale.

In this study, we estimate vulnerability focusing on two of its components, adaptive capacity and exposure. We centre our approach on widely distributed forest trees and aim to provide a tool to support conservation management decisions—as well as forest management as a whole—directed at maintaining species' evolutionary potential and consequently increasing its probabilities to cope with climate change. We use genetic groups as a proxy to deal with genetically driven adaptive differentiation, providing an extension of Lefèvre et al. (2013)'s environmental zonation. Exposure is assessed individually for each genetic group, assessing the risk of habitat loss due to climate change. To account for future climate uncertainty, we use infra-specific SDMs combining 42 different future climate projections. We then utilize exposure to select the most adequate among the different available conservation strategies (i.e., in situ vs. ex situ conservation; see Ledig, 1986; Figure 1).

We focus on Mediterranean forest species, which are in particular need of conservation plans, as they inhabit regions that are expected to suffer intensely the effects of climate change (Lindner et al., 2010), especially with increased risk of drought and fire (Mouillot, Rambal & Joffre, 2002; Pausas, 2004). Despite this heightened threat, Mediterranean species remain under-represented in the current European conservation network (Lefèvre et al., 2013) as well as in earlier conservation studies (e.g., Schueler et al., 2014). Our two target species, maritime pine (*Pinus pinaster* Ait.) and Aleppo pine (*Pinus halepensis* Mill.), are characterized by widespread, highly fragmented distribution ranges that may compromise their responses to climate change. Each exhibits very contrasted evolutionary histories and genetic structure patterns (see Burban & Petit, 2003; Bucci et al., 2007; Jaramillo-Correa et al., 2015 for *P. pinaster*; and Morgante, Felice & Vendramin, 1998; Gómez, Alía & Bueno, 2001; Grivet, Sebastiani, González-Martínez & Vendramin, 2009 for *P. halepensis*), and are good candidates for dynamic conservation of forest genetic resources, as their genetic groups are likely to have enough adaptive capacity to cope with climatic changes, both pines presenting high levels of adaptive trait differentiation (see Rodríguez-Quilón et al., 2016 for *P. pinaster*, and Voltas, Chambel, Prada & Ferrio, 2008 for *P. halepensis*), and

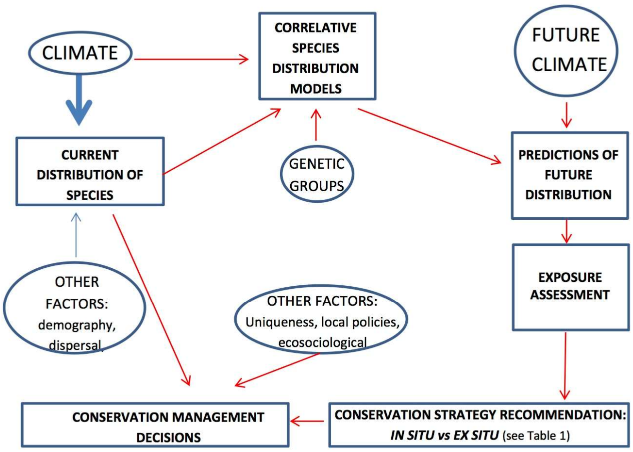

CLIMATE
CORRELATIVE SPECIES DISTRIBUTION MODELS
FUTURE CLIMATE
CURRENT DISTRIBUTION OF SPECIES
GENETIC GROUPS
PREDICTIONS OF FUTURE DISTRIBUTION
EXPOSURE ASSESSMENT
OTHER FACTORS: demography, dispersal
OTHER FACTORS: Uniqueness, local policies, ecosociological
CONSERVATION MANAGEMENT DECISIONS
CONSERVATION STRATEGY RECOMMENDATION: IN SITU vs EX SITU (see Table 1)

FIGURE 1 Framework for conservation management decisions. [Colour figure can be viewed at wileyonlinelibrary.com]

significant levels of phenotypic plasticity (see Corcuera, Gil-Pelegrin & Notivol, 2010; Corcuera, Cochard, Gil-Pelegrin & Notivol, 2011 for *P. pinaster*, and Baquedano, Valladares & Castillo, 2008; Santos-del-Blanco, Bonser, Valladares, Chambel & Climent, 2013 for *P. halepensis*). Despite their potential adaptive capacity, it is challenging to forecast how these two pines will respond to the pressures of climate change, as it highly depends on the environmental conditions and on the traits under consideration (Alía, Chambel, Notivol, Climent & González-Martínez, 2014). Therefore, as we still cannot quantify the amount of adaptive capacity necessary to adapt in situ, we propose here, as an alternative, a monitoring program to check for signals of maladaptation. Finally, focusing on these two target species enables the evaluation of current efforts in conserving their genetic resources in Europe (as defined in EUFORGEN), by assessing the exposure of their currently defined dynamic conservation units to future climate change and whether genetic groups are appropriately represented.

2 | METHODS

2.1 | Molecular data and definition of genetic groups

We obtained eight genetically defined genetic groups for the full distribution range of *P. pinaster* from Serra-Varela et al. (2015) (Figure 2a) namely Atlantic Iberian Peninsula (G1-pin), Eastern populations (G2-pin), Atlantic France (G3-pin), Morocco (G4-pin), eastern (G5-pin), central (G6-pin) and southern (G7-pin) Spain, and Tunisia (G8-pin) based on mitochondrial, chloroplast and nuclear (simple sequence repeats(SSRs) and single nucleotide polymorphisms (SNPs)) molecular markers. The SNP dataset comprised 772 individuals from 36 populations (see Jaramillo-Correa et al., 2015 for more details).

For *P. halepensis* we detected seven different genetic groups namely central and southern Spain (G1-hal), Balearic and southern France (G2-hal), Tunisian and northern Italian (G3-hal), Moroccan and southern Spain (G4-hal), Greek (G5-hal), central and northern Spain (G6-hal) and northern Spain and southern France (G7-hal; Figure 2b), by performing a Bayesian clustering analysis using STRUCTURE (Pritchard, Stephens & Donnelly, 2000) on a SNP dataset (1325 individuals from 49 populations) covering most of the species natural range (Appendix S1 in Supporting Information for more details). Contrary to *P. pinaster* where the genetic groups were spatially differentiated, *P. halepensis* presented transition areas occupied by more than one genetic group simultaneously and that we defined based on *Q* values as explained in Appendix S1. Some small areas of the distribution of the species (located in Algeria) could not be assigned to any genetic group due to insufficient sampling (see Figure 2b).

For both, *P. pinaster* and *P. halepensis*, the genetic groups were based on a priori neutral molecular markers, which are primarily influenced by demographic processes and not by adaptation. However, some of these markers may also be influenced by adaptive selection (especially the SNPs, e.g., Jaramillo-Correa et al., 2015 for *P. pinaster*).

2.2 | Species data

The complete native range for both species was obtained by combining the Tree Species Distribution for Europe (TSDE; Köble & Seufert, 2001) from the Joint Research Centre's AFOLU data portal (ftp://mars.jrc.ec.europa.eu/Afoludata/Public/DS66/) and the EUFORGEN database from the European forest genetic resources programme (http://www.euforgen.org/distribution-maps/; see Appendix S2 for further details).

We prepared a presence–absence dataset for each genetic group individually. Presences of genetic groups were defined as the subset of the overall presences records that belonged to one specific genetic group. In the case of *P. halepensis* presence records of transition zones were considered as presence records for both genetic groups inhabiting that territory. Possible absences corresponded to all the rest of the territory within the study area where TSDE reported 0% occupancy as well as to presences from other genetic groups. Note that

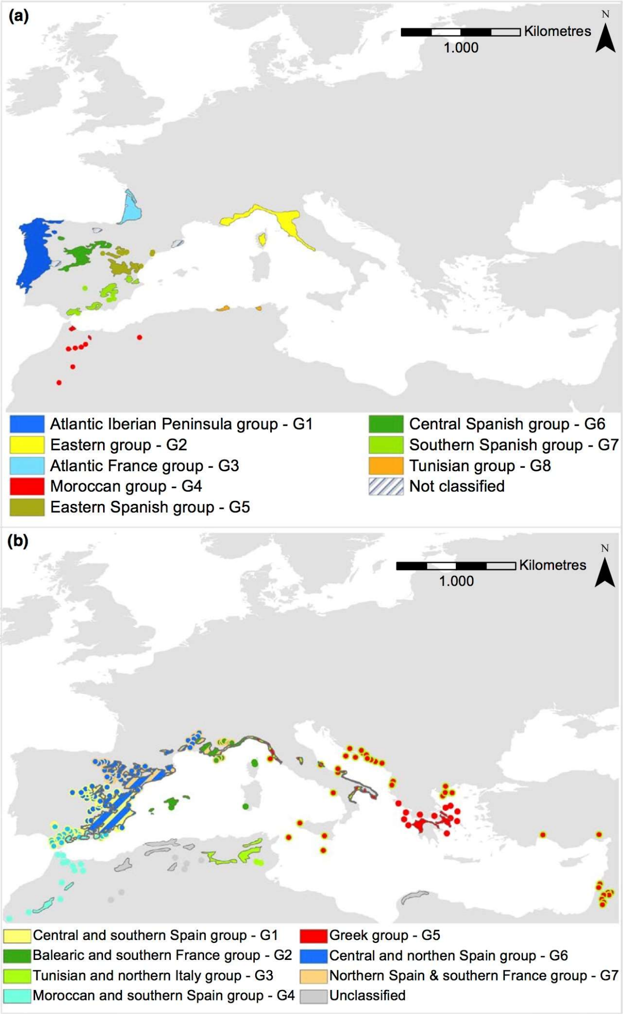

(a)
Kilometres
1.000
N
Atlantic Iberian Peninsula group - G1
Eastern group - G2
Atlantic France group - G3
Moroccan group - G4
Eastern Spanish group - G5
Central Spanish group - G6
Southern Spanish group - G7
Tunisian group - G8
Not classified
(b)
Kilometres
1.000
N
Central and southern Spain group - G1
Balearic and southern France group - G2
Tunisian and northern Italy group - G3
Moroccan and southern Spain group - G4
Greek group - G5
Central and northen Spain group - G6
Northern Spain & southern France group - G7
Unclassified

FIGURE 2 Distribution of the genetic groups of *Pinus pinaster* (modified from Serra-Varela et al., 2015) (a) and *Pinus halepensis* (b) along the natural distribution of the species

areas from the distribution that had not been classified in any of the genetic groups were not considered in the analysis. The numbers of presences for the genetic groups are specified in Appendix S2 while the selection method and the number of selected absences are specified below.

2.3 | Bioclimatic data

We used the 19 bioclimatic variables available in WORLDCLIM (1 Km grid; Hijmans, Cameron, Parra, Jones & Jarvis, 2005) representative of the period between 1950–2000 for the analysis. After screening for correlations and variance inflation, the final set of relevant variables that correlate < 0.75 among each other contained: BIO4 (Temperature Seasonality), BIO11 (Mean Temperature of Coldest Quarter), BIO12 (Annual Precipitation) and BIO18 (Precipitation of Warmest Quarter; see Appendix S2 for more details on variable selection).

Future bioclimatic predictions were also obtained from WORLDCLIM, as these predictions are based on the most recent global climate models (GCMs) projections that have been used in the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment report (IPCC, 2013). We used all GCMs that were simultaneously available for the scenarios of representative concentration pathways (RCP) 2.6, 4.5 and 8.5. This included 14 GCMs (see Appendix S2) and totalled to 42 different future climate predictions.

2.4 | Species distribution models

Five different statistical methods namely general linear model (GLM—McCullagh & Nelder, 1989), generalized additive model (GAM—Hastie & Tibshirani, 1990), Random Forest (RF—Breiman, 2001), classification tree analysis (CTA—Breiman, Friedman, Olshen & Stone, 1984) and MaxEnt (Phillips, Anderson & Schapire, 2006) were individually calibrated, evaluated, converted to binary predictions and then combined to build an ensemble SDM (see more details in Appendix S2). Using an ensemble model, we benefitted from the powerfulness of some algorithms, such as RF, which has shown its high performance in several studies (see Rehfeldt, Crookston, Warwell & Evans, 2006; Ledig, Rehfeldt, Sáenz-Romero & Flores-López, 2010; Leites, Robinson, Rehfeldt, Marshall & Crookston, 2012 for examples) as well as from others, such as GLM and GAM, that could help to circumvent possible fitting to particular cases as well as capturing the general relationship between climate and the distribution of species.

The performance of each algorithm was assessed from the area under the ROC curve (AUC; Fielding & Bell, 1997) and the true skill statistic (TSS; Allouche, Tsoar & Kadmon, 2006), as well as by their sensitivity and specificity, that is, true positive and true negative rates, respectively, from the confusion matrices. While AUC is based on the ROC curve—which plots sensitivity as a function of commission error (1—specificity) along different thresholds—the estimation of TSS is simple (TSS = sensitivity + specificity −1) and has been used as an alternative to Cohen's kappa and the ROC curve (see Mouton, De Baets, Van Broekhoven & Goethals, 2009; Pearman, D'Amen, Graham, Thuiller & Zimmermann, 2010; Benito-Garzón, Ruiz-Benito & Zavala, 2013 for examples).

The probabilistic model outputs of each algorithm were converted to binary maps by defining thresholds that maximized TSS values in a test of initial models against the half of the data not used for model building (split-sample test). Then, to deal with the uncertainty derived from algorithm selection, we combined their predictions by means of a weighted sum, the weights being the TSS scores obtained by each algorithm. Thus, algorithms with better performance had higher influence on the final projections. Following this procedure, we obtained 43 suitability maps for each genetic group (corresponding to current climate and to 42 future climate predictions representing 2050) that ranged from 0.0 to 1.0, with higher scores indicating higher performance and agreement among algorithms.

We set the number of randomly selected absences for all the models to five times the number of total presences used per species (see Appendix S2 for more details). Presences and absences were given weights inversely proportional to their respective numbers, so as to give equal total weights to the two sets as recommended by Barbet-Massin, Jiguet, Albert and Thuiller (2012).

The large number of occurrence records available permitted a random division of each dataset (corresponding to both species and to each genetic group) into two equally sized subsets for training and evaluation, and maintaining the prevalence.

2.5 | Future suitability maps

For each genetic group, we obtained 42 binary (0/1) future projections. None of the 42 future climate predictions is better than another (but see Fordham, Wigley & Brook, 2011). Instead, all are equally likely and we used the sum of all the projections as an indicator of the degree of agreement among models and future climate predictions that the future habitat will be suitable in each cell. Thereby, combining the 42 binary projections, we obtained a map in which values could possibly range from 0.0 (none of the algorithms and future climate predictions was projected to be suitable) to 1.0 (all algorithms and future climate predictions tested were projected to be suitable) with higher scores indicating higher performance and agreement of suitable habitat in the future. Then, we defined three different future suitability categories: "likely suitable" with suitability scores > 0.7 (suitable habitat for more than 70% of future projections), "uncertain" with a suitability score of 0.4–0.7, and "likely unsuitable" with suitability scores < 0.4. We performed this analysis for each genetic group individually.

2.6 | Defining conservation strategies based on exposure to climate change

Future suitability maps were overlaid with maps of current distribution to assess risks of habitat loss (a proxy of exposure), and set the basis for the recommendation of the different conservation strategies within the current distribution of each genetic group (Table 1). In situ conservation is an effective conservation strategy for populations highlighted as "likely suitable" and "uncertain" by future suitability maps, as the high level of standing variation and phenotypic plasticity reported for the two target species should enable them to adapt in situ. However, it should always be associated to a monitoring programme (see Graudal et al., 2014) in order to: (i) ascertain that genetic processes linked to the adaptability of the species are maintained over time and (ii) apply more intense management aimed at supporting local adaptation processes if necessary (see Prieto-Recio, Martín-García, Bravo & Diez, 2015 for recommendations for *P. pinaster*). In particular, a more targeted and intense monitoring scheme, aiming (iii) to identify threats and forest decay, would be necessary in "uncertain" locations, as they are more susceptible of suffering the consequences of climate change. On the contrary, ex situ conservation is recommended in the case of exposed populations (Schueler et al., 2014) that are classified as "likely unsuitable" as we assume that the adaptive capacity of the populations do not allow them to adapt in situ. This strategy includes translocation (Leech, Almuedo & Neill, 2011) and/or conservation in germplasm banks. Given a choice, translocation is the preferred option as it maintains dynamic evolution within populations, and future suitability maps can identify suitable locations for such a purpose. Only when future suitability maps do not highlight available locations for translocation, germplasm banks are recommended.

Finally, for each genetic group, we calculated the percentage of the currently occupied territory proposed for ex situ conservation to assess the risk of habitat loss of the genetic group.

TABLE 1 Conservation strategy recommendations based on current distribution of the genetic group, current projection of its species distribution model (SDM) and future suitability map. Note that we considered as currently suitable those areas that obtained a suitability score > 70% of the maximum

Current distribution	SDM—Current projection	Future suitability map 2050	Conservation strategy
Present	—	Likely unsuitable	Ex situ
Present	—	Likely suitable	In situ with monitoring
Present	—	Uncertain	In situ with an intense and targeted monitoring scheme
Absent	Suitable	Likely suitable	Current first option area for translocation
Absent	Suitable	Uncertain	Current second option area for translocation
Absent	Unsuitable	Likely suitable	Midterm first option area for translocation
Absent	Unsuitable	Uncertain	Midterm second option area for translocation

2.7 | Exposure assessment of dynamic conservation units

We assessed the exposure of the currently defined dynamic conservation units covering each species' range: (i) the units from the EUFGIS database (http://portal.eufgis.org/) based on ecological zonation and expert knowledge, summing to 36 for *P. pinaster* and 19 for *P. halepensis* (see Tables S4 & S5 in Appendix S3); (ii) 10 additional units for *P. pinaster* based on molecular and quantitative trait information (Rodríguez-Quilón et al., 2016), composed from one to ten populations per unit (Table S6 in Appendix S3). Dynamic conservation units from EUFGIS without a clear association to a particular genetic group were excluded from the analysis.

We associated each unit to its corresponding genetic group and assessed the degree to which different gene pools (genetic groups) were represented in the European network. We also evaluated the risk of each dynamic conservation units to fail in finding suitable habitat in the future using future suitability maps. In the case of *P. halepensis*, when dynamic conservation units represented two genetic groups simultaneously (transition zones) risk evaluations of habitat (and genetic group) loss were performed separately for both genetic groups.

3 | RESULTS

3.1 | Species distribution models

Models performed similarly to those described by Serra-Varela et al. (2015) in cross-validation tests: (i) models built by means of individual algorithms performed well (TSS and AUC values above 0.80, and sensitivity and specificity above 90% in all cases except for GLM G4-hal; see Table S2 and S3 in Appendix S2); (ii) RF displayed the highest AUC and TSS scores in general.

All geographical projections can be checked in Appendix S4. We detected high agreement among the different algorithms used in the projected niches coinciding with the current distribution of the genetic groups. However, other locations were also highlighted as suitable obtaining relatively high scores (0.7–0.8), broadening the projected niches of the genetic groups (see Figures in Appendix S4).

3.2 | Future suitability maps

Likely suitable areas at genetic group level generally concentrated around the genetic group's current distribution (see Appendix S4). Away from the current distribution, uncertain habitat suitability areas were majorly found in other locations around the Mediterranean basin (e.g., G2-pin and G1-hal). Only in the case of G6/G7-hal, we found larger regions of likely suitable areas in northern central Europe. For both species, there were several other genetic groups that additionally found suitable areas in northern central Europe, although with a medium to low probability of suitability (which was classified as uncertain or likely unsuitable).

3.3 | Defining conservation strategies based on exposure to climate change

According to the results obtained, we proposed conservation guidelines for each genetic group based on its characteristics, as well as taking into account their current distribution as well as its future distribution (see Appendix S4).

The risk of habitat loss of the genetic groups varied widely from one genetic group to another within each of the two species. We detected very slightly exposed genetic groups, in which < 5% of the currently realized niche was classified as highly exposed (e.g., G8-pin and G3-hal; see Table S7 in Appendix S4), as well as cases in which this percentage (almost) exceeded 90% of the currently realized niche (e.g., G5/G6/G7-pin or G4-hal). This analysis revealed very diverging patterns of exposure to possible future habitat loss for both species along the Mediterranean coast of Spain: while *P. pinaster*'s genetic groups inhabiting this area (G5/G6/G7-pin) were highly exposed, *P. halepensis*' genetic groups occupying the same territory (G1/G4/G6/G7-hal) were not (percentage below 60%, except for G4-hal in southern Spain and Morocco, which was also found to be highly exposed).

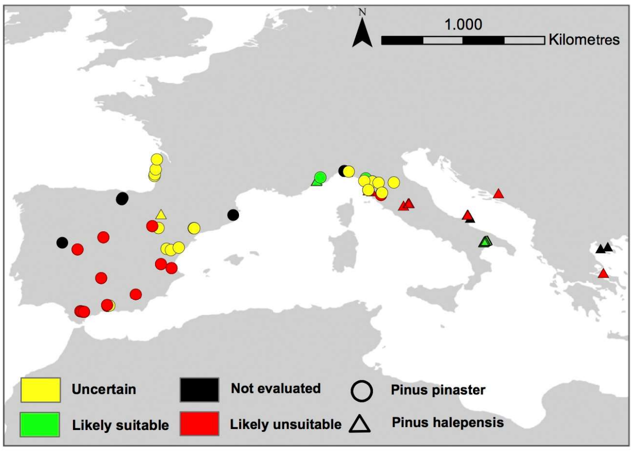


FIGURE 3 Exposure assessment for the EUFGIS dynamic conservation units of *Pinus pinaster* and *Pinus halepensis* along their natural distribution

Overall, we found that in situ conservation should suffice to guarantee conservation of the different genetic groups, as large areas were found in which in situ conservation will likely be successful when combined with a monitoring programme. Nevertheless, there were some cases in which planning a more targeted monitoring scheme seemed necessary to ensure that the areas detected as uncertainly exposed will be preserved (i.e., the cases of G3-pin, G4-pin and G5-hal).

Few genetic groups lacked areas that could be proposed for in situ conservation (i.e., G5/G6/G7-pin and G4-hal). In these cases it was necessary to consider ex situ conservation strategies either by translocation or by conservation in germplasm banks. For three of them (G5/G7-pin and G4-hal) translocation was the recommended ex situ conservation strategy, while for G6-pin that lacked translocation areas, germplasm bank conservation had to be recommended. In addition, because there were only few translocation areas available for G7-pin, the conservation of this genetic group would also benefit from germplasm conservation.

3.4 | Exposure assessment of dynamic conservation units

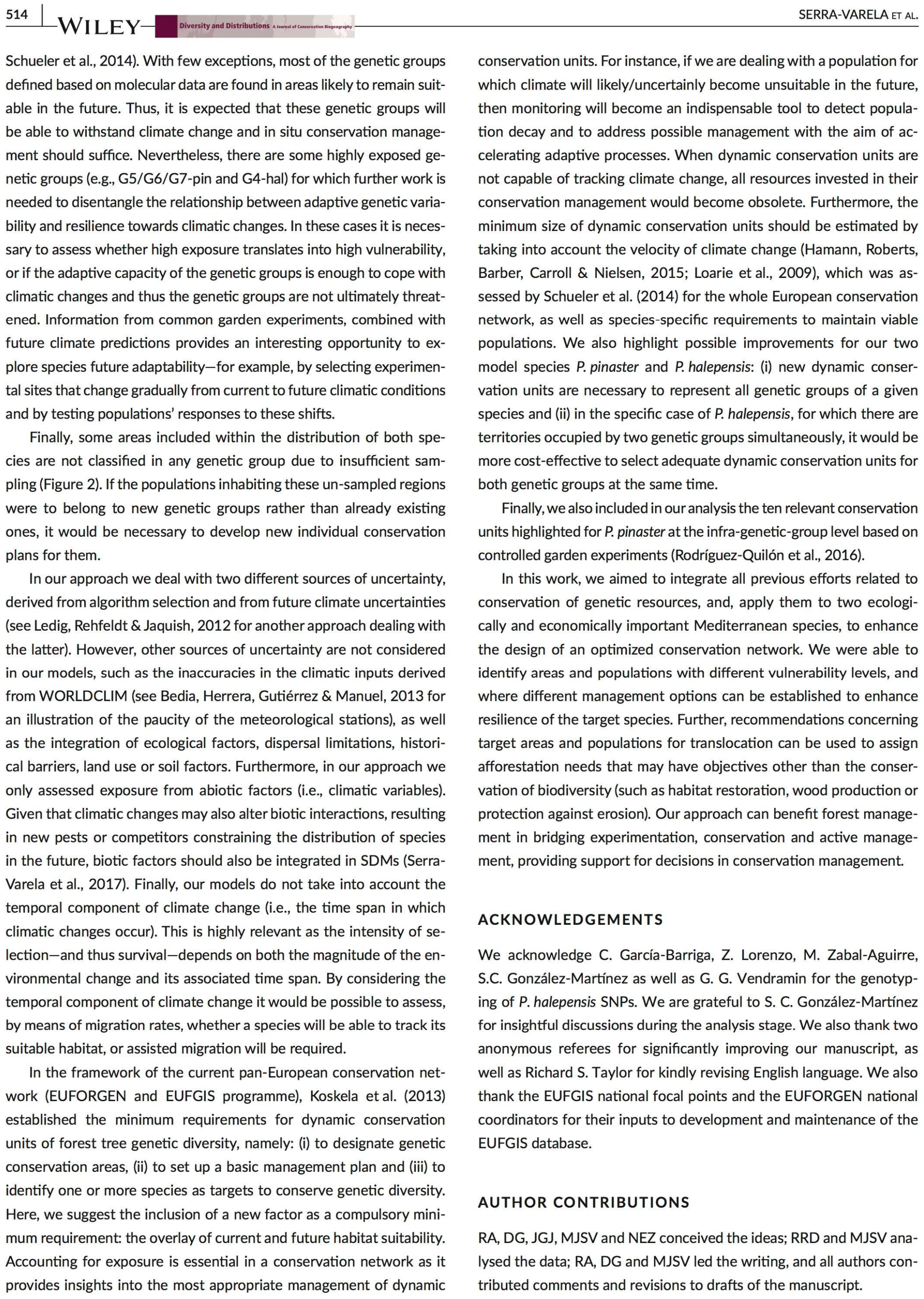
First, we analysed whether the already established EUFORGEN dynamic conservation units for *P. pinaster* and *P. halepensis* adequately represented the different gene pools detected in both species. In the case of *P. halepensis*, dynamic conservation units located within transition zones were considered representative of both genetic groups. We found that while some genetic groups were overrepresented (e.g., in *P. pinaster* 13 and 10 of 36 dynamic conservation units harbour G2/G7-pin populations, respectively, and in *P. halepensis* 14 of 19 dynamic conservation units combined admixtures of G3/G5-hal populations), other genetic groups were not included in the dynamic conservation units network (G1/G4/G8-pin and G1/G4-hal). The remaining genetic groups, although present within the dynamic conservation units network, were underrepresented (e.g., G6-pin and G6-hal with just 2 and 1 dynamic conservation units, respectively; Tables S4 & S5 in Appendix S3 for more details).

Second, we evaluated the risk of habitat loss within the different dynamic conservation units. In the case of *P. pinaster*, we detected two dynamic conservation units classified as "likely suitable" in 2050 both of which belonged to G2-pin. The other dynamic conservation units were classified as "likely unsuitable" (15) or "uncertain" (21) (Figure 3 and Appendix S4 for more details). As for *P. halepensis*, we detected three dynamic conservation units classified as "likely suitable" in 2050, which were located in Italy and represented G3-hal. The rest were classified as "likely unsuitable" except for eight dynamic conservation units corresponding to G2/G3/G7-hal that were classified as "uncertain." None of the dynamic conservation units located in transition zones, and thus representing two genetic groups at the same time, were classified as "likely suitable" for both represented groups (Figure 3 and Tables S4 & S5 in Appendix S3 for more details).

Finally, the exposure of the 10 genetically homogeneous conservation relevant units for *P. pinaster* (Rodríguez-Quilón et al., 2016), which span all genetic groups but G8-pin (Table S6 in Appendix S3), varied widely across populations within units composed of more than one population. For example, the ten populations comprising the conservation group SpAtl obtained three possible classifications (Table S6 in Appendix S3). Meanwhile among single population groups, the classification ranged from "likely unsuitable" to "uncertain" (Table S6 in Appendix S3).

4 | DISCUSSION

By individually analysing genetic groups and assessing their exposure to climate change, taking into account future climatic uncertainty, our approach constitutes a step forward in the preservation of the species' adaptive capacity, building on previous studies (Hamann, Aitken & Yanchuk, 2004; Kapeller, Lexer, Geburek, Hiebl & Schueler, 2012; and



Schueler et al., 2014). With few exceptions, most of the genetic groups defined based on molecular data are found in areas likely to remain suitable in the future. Thus, it is expected that these genetic groups will be able to withstand climate change and in situ conservation management should suffice. Nevertheless, there are some highly exposed genetic groups (e.g., G5/G6/G7-pin and G4-hal) for which further work is needed to disentangle the relationship between adaptive genetic variability and resilience towards climatic changes. In these cases it is necessary to assess whether high exposure translates into high vulnerability, or if the adaptive capacity of the genetic groups is enough to cope with climatic changes and thus the genetic groups are not ultimately threatened. Information from common garden experiments, combined with future climate predictions provides an interesting opportunity to explore species future adaptability—for example, by selecting experimental sites that change gradually from current to future climatic conditions and by testing populations' responses to these shifts.

Finally, some areas included within the distribution of both species are not classified in any genetic group due to insufficient sampling (Figure 2). If the populations inhabiting these un-sampled regions were to belong to new genetic groups rather than already existing ones, it would be necessary to develop new individual conservation plans for them.

In our approach we deal with two different sources of uncertainty, derived from algorithm selection and from future climate uncertainties (see Ledig, Rehfeldt & Jaquish, 2012 for another approach dealing with the latter). However, other sources of uncertainty are not considered in our models, such as the inaccuracies in the climatic inputs derived from WORLDCLIM (see Bedia, Herrera, Gutiérrez & Manuel, 2013 for an illustration of the paucity of the meteorological stations), as well as the integration of ecological factors, dispersal limitations, historical barriers, land use or soil factors. Furthermore, in our approach we only assessed exposure from abiotic factors (i.e., climatic variables). Given that climatic changes may also alter biotic interactions, resulting in new pests or competitors constraining the distribution of species in the future, biotic factors should also be integrated in SDMs (Serra-Varela et al., 2017). Finally, our models do not take into account the temporal component of climate change (i.e., the time span in which climatic changes occur). This is highly relevant as the intensity of selection—and thus survival—depends on both the magnitude of the environmental change and its associated time span. By considering the temporal component of climate change it would be possible to assess, by means of migration rates, whether a species will be able to track its suitable habitat, or assisted migration will be required.

In the framework of the current pan-European conservation network (EUFORGEN and EUFGIS programme), Koskela et al. (2013) established the minimum requirements for dynamic conservation units of forest tree genetic diversity, namely: (i) to designate genetic conservation areas, (ii) to set up a basic management plan and (iii) to identify one or more species as targets to conserve genetic diversity. Here, we suggest the inclusion of a new factor as a compulsory minimum requirement: the overlay of current and future habitat suitability. Accounting for exposure is essential in a conservation network as it provides insights into the most appropriate management of dynamic conservation units. For instance, if we are dealing with a population for which climate will likely/uncertainly become unsuitable in the future, then monitoring will become an indispensable tool to detect population decay and to address possible management with the aim of accelerating adaptive processes. When dynamic conservation units are not capable of tracking climate change, all resources invested in their conservation management would become obsolete. Furthermore, the minimum size of dynamic conservation units should be estimated by taking into account the velocity of climate change (Hamann, Roberts, Barber, Carroll & Nielsen, 2015; Loarie et al., 2009), which was assessed by Schueler et al. (2014) for the whole European conservation network, as well as species-specific requirements to maintain viable populations. We also highlight possible improvements for our two model species *P. pinaster* and *P. halepensis*: (i) new dynamic conservation units are necessary to represent all genetic groups of a given species and (ii) in the specific case of *P. halepensis*, for which there are territories occupied by two genetic groups simultaneously, it would be more cost-effective to select adequate dynamic conservation units for both genetic groups at the same time.

Finally, we also included in our analysis the ten relevant conservation units highlighted for *P. pinaster* at the infra-genetic-group level based on controlled garden experiments (Rodríguez-Quilón et al., 2016).

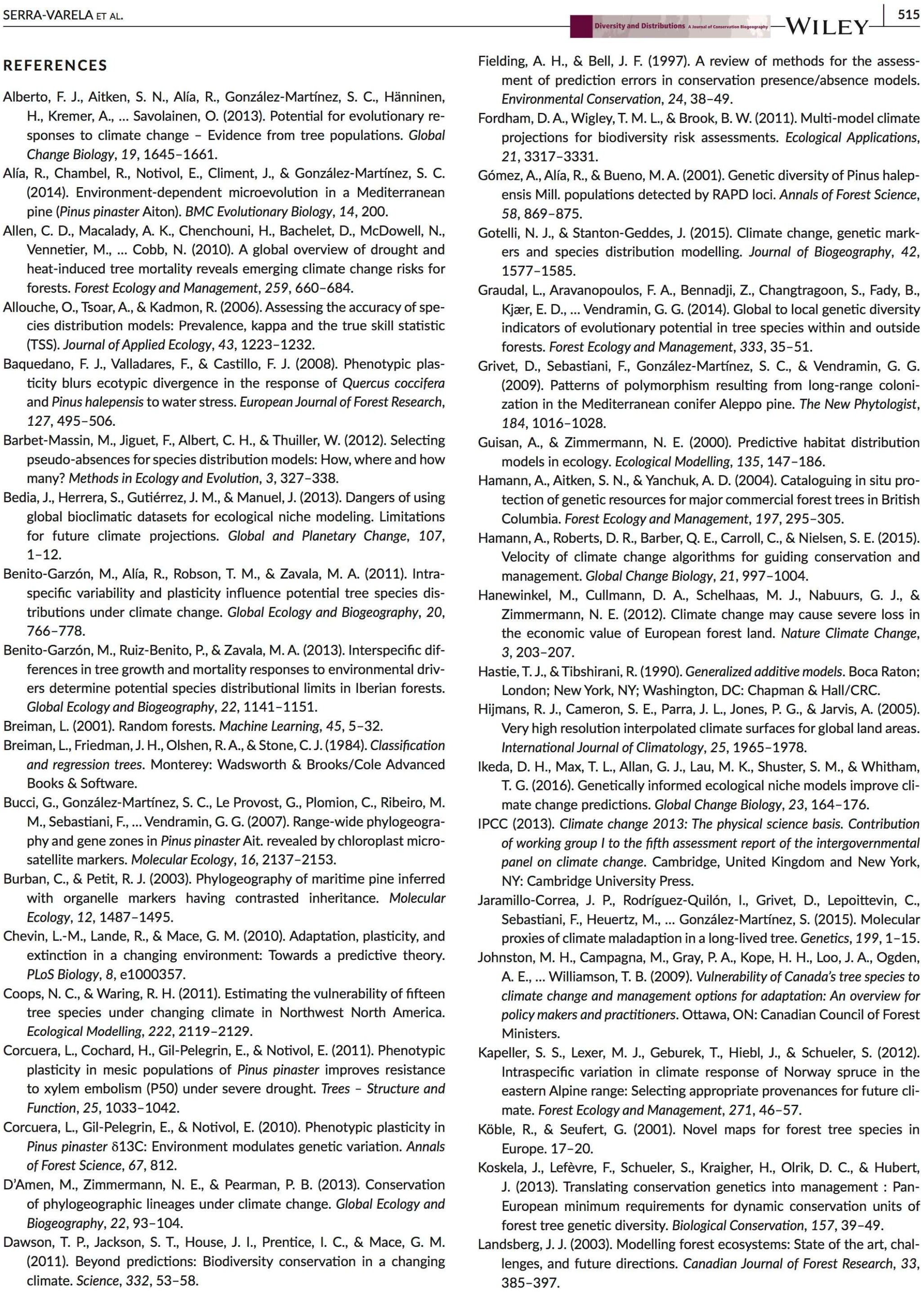
In this work, we aimed to integrate all previous efforts related to conservation of genetic resources, and, apply them to two ecologically and economically important Mediterranean species, to enhance the design of an optimized conservation network. We were able to identify areas and populations with different vulnerability levels, and where different management options can be established to enhance resilience of the target species. Further, recommendations concerning target areas and populations for translocation can be used to assign afforestation needs that may have objectives other than the conservation of biodiversity (such as habitat restoration, wood production or protection against erosion). Our approach can benefit forest management in bridging experimentation, conservation and active management, providing support for decisions in conservation management.

ACKNOWLEDGEMENTS

We acknowledge C. García-Barriga, Z. Lorenzo, M. Zabal-Aguirre, S.C. González-Martínez as well as G. G. Vendramin for the genotyping of *P. halepensis* SNPs. We are grateful to S. C. González-Martínez for insightful discussions during the analysis stage. We also thank two anonymous referees for significantly improving our manuscript, as well as Richard S. Taylor for kindly revising English language. We also thank the EUFGIS national focal points and the EUFORGEN national coordinators for their inputs to development and maintenance of the EUFGIS database.

AUTHOR CONTRIBUTIONS

RA, DG, JGJ, MJSV and NEZ conceived the ideas; RRD and MJSV analysed the data; RA, DG and MJSV led the writing, and all authors contributed comments and revisions to drafts of the manuscript.



REFERENCES

Alberto, F. J., Aitken, S. N., Alía, R., González-Martínez, S. C., Hänninen, H., Kremer, A., … Savolainen, O. (2013). Potential for evolutionary responses to climate change – Evidence from tree populations. *Global Change Biology*, *19*, 1645–1661.

Alía, R., Chambel, R., Notivol, E., Climent, J., & González-Martínez, S. C. (2014). Environment-dependent microevolution in a Mediterranean pine (*Pinus pinaster* Aiton). *BMC Evolutionary Biology*, *14*, 200.

Allen, C. D., Macalady, A. K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., … Cobb, N. (2010). A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *Forest Ecology and Management*, *259*, 660–684.

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*, 1223–1232.

Baquedano, F. J., Valladares, F., & Castillo, F. J. (2008). Phenotypic plasticity blurs ecotypic divergence in the response of *Quercus coccifera* and *Pinus halepensis* to water stress. *European Journal of Forest Research*, *127*, 495–506.

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*, 327–338.

Bedia, J., Herrera, S., Gutiérrez, J. M., & Manuel, J. (2013). Dangers of using global bioclimatic datasets for ecological niche modeling. Limitations for future climate projections. *Global and Planetary Change*, *107*, 1–12.

Benito-Garzón, M., Alía, R., Robson, T. M., & Zavala, M. A. (2011). Intraspecific variability and plasticity influence potential tree species distributions under climate change. *Global Ecology and Biogeography*, *20*, 766–778.

Benito-Garzón, M., Ruiz-Benito, P., & Zavala, M. A. (2013). Interspecific differences in tree growth and mortality responses to environmental drivers determine potential species distributional limits in Iberian forests. *Global Ecology and Biogeography*, *22*, 1141–1151.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey: Wadsworth & Brooks/Cole Advanced Books & Software.

Bucci, G., González-Martínez, S. C., Le Provost, G., Plomion, C., Ribeiro, M. M., Sebastiani, F., … Vendramin, G. G. (2007). Range-wide phylogeography and gene zones in *Pinus pinaster* Ait. revealed by chloroplast microsatellite markers. *Molecular Ecology*, *16*, 2137–2153.

Burban, C., & Petit, R. J. (2003). Phylogeography of maritime pine inferred with organelle markers having contrasted inheritance. *Molecular Ecology*, *12*, 1487–1495.

Chevin, L.-M., Lande, R., & Mace, G. M. (2010). Adaptation, plasticity, and extinction in a changing environment: Towards a predictive theory. *PLoS Biology*, *8*, e1000357.

Coops, N. C., & Waring, R. H. (2011). Estimating the vulnerability of fifteen tree species under changing climate in Northwest North America. *Ecological Modelling*, *222*, 2119–2129.

Corcuera, L., Cochard, H., Gil-Pelegrin, E., & Notivol, E. (2011). Phenotypic plasticity in mesic populations of *Pinus pinaster* improves resistance to xylem embolism (P50) under severe drought. *Trees – Structure and Function*, *25*, 1033–1042.

Corcuera, L., Gil-Pelegrin, E., & Notivol, E. (2010). Phenotypic plasticity in *Pinus pinaster* δ13C: Environment modulates genetic variation. *Annals of Forest Science*, *67*, 812.

D'Amen, M., Zimmermann, N. E., & Pearman, P. B. (2013). Conservation of phylogeographic lineages under climate change. *Global Ecology and Biogeography*, *22*, 93–104.

Dawson, T. P., Jackson, S. T., House, J. I., Prentice, I. C., & Mace, G. M. (2011). Beyond predictions: Biodiversity conservation in a changing climate. *Science*, *332*, 53–58.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*, 38–49.

Fordham, D. A., Wigley, T. M. L., & Brook, B. W. (2011). Multi-model climate projections for biodiversity risk assessments. *Ecological Applications*, *21*, 3317–3331.

Gómez, A., Alía, R., & Bueno, M. A. (2001). Genetic diversity of Pinus halepensis Mill. populations detected by RAPD loci. *Annals of Forest Science*, *58*, 869–875.

Gotelli, N. J., & Stanton-Geddes, J. (2015). Climate change, genetic markers and species distribution modelling. *Journal of Biogeography*, *42*, 1577–1585.

Graudal, L., Aravanopoulos, F. A., Bennadji, Z., Changtragoon, S., Fady, B., Kjær, E. D., … Vendramin, G. G. (2014). Global to local genetic diversity indicators of evolutionary potential in tree species within and outside forests. *Forest Ecology and Management*, *333*, 35–51.

Grivet, D., Sebastiani, F., González-Martínez, S. C., & Vendramin, G. G. (2009). Patterns of polymorphism resulting from long-range colonization in the Mediterranean conifer Aleppo pine. *The New Phytologist*, *184*, 1016–1028.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*, 147–186.

Hamann, A., Aitken, S. N., & Yanchuk, A. D. (2004). Cataloguing in situ protection of genetic resources for major commercial forest trees in British Columbia. *Forest Ecology and Management*, *197*, 295–305.

Hamann, A., Roberts, D. R., Barber, Q. E., Carroll, C., & Nielsen, S. E. (2015). Velocity of climate change algorithms for guiding conservation and management. *Global Change Biology*, *21*, 997–1004.

Hanewinkel, M., Cullmann, D. A., Schelhaas, M. J., Nabuurs, G. J., & Zimmermann, N. E. (2012). Climate change may cause severe loss in the economic value of European forest land. *Nature Climate Change*, *3*, 203–207.

Hastie, T. J., & Tibshirani, R. (1990). *Generalized additive models*. Boca Raton; London; New York, NY; Washington, DC: Chapman & Hall/CRC.

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*, 1965–1978.

Ikeda, D. H., Max, T. L., Allan, G. J., Lau, M. K., Shuster, S. M., & Whitham, T. G. (2016). Genetically informed ecological niche models improve climate change predictions. *Global Change Biology*, *23*, 164–176.

IPCC (2013). *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge, United Kingdom and New York, NY: Cambridge University Press.

Jaramillo-Correa, J. P., Rodríguez-Quilón, I., Grivet, D., Lepoittevin, C., Sebastiani, F., Heuertz, M., … González-Martínez, S. (2015). Molecular proxies of climate maladaption in a long-lived tree. *Genetics*, *199*, 1–15.

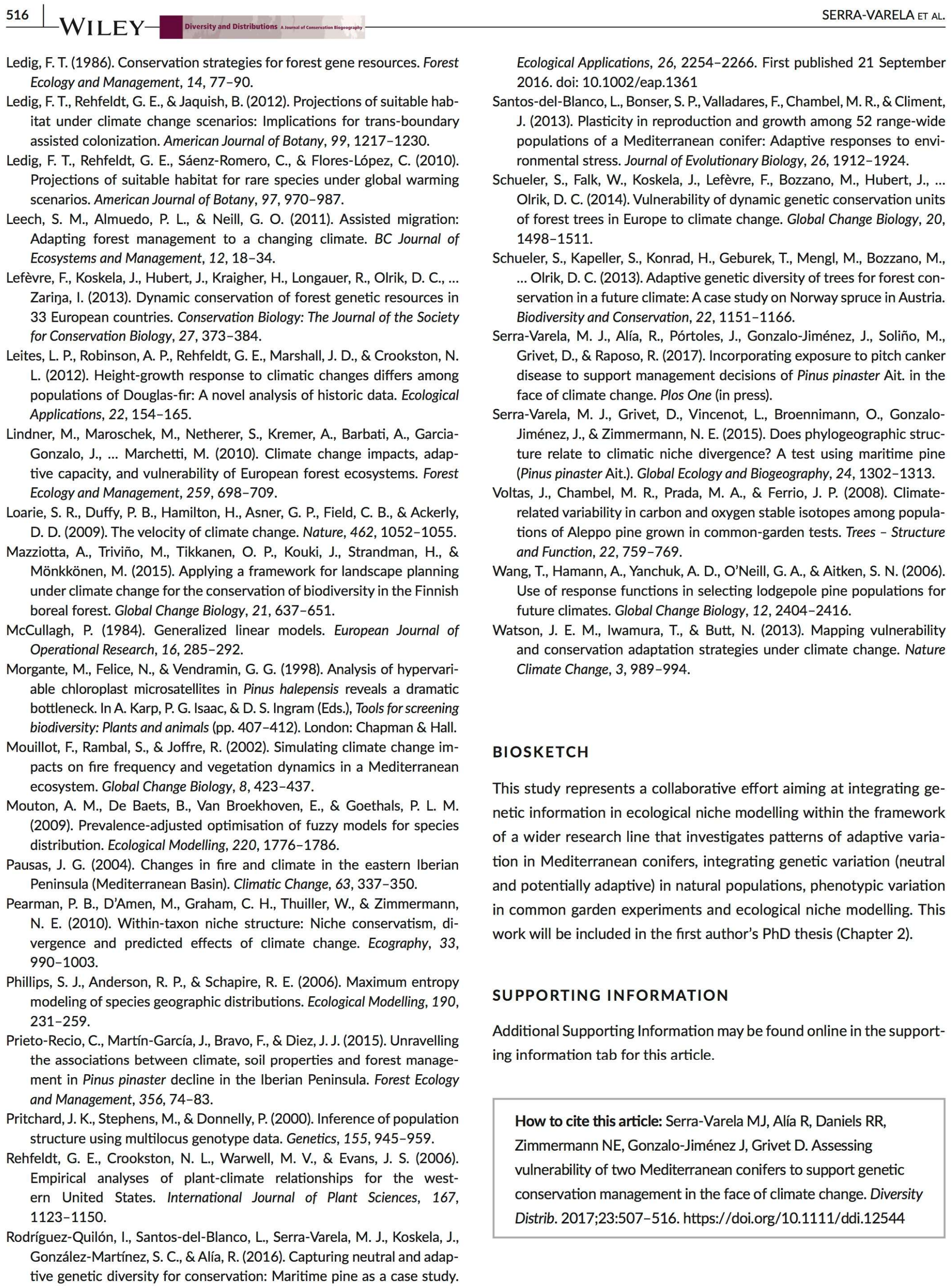
Johnston, M. H., Campagna, M., Gray, P. A., Kope, H. H., Loo, J. A., Ogden, A. E., … Williamson, T. B. (2009). *Vulnerability of Canada's tree species to climate change and management options for adaptation: An overview for policy makers and practitioners*. Ottawa, ON: Canadian Council of Forest Ministers.

Kapeller, S. S., Lexer, M. J., Geburek, T., Hiebl, J., & Schueler, S. (2012). Intraspecific variation in climate response of Norway spruce in the eastern Alpine range: Selecting appropriate provenances for future climate. *Forest Ecology and Management*, *271*, 46–57.

Köble, R., & Seufert, G. (2001). Novel maps for forest tree species in Europe. 17–20.

Koskela, J., Lefèvre, F., Schueler, S., Kraigher, H., Olrik, D. C., & Hubert, J. (2013). Translating conservation genetics into management : Pan-European minimum requirements for dynamic conservation units of forest tree genetic diversity. *Biological Conservation*, *157*, 39–49.

Landsberg, J. J. (2003). Modelling forest ecosystems: State of the art, challenges, and future directions. *Canadian Journal of Forest Research*, *33*, 385–397.



Ledig, F. T. (1986). Conservation strategies for forest gene resources. *Forest Ecology and Management*, *14*, 77–90.

Ledig, F. T., Rehfeldt, G. E., & Jaquish, B. (2012). Projections of suitable habitat under climate change scenarios: Implications for trans-boundary assisted colonization. *American Journal of Botany*, *99*, 1217–1230.

Ledig, F. T., Rehfeldt, G. E., Sáenz-Romero, C., & Flores-López, C. (2010). Projections of suitable habitat for rare species under global warming scenarios. *American Journal of Botany*, *97*, 970–987.

Leech, S. M., Almuedo, P. L., & Neill, G. O. (2011). Assisted migration: Adapting forest management to a changing climate. *BC Journal of Ecosystems and Management*, *12*, 18–34.

Lefèvre, F., Koskela, J., Hubert, J., Kraigher, H., Longauer, R., Olrik, D. C., … Zariņa, I. (2013). Dynamic conservation of forest genetic resources in 33 European countries. *Conservation Biology: The Journal of the Society for Conservation Biology*, *27*, 373–384.

Leites, L. P., Robinson, A. P., Rehfeldt, G. E., Marshall, J. D., & Crookston, N. L. (2012). Height-growth response to climatic changes differs among populations of Douglas-fir: A novel analysis of historic data. *Ecological Applications*, *22*, 154–165.

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*, 698–709.

Loarie, S. R., Duffy, P. B., Hamilton, H., Asner, G. P., Field, C. B., & Ackerly, D. D. (2009). The velocity of climate change. *Nature*, *462*, 1052–1055.

Mazziotta, A., Triviño, M., Tikkanen, O. P., Kouki, J., Strandman, H., & Mönkkönen, M. (2015). Applying a framework for landscape planning under climate change for the conservation of biodiversity in the Finnish boreal forest. *Global Change Biology*, *21*, 637–651.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, *16*, 285–292.

Morgante, M., Felice, N., & Vendramin, G. G. (1998). Analysis of hypervariable chloroplast microsatellites in *Pinus halepensis* reveals a dramatic bottleneck. In A. Karp, P. G. Isaac, & D. S. Ingram (Eds.), *Tools for screening biodiversity: Plants and animals* (pp. 407–412). London: Chapman & Hall.

Mouillot, F., Rambal, S., & Joffre, R. (2002). Simulating climate change impacts on fire frequency and vegetation dynamics in a Mediterranean ecosystem. *Global Change Biology*, *8*, 423–437.

Mouton, A. M., De Baets, B., Van Broekhoven, E., & Goethals, P. L. M. (2009). Prevalence-adjusted optimisation of fuzzy models for species distribution. *Ecological Modelling*, *220*, 1776–1786.

Pausas, J. G. (2004). Changes in fire and climate in the eastern Iberian Peninsula (Mediterranean Basin). *Climatic Change*, *63*, 337–350.

Pearman, P. B., D'Amen, M., Graham, C. H., Thuiller, W., & Zimmermann, N. E. (2010). Within-taxon niche structure: Niche conservatism, divergence and predicted effects of climate change. *Ecography*, *33*, 990–1003.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259.

Prieto-Recio, C., Martín-García, J., Bravo, F., & Diez, J. J. (2015). Unravelling the associations between climate, soil properties and forest management in *Pinus pinaster* decline in the Iberian Peninsula. *Forest Ecology and Management*, *356*, 74–83.

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

Rehfeldt, G. E., Crookston, N. L., Warwell, M. V., & Evans, J. S. (2006). Empirical analyses of plant-climate relationships for the western United States. *International Journal of Plant Sciences*, *167*, 1123–1150.

Rodríguez-Quilón, I., Santos-del-Blanco, L., Serra-Varela, M. J., Koskela, J., González-Martínez, S. C., & Alía, R. (2016). Capturing neutral and adaptive genetic diversity for conservation: Maritime pine as a case study. *Ecological Applications*, *26*, 2254–2266. First published 21 September 2016. doi: 10.1002/eap.1361

Santos-del-Blanco, L., Bonser, S. P., Valladares, F., Chambel, M. R., & Climent, J. (2013). Plasticity in reproduction and growth among 52 range-wide populations of a Mediterranean conifer: Adaptive responses to environmental stress. *Journal of Evolutionary Biology*, *26*, 1912–1924.

Schueler, S., Falk, W., Koskela, J., Lefèvre, F., Bozzano, M., Hubert, J., … Olrik, D. C. (2014). Vulnerability of dynamic genetic conservation units of forest trees in Europe to climate change. *Global Change Biology*, *20*, 1498–1511.

Schueler, S., Kapeller, S., Konrad, H., Geburek, T., Mengl, M., Bozzano, M., … Olrik, D. C. (2013). Adaptive genetic diversity of trees for forest conservation in a future climate: A case study on Norway spruce in Austria. *Biodiversity and Conservation*, *22*, 1151–1166.

Serra-Varela, M. J., Alía, R., Pórtoles, J., Gonzalo-Jiménez, J., Soliño, M., Grivet, D., & Raposo, R. (2017). Incorporating exposure to pitch canker disease to support management decisions of *Pinus pinaster* Ait. in the face of climate change. *Plos One* (in press).

Serra-Varela, M. J., Grivet, D., Vincenot, L., Broennimann, O., Gonzalo-Jiménez, J., & Zimmermann, N. E. (2015). Does phylogeographic structure relate to climatic niche divergence? A test using maritime pine (*Pinus pinaster* Ait.). *Global Ecology and Biogeography*, *24*, 1302–1313.

Voltas, J., Chambel, M. R., Prada, M. A., & Ferrio, J. P. (2008). Climate-related variability in carbon and oxygen stable isotopes among populations of Aleppo pine grown in common-garden tests. *Trees – Structure and Function*, *22*, 759–769.

Wang, T., Hamann, A., Yanchuk, A. D., O'Neill, G. A., & Aitken, S. N. (2006). Use of response functions in selecting lodgepole pine populations for future climates. *Global Change Biology*, *12*, 2404–2416.

Watson, J. E. M., Iwamura, T., & Butt, N. (2013). Mapping vulnerability and conservation adaptation strategies under climate change. *Nature Climate Change*, *3*, 989–994.

BIOSKETCH

This study represents a collaborative effort aiming at integrating genetic information in ecological niche modelling within the framework of a wider research line that investigates patterns of adaptive variation in Mediterranean conifers, integrating genetic variation (neutral and potentially adaptive) in natural populations, phenotypic variation in common garden experiments and ecological niche modelling. This work will be included in the first author's PhD thesis (Chapter 2).

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Serra-Varela MJ, Alía R, Daniels RR, Zimmermann NE, Gonzalo-Jiménez J, Grivet D. Assessing vulnerability of two Mediterranean conifers to support genetic conservation management in the face of climate change. *Diversity Distrib*. 2017;23:507–516. https://doi.org/10.1111/ddi.12544