

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL



Analysis of Population Structure and Local Adaptation in Holm Oak

Davide Miguel Lourenço Galhofa

Mestrado em Biologia Evolutiva e do Desenvolvimento

Dissertação orientada por:
Prof. Doutor Octávio S. Paulo

Resumo

As alterações climáticas são atualmente tidas como uma das maiores ameaças à biodiversidade a nível global. É esperado que estas alterações envolvam não só variações ao nível da temperatura e precipitação, mas também variações ao nível da sazonalidade. Espera-se também que haja um aumento da frequência de períodos de seca e um aumento da intensidade e frequência de eventos meteorológicos extremos. Para além disto, é expectável que a intensidade e o ritmo destas alterações tenda a aumentar. Deste modo, torna-se importante entender a maneira como as espécies respondem às alterações climáticas e a sua capacidade de resposta face à velocidade e intensidade a que se espera que estas alterações aconteçam nas próximas décadas. Este conhecimento é essencial, e pode ajudar a informar futuros esforços de conservação, de forma a amortecer as grandes perdas de biodiversidade que são esperadas.

De uma forma geral, as espécies podem adaptar-se às alterações climáticas de três formas: alterando a sua distribuição ao longo do espaço de forma a permanecer no seu nicho; acomodando-se às alterações no seu novo ambiente, através de plasticidade fenotípica ou mecanismos epigenéticos; ou através de adaptação genética com base na variabilidade genética presente nas populações. Dentro deste tema, o estudo de espécies de árvores é de particular interesse, dado serem organismos sésseis e com longos tempos geracionais, características que podem afetar a sua capacidade de resposta.

As azinheiras são árvores perenes do género *Quercus*, sendo uma das espécies dominantes nos bosques e matagais Mediterrânicos, onde são parte integrante do ecossistema, servindo como abrigo para um grande nível de biodiversidade. Na Península Ibérica, as Azinheiras são parte integral de ecossistemas de origem humana, conhecidos como montado em Portugal e dehesas em Espanha. Encontram-se distribuídas maioritariamente no centro-oeste da bacia do Mediterrânico, desde Portugal e Marrocos, ao extremo oeste da Turquia, e a norte na costa atlântica de França. Duas espécies de Azinheira são normalmente reconhecidas: *Quercus ilex* e *Quercus rotundifolia*. A espécie *Quercus ilex* habita zonas húmidas e sub-húmidas do este do Mediterrâneo à costa atlântica de França, enquanto que a espécie *Quercus rotundifolia*, habita zonas continentais e mais secas como a Península Ibérica e o norte de África. Apesar da distinção entre as duas espécies ser suportada por diferenças em caracteres morfológicos e bioquímicos das folhas, diferenças em termos de haplótipos cloroplastidiais e diferenças ecológicas no que toca à resposta ao stress hídrico, a sua classificação taxonómica é relativamente inconsistente, podendo aparecer classificadas como duas subespécies, ou até variedades de uma única espécie. Ambas possuem também várias adaptações que lhes conferem um alto grau de resistência à temperatura e à seca, e lhes permite lidar com as elevadas temperaturas e períodos de seca comuns no clima Mediterrânico.

Neste trabalho foram utilizadas amostras de Azinheira coletadas em 18 localizações da Bacia do Mediterrânico. Procedeu-se à extração e purificação do DNA de 94 amostras, que foram depois sequenciadas através de um protocolo de Genotyping by Sequencing (GBS). GBS é uma abordagem que envolve a sequenciação de frações específicas do genoma através da utilização de enzimas de restrição e Sequenciação de Nova Geração (NGS), permitindo identificar e genotipar um elevado número posições variáveis ao longo do genoma (SNPs - Single Nucleotide Polymorphisms). Foram gerados múltiplos

datasets utilizando diferentes programas de SNP calling, que foram utilizados para: (i) investigar a estrutura populacional de Azinheiras do Mediterrâneo através de vários métodos complementares e (ii) detetar sinais de adaptação local em *Q. rotundifolia* numa abordagem de Landscape Genomics, uma estratégia cujo objetivo é identificar fatores ambientais responsáveis por moldar a variação genética adaptativa responsável pela adaptação local. Isto foi feito através da deteção de associações entre SNPs e um conjunto de variáveis bioclimáticas relacionadas com temperatura e precipitação. Foi também feita a anotação das regiões genómicas sob possível efeito de seleção de forma a identificar potenciais genes relacionados com adaptação local ao stress-hídrico ou temperatura. Pelo que sabemos, esta é o primeiro estudo que usa marcadores nucleares ao nível do genoma para estudar estrutura populacional e detetar sinais de adaptação local em Azinheira.

Os nossos resultados revelam um padrão marcado de estrutura populacional e diferenciação, e destacam a separação das amostras em três populações genéticas: *Q. ilex*, *Q. rotundifolia* Ibérica e *Q. rotundifolia* Marroquina. O maior nível de diferenciação é observado entre amostras de *Q. ilex* e *Q. rotundifolia*, resultado que reforça a classificação destas taxa como duas espécies genética, morfológica e ecologicamente distintas. Dentro de *Q. rotundifolia*, foram observados valores consideráveis de diferenciação entre amostras Ibéricas e Marroquinas, evidenciando a existência de duas populações distintas. Este nível de diferenciação parece evidenciar o já proposto isolamento destas duas populações em refúgios distintos durante as glaciações do Quaternário, e a colonização pós-glacial do Norte de África e da Península Ibérica a partir destes refúgios. Este resultado reforça também a possível classificação de uma das populações como uma subespécie distinta, algo anteriormente proposto com base em caracteres fenotípicos. A população Ibérica de *Q. rotundifolia* foi identificada como sendo relativamente pouco estruturada, o que pode possivelmente ser atribuído à colonização pós-glacial a partir de um ou mais refúgios no sul da península, bem como um equilíbrio entre fluxo génico e adaptação.

Os resultados de estrutura populacional revelam também o agrupamento de amostras Tunisinas com as restantes amostras de *Q. ilex*, algo que parece discordar da literatura existente relativa à distribuição, morfologia e genética das Azinheiras nesta zona da distribuição, que até agora se viram classificadas como *Q. rotundifolia*. Em particular, este resultado evidencia uma incongruência entre a distribuição geográfica de haplótipos cloroplastidiais conhecida e os resultados de estrutura obtidos com base em marcadores nucleares. Uma possível explicação para este padrão será a existência de hibridação com base num fluxo de pólen de longa distância com origem algures na distribuição de *Q. ilex*, potencialmente Sicília ou Sardenha.

Em relação à deteção de sinais de adaptação em *Q. rotundifolia*, foi possível identificar um número considerável de SNPs sob o possível efeito de seleção natural. A maioria dos SNPs identificados mostrou evidências de estar associada a várias variáveis bioclimáticas relacionadas com temperatura e precipitação. Observou-se também uma diferença no tipo de variáveis identificadas como associadas quando são analisadas amostras Ibéricas e Marroquinas ou apenas amostras Ibéricas. Em particular, observou-se uma tendência a que variáveis relacionadas com a precipitação fossem maioritariamente identificadas como associadas quando incluídas amostras Marroquinas. A alteração das frequências alélicas dos SNPs associados ao longo das respetivas variáveis ambientais parece seguir um padrão específico, de acordo com os nossos resultados: segundo os nossos resultados, existe uma tendência para que um dos alelos esteja fixo acima ou abaixo de um valor ou intervalo de valores, que tende a ser consistente para cada variável ambiental. Finalmente, a anotação das regiões genómicas em que estão inseridos os SNPs sob o possível efeito da seleção revelou vários genes putativamente relacionados com a adaptação local à temperatura e ao stress hídrico.

De uma forma geral, estes resultados contribuem não só para uma clarificação da organização

taxonómica das Azinheiras, que até ao momento tem sido relativamente inconsistente, e também para o esclarecimento da estrutura populacional das Azinheiras na Bacia do Mediterrâneo. Adicionalmente, configuram um primeiro passo no estudo da adaptação local às alterações climáticas nestes taxa, abrindo caminho para a realização de análises mais complexas, que nos podem ajudar a entender a capacidade das populações das Azinheiras de responder a alterações nos seus ambientes, e informar possíveis esforços de conservação.

Palavras-chave: Azinheira; Mediterrâneo; Genómica Populacional; Associação ambiental; Adaptação local

Abstract

Climate change is currently one of the biggest threats to biodiversity. Understanding how species respond to climate change, and their capacity to respond to the ever-increasing pace and intensity at which these changes are taking place, is critical for the preservation of ecosystems. Tree species make a case study of particular interest, as they are sedentary, display long generation times, and have high ecological importance. The Mediterranean region has been described as a ‘climate change hotspot’, with increased temperatures and decreased precipitation expected to affect the region in the coming decades. Holm oaks (*Quercus* genus) are considered keystone tree species of the Mediterranean basin and display several adaptations that award them a high degree of temperature and drought resistance, which helps them deal with the hot, dry summers common to the Mediterranean climate.

This work involved samples from two Holm oak species, *Quercus ilex* and *Quercus rotundifolia*, collected throughout the Mediterranean basin. Genotyping by Sequencing was used to derive several nuclear Single Nucleotide Polymorphism datasets, which were used (i) to investigate population structure using several complementary approaches, and (ii) to detect evidence of local adaptation through a Landscape Genomics approach involving the detection of *Fst* outliers and genetic-environmental associations. This was done using samples of Holm oak collected from 18 sampling locations across the western Mediterranean basin. This is, to our knowledge, the first attempt to use genome-wide, nuclear genetic markers to uncover the existence of population structure and signatures of local adaptation in Holm oak.

Our results reveal a marked pattern of population structure and differentiation, which supports the status of *Q. ilex* and *Q. rotundifolia* as two genetically, morphologically, and ecologically distinct species. A considerable degree of differentiation is also observed between Iberian and Moroccan samples of *Q. rotundifolia*, indicating their status as distinct populations. On the other hand, Iberian samples and Moroccan samples geographically closer to Iberia show lower levels of differentiation, indicating either the presence of gene flow or incomplete lineage sorting. We also uncover the relatively unstructured nature of the Iberian *Q. rotundifolia* population. Additionally, we uncover the presence of *Q. ilex* in Tunisia, where only *Q. rotundifolia* was thought to inhabit, revealing an incongruity between the known chlorotype distribution and population structure inferred from nuclear DNA, which we hypothesize to be the result of hybridization through long-distance pollen flow. Regarding the detection of local adaptation, we identified a considerable number of putative SNPs under selection in *Q. rotundifolia*. The majority of these SNPs showed evidence of association with multiple temperature and precipitation-related bioclimatic variables. Annotation of the genomic areas in which the putative SNPs under selection were located, through comparison with the reference genome of *Quercus lobata*, revealed several genes putatively related to the adaptive process to heat and water stress.

In general, these results contribute towards the clarification of the taxonomy of Holm oak, which up until this point has suffered from a lack of consistency, bring new insights into the population structure of Holm oak, and also represent a first step in understanding the capacity of Mediterranean Holm oak

populations to respond to future climate change.

Keywords: Holm oak; Mediterranean; Population genomics; Landscape genomics; Local adaptation

Aos meus pais, cujo trabalho e esforço me permitiram chegar até aqui. Aos membros do CoBiG2, Professor Octávio Paulo, André Henriques, Filipe Sousa, Francisco Pina-Martins, Mariana Viegas, Paulo Sousa e Waira Machida, por todo o apoio durante estes dois anos. Aos meus amigos do mestrado em Biologia Evolutiva e do Desenvolvimento, Afonso Magalhães, Ana Carolina Pereira, Marta Ferreira e Vanessa Ribeiro, por toda a companhia durante estes três anos.

Keep Pace: Selection of trees keeping pace with fast environmental changes, a science-based approach for sustainable XXI century Oak forests – KeePace. EU ERDF funds. Portugal 2020 Partnership Agreement and Programa Operacional Regional Algarve. National funds through Fundação para a Ciência e Tecnologia (FCT) Portugal: ALG-01- 0145-FEDER-029263 / PTDC/ASP/SIL/29263/2017.

Table of contents

Resumo	I
Abstract	V
Table of contents	IX
List of Figures	XI
List of Tables	XV
1 Introduction	1
1.1 Climate change as a threat to biodiversity	2
1.2 The Mediterranean Basin and its climate	2
1.3 How species may adapt to climate change	3
1.4 General characteristics of Holm oaks	5
1.4.1 Taxonomic classification of Holm oaks	5
1.5 Obtaining genome-wide SNP data using Genotyping by Sequencing	8
1.6 Complementary approaches for the inference of population structure	9
1.7 Understanding the adaptive process through Landscape Genomics	10
1.8 Objectives of this thesis	11
2 Materials and Methods	13
2.1 Study species and sampling sites	14
2.2 DNA extraction, quantification and quality assessment	14
2.3 Sequencing, raw data filtering, and SNP calling	14
2.3.1 Demultiplexing and removal of mitochondrial and chloroplastidial reads	15
2.3.2 SNP calling and filtering	15
2.4 GBS data analysis	20
2.4.1 Population Structure	20
2.4.2 Summary Statistics	20
2.4.3 Outlier detection and genetic-environmental association	20
3 Results	23
3.1 Filtering and SNP calling	24
3.2 Population Structure	25
3.3 Genetic diversity and differentiation	29
3.4 Outlier detection	31

3.5	Genetic-environmental association	31
3.5.1	Associated SNPs and bioclimatic variables	31
3.5.2	Allele frequency analysis	35
3.6	Loci annotation and enrichment analyses	39
4	Discussion	41
4.1	Observed differences between SNP calling programs	42
4.2	Genetic diversity and population structure of Holm oak	43
4.3	Detection of selection in Holm oak	45
4.4	Conclusion and final remarks	49
	Bibliography	51
	Supplementary Material	61

List of Figures

2.1	Sampling locations and species distribution range of <i>Q. rotundifolia</i> and <i>Q. ilex</i> (as presented in Cuadullo et al. (2017) [22]). Sampling location codes as in Table 2.1.	14
2.2	Simplified workflow of the filtering and SNP calling process using <i>ipyrad</i> and <i>Stacks</i> . Represented are the main steps in the filtering and SNP calling process for different assemblies, including the intermediate datasets. Only the eight final datasets are shown (Table 2.2), as well as the analyses for which they were used. The remaining assemblies run with different parameter combinations are not shown (Supplementary Tables S2 and S3).	18
3.1	Population structure results for <i>ipyrad</i> datasets, highlighting the three genetic groups found. (A) <i>fineRADstructure</i> results for dataset 'ipr_allsnps_allsamps'. Genetic groups are identified in the following way: I - Iberian <i>Q. rotundifolia</i> ; II - Moroccan <i>Q. rotundifolia</i> ; and III - <i>Q. ilex</i> . (B) <i>ALStructure</i> results for dataset 'ipr_centersnp_allsamps', from K2 to K5.	26
3.2	Population structure results for <i>Stacks</i> datasets, highlighting the three genetic groups found. (A) <i>fineRADstructure</i> results for dataset 'stk_allsnps_allsamps'. Genetic groups are identified in the following way: I - Iberian <i>Q. rotundifolia</i> ; II - Moroccan <i>Q. rotundifolia</i> ; and III - <i>Q. ilex</i> . (B) <i>ALStructure</i> results for dataset 'stk_centersnp_allsamps', from K2 to K5.	27
3.3	PCA results for datasets (a) 'ipr_centersnp_allsamps' and (b) 'stk_centersnp_allsamps'. Points represent the samples. Each sampling location is represented by a different point color/shape. Samples belonging to each of the genetic groups revealed in the <i>fineRADstructure</i> and <i>ALStructure</i> results are highlighted.	28
3.4	Frequency of the major allele for the outlier SNPs of datasets (a) 'ipr_centersnp_ib_mor' and (b) 'stk_centersnp_ib_mor', calculated for each <i>Q. rotundifolia</i> sampling location.	32
3.5	Representation of allele frequency turnover for loci associated with bioclimatic variable bio4 (temperature seasonality). The three strongest associated loci are shown for datasets ipr_centersnp_ib and stk_centersnp_ib. Allele frequencies for each sampling location are plotted as a function of temperature seasonality values (above). The scale of the yy axis may differ across plots. Allele frequencies for each sampling location are represented along the geographical area of the Iberian Peninsula (below). The range of temperature seasonality values is represented by a color scale. Values for each sampling location are represented above their geographical position.	36

3.6	Representation of allele frequency turnover for loci associated with variable bio12 (annual precipitation amount). The three strongest associated loci are shown for datasets <i>ipr_centersnp_ib_mor</i> and <i>stk_centersnp_ib_mor</i> . Allele frequencies for each sampling location are plotted as a function of annual precipitation values (above). The scale of the yy axis may differ across plots. Allele frequencies for each sampling location are represented along the geographical area of the Iberian Peninsula and northern Morocco (below). The range of annual precipitation is represented by a color scale. Values for each sampling location are represented above their geographical position.	37
3.7	Average pattern of allele turnover along the associated bioclimatic variables common (a) to datasets <i>ipr_centersnp_ib</i> and <i>stk_centersnp_ib</i> , and (b) to datasets <i>ipr_centersnp_ib_mor</i> and <i>stk_centersnp_ib_mor</i> . Each SNP associated with an environmental variable, for all levels of association strength, was classified according to its pattern of allele turnover (upward pattern, with a tendency of allele frequencies to be fixed above turnover threshold; downward pattern, with a tendency of allele frequencies to be fixed below turnover threshold). The average allele frequency of the major alleles was calculated for SNPs in each turnover pattern, for each sampling location/environmental value. Colored lines represent the average allele frequencies along the bioclimatic variable. Standard deviation is represented by the colored shadow. .	38
S1	Distribution of the number of raw reads per sample after the GBS sequencing protocol. The total number of reads was 293,145,449. The mean number of reads per sample was 3,118,569.	61
S2	Distribution of the number of paired-end reads per sample for (a) the raw dataset after demultiplexation and after the removal of mitochondrial and chloroplastidial reads with <i>BBMap</i> , and (b) after <i>ipyrad</i> and <i>Stacks</i> filtering steps.	64
S3	Spearman's correlation matrices for the 19 bioclimatic variables, calculated for the geographic areas of (a) Iberia and (b) Iberia and Morocco. Color corresponds to the strength of the correlation.	67
S4	Environmental values for all 19 bioclimatic variables. The average environmental value and standard deviation are presented for each <i>Q. rotundifolia</i> sampling location. Sampling locations are colored according to their genetic population (QRO_IBERIA - Iberian <i>Q. rotundifolia</i> ; QRO_MOROCCO - Moroccan <i>Q. rotundifolia</i>). The bioclimatic variables adopted for the genetic-environmental analyses were bio1, bio2, bio3, bio4, bio6, bio7, bio8, bio9, bio10, bio12, bio15, and bio17. Bioclimatic variable names can be consulted in Supplementary Table S4.	68
S5	Full <i>ALStructure</i> results for K2 - K6 for datasets (a) <i>ipr_centersnp_ib</i> and (b) <i>stk_centersnp_ib</i> . Samples are organized by sampling location.	69
S6	Bootstrap distributions of F_{st}/F_{st}^{max} values for each population, for datasets (a) <i>ipr_centersnp_ib</i> and (b) <i>stk_centersnp_ib</i>	70
S7	Frequency of the major allele for the outlier SNPs detected by <i>BayPass</i> core model mode, for datasets (a) <i>ipr_centersnp_ib</i> and (b) <i>stk_centersnp_ib</i>	72
S8	Proportion of associated SNPs (a) per bioclimatic variable, per dataset, and (b) per bioclimatic variable, per dataset, per strength of association, for datasets <i>ipr_centersnp_ib</i> , <i>ipr_centersnp_ib_mor</i> , <i>stk_centersnp_ib</i> , and <i>stk_centersnp_ib_mor</i>	73
S8	<i>cont.</i>	74

- S9 Average pattern of allele turnover along all associated bioclimatic variables, for the datasets `ipr.centersnp_ib`, `stk.centersnp_ib`, `ipr.centersnp_ib_mor`, and `stk.centersnp_ib_mor`. Each SNP associated with an environmental variable, for all levels of association strength, was classified according to its pattern of allele turnover (upward pattern, a tendency of allele frequencies to be fixed above turnover threshold; downward pattern, a tendency of allele frequencies to be fixed below turnover threshold). The average allele frequency of the major alleles was calculated for SNPs in each turnover pattern, for each sampling location/environmental value. The number of associated SNPs is represented for each dataset/bioclimatic variable. Colored lines represent the average allele frequencies along the bioclimatic variable. Standard deviation is represented by the colored shadow. 75
- S10 Average pattern of allele turnover along associated bioclimatic variables, for the datasets `ipr.centersnp_ib`, `stk.centersnp_ib`, `ipr.centersnp_ib_mor`, and `stk.centersnp_ib_mor`. Each SNP associated with an environmental variable, for the decisive level of association strength, was classified according to its pattern of allele turnover (upward pattern, a tendency of allele frequencies to be fixed above turnover threshold; downward pattern, a tendency of allele frequencies to be fixed below turnover threshold). The average allele frequency of the major alleles was calculated for SNPs in each turnover pattern, for each sampling location/environmental value. The number of associated SNPs is represented for each dataset/bioclimatic variable. Colored lines represent the average allele frequencies along the bioclimatic variable. Standard deviation is represented by the colored shadow. 76
- S11 Representation of allele frequency turnover for loci associated with bioclimatic variable `bio9` (mean daily mean air temperatures of the driest quarter). The three strongest associated loci are shown for datasets `ipr.centersnp_ib` and `stk.centersnp_ib`. Allele frequencies for each sampling location are plotted as a function of mean daily mean air temperature values (above). The scale of the yy axis may differ across plots. Allele frequencies for each sampling location are represented along the geographical area of the Iberian Peninsula (below). The range of mean daily mean air temperature is represented by a color scale. Values for each sampling location are represented above their geographical position. 77
- S12 Representation of allele frequency turnover for loci associated with bioclimatic variable `bio15` (precipitation seasonality). The three strongest associated loci are shown for datasets `ipr.centersnp_ib` and `stk.centersnp_ib`. Allele frequencies for each sampling location are plotted as a function of precipitation seasonality values (above). The scale of the yy axis may differ across plots. Allele frequencies for each sampling location are represented along the geographical area of the Iberian Peninsula (below). The range of precipitation seasonality is represented by a color scale. Values for each sampling location are represented above their geographical position. 78

- S13 Representation of allele frequency turnover for loci associated with bioclimatic variable bio17 (mean monthly precipitation amount of the driest quarter). The three strongest associated loci are shown for datasets `ipr_centersnp_ib` and `stk_centersnp_ib`. Allele frequencies for each sampling location are plotted as a function of mean monthly precipitation values (above). The scale of the yy axis may differ across plots. Allele frequencies for each sampling location are represented along the geographical area of the Iberian Peninsula (below). The range of mean monthly precipitation is represented by a color scale. Values for each sampling location are represented above their geographical position. 79
- S14 Representation of allele frequency turnover for SNPs/loci associated with bioclimatic variable bio17 (mean monthly precipitation amount of the driest quarter). The three strongest associated SNPs/loci are shown for datasets `ipr_centersnp_ib_mor` and `stk_centersnp_ib_mor`. Allele frequencies for each sampling location are plotted as a function of mean monthly precipitation values (above). The scale of the yy axis may differ across plots. Allele frequencies for each sampling location are represented along the geographical area of the Iberian Peninsula and northern Morocco (below). The range of mean monthly precipitation is represented by a color scale. Values for each sampling location are represented above their geographical position. 80
- S15 Simplified workflow of the outlier detection and genetic-environmental association analyses using *BayPass*, BLAST against the reference genome of *Q. lobata* and enrichment analysis using *g:Profiler*. This process is shown for datasets '`ipr_centersnp_ib`', '`stk_centersnp_ib`', '`ipr_centersnp_ib_mor`', and '`stk_centersnp_ib_mor`'. Represented are the number of outlier or associated SNPs/loci for each dataset after the respective analysis using *BayPass*, the number of unique *Q. lobata* genes found to be similar to the datasets' loci, and the number of enriched GO terms for the datasets that returned results. Also represented are the sets of genes that were explored in terms of biological function. 81
- S16 Venn diagrams representing the overlap of genes found for the different datasets (`ipr_centersnp_ib`, `ipr_centersnp_ib_mor`, `stk_centersnp_ib`, and `stk_centersnp_ib_mor`) after annotation of outlier/ associated loci against the reference genome *Q. lobata*. Represented is the overlap for (a) *ipyrad* datasets after outlier detection and genetic-environmental association, (b) *Stacks* datasets after outlier detection and genetic-environmental association, and (c) the four datasets after genetic-environmental association analyses. 82
- S17 Enrichment analyses results for datasets `ipr_centersnp_ib`, `ipr_centersnp_ib_mor`, `stk_centersnp_ib`, and `stk_centersnp_ib_mor`. Analyses were run with *g:Profiler* for the list of *Q. lobata* genes similar to the loci obtained from the outlier detection and genetic-environmental analyses. Gene ratio represents the number of observed genes in our datasets divided by the number of genes in the reference genome annotated for each GO term). The gene count is the number of genes related to each GO term and is represented by the size of points. The adjusted p-values are a measure of the randomness of the intersection between the query and the ontology term and are represented by a color gradient. Results are grouped by the three GO term categories: *GO:BP* = Biological process; *GO:CC* = Cellular component; *GO:MF* = Molecular function. 83

List of Tables

2.1	Sampling information. Additional information for each sample, including geographical coordinates, can be found in Supplementary Table S1	15
2.2	Summary table of all datasets used in the data analyses, showing the dataset names, included samples, SNPs used in final datasets, SNP calling software parameters, and <i>VCFTools</i> filtering parameters for (a) <i>ipyrad</i> and (b) <i>Stacks</i> . Datasets iii, iv, vii, and viii contain only a subset of the total samples, corresponding to either Iberian <i>Q. rotundifolia</i> or Iberian and Moroccan <i>Q. rotundifolia</i> samples.	19
3.1	Summary of final datasets used for (a) the population structure and genetic diversity analyses and (b) outlier detection and genetic-environmental association. Presented is the number of SNPs after <i>ipyrad/Stacks</i> SNP calling process (and <i>vcf_parser.py</i> for Center-SNP datasets), and the number of SNPs and samples in the final datasets, after <i>VCFTools</i> filtering.	25
3.2	Summary statistics, for each population and overall, and global F_{st} values calculated for datasets (a) 'ipr_centersnp_allsamp' and (b) 'stk_centersnp_allsamp'. Represented is the number of individuals (N), number of alleles in each population (N_A), allelic richness A_R , observed heterozygosity (H_o), expected heterozygosity (H_e), inbreeding coefficient (F_{is}), F_{st}/F_{st}^{max} for each population, and overall F_{st}	30
3.3	Pairwise F_{st} values calculated for datasets (a) 'ipr_centersnp_allsamp' and (b) 'stk_centersnp_allsamp'. F_{st} values were calculated pairwise considering two populations (<i>Q. rotundifolia</i> and <i>Q. ilex</i>) and three populations (Iberian <i>Q. rotundifolia</i> (<i>Q. rotund. Ib.</i>), Moroccan <i>Q. rotundifolia</i> (<i>Q. rotund. Mor.</i>), and <i>Q. ilex</i>).	30
3.4	Summary of results for (a) the outlier detection and (b) genetic-environmental association analyses, and subsequent annotation process, for datasets 'ipr_centersnp_ib', 'ipr_centersnp_ib_mor', 'stk_centersnp_ib' and 'stk_centersnp_ib_mor'. Columns show the number of outlier and associated SNPs/loci in each dataset, the total number of sequence matches against the reference genome of <i>Q. lobata</i> , the number of unique loci in the dataset with at least a match, and the number of unique genes matched in <i>Q. lobata</i> 's reference genome.	33

3.5	Genetic-environmental association results for datasets (a) 'ipr_centersnp_ib', (b) 'ipr_centersnp_ib_mor', (c) 'stk_centersnp_ib' and (d) 'stk_centersnp_ib_mor', analyzed under <i>BayPass</i> v.2.31's auxiliary covariate mode. Columns show the number of associations, the number of associated SNPs, and their percentage in relation to the total number of SNPs in the dataset (Table 3.1), and the associated variables for the different levels of association strength and in total. Association strength is defined by the BF value, based on Jeffreys' rule according to Gautier et al. (2015) [76].	34
S1	Identification, country of origin, sampling location name, and geographical coordinates for the 94 samples used in this study. Bold Sample IDs indicate samples removed due to missing data.	62
S1	(cont.)	63
S2	Results for <i>ipyrad</i> assemblies run with different values for the <i>clustering_threshold</i> parameter. Represented for all-SNPs and center-SNPs datasets are the number of variable sites directly out of <i>ipyrad</i> (and <i>vcf_parser.py</i> for center-SNP datasets) and after <i>VCFTools</i> filtering, and the number of samples remaining in the datasets after removing those with >65% of missing data.	65
S3	Results for <i>Stacks</i> assemblies run with different values for parameters <i>-m</i> , <i>-M</i> , <i>-n</i> , and <i>-R</i> . Represented for all-SNPs and center-SNPs datasets are the number of variable sites directly out of <i>ipyrad</i> (and <i>vcf_parser.py</i> for center-SNP datasets) and after <i>VCFTools</i> filtering, and the number of samples remaining in the datasets after removing those with >65% of missing data.	65
S4	List of bioclimatic variables kept for each geographical area after the correlation analyses. Represented for each variable are the short name, full name, unit, and a brief explanation.	66
S5	Pairwise <i>F_{st}</i> values calculated for datasets 'ipr_centersnp_allsamps' (below diagonal) and 'stk_centersnp_allsamps' (above diagonal). <i>F_{st}</i> values were calculated pairwise for all sampling locations. Sampling location codes as in Table 2.1.	71
S6	Summary of outlier SNP/loci hits against the reference genome of <i>Q. lobata</i> for datasets 'ipr_centersnp_ib_mor' and 'stk_centersnp_ib_mor'.	77
S7	Summary of <i>Q. lobata</i> genes with a potentially interesting biological function relating to heat and water stress response, among the 50 common genes between datasets (a) 'stk_centersnp_ib' and (b) 'stk_centersnp_ib_mor'. For each of the genes, the respective associated SNPs and bioclimatic variables are shown, as well as their putative role in the response to abiotic stressors such as temperature or drought.	84

Chapter 1

Introduction

1.1 Climate change as a threat to biodiversity

Climate change is currently one of the biggest drivers of biodiversity loss [1]. Its impacts are being felt across biological kingdoms, latitudes, ecosystems, biomes, and habitats [2]. According to the 2022 Intergovernmental Panel on Climate Change (IPCC) [3], the consequences of climate change have already altered marine, terrestrial, and freshwater ecosystems worldwide, with their effects being experienced earlier and more widely than ever predicted. Local plant and animal populations have already suffered losses, many of them due to considerable increases in the hottest yearly temperatures and heatwave events. [3]

The most prominent manifestations of climate change include changes in temperature, precipitation, drought frequency, seasonality, shifts of climate zones, and extreme weather events, which have a tendency to not only become more intense but also more frequent [2, 3]. At the level of the individual, these pressures can be felt through changes in life history, physiological, morphological, phenological, and genetic characteristics, which affect survival and/or reproduction [2, 4]. This in turn influences populations through changes in their distribution and genetic characteristics, which affects their abundance and the dynamics between populations. At the species level, this may culminate in an increase in their vulnerability to extinction [2] and, ultimately, in alterations in the ecosystem's structure [4].

1.2 The Mediterranean Basin and its climate

The Mediterranean climate can be characterized by mild, wet winters and warm to hot, dry summers. This type of climate can be found in other parts of the globe, generally on the western side of a continent, between latitudes of 30° to 40°, but the Mediterranean region has some particular characteristics that set it apart. The presence of such a large body of water, such as the Mediterranean Sea, is unique to this region. The sea has an area of about 2.5 million km², with an average depth of 1,500m. It is almost completely closed, connecting to the Atlantic Ocean through the Gibraltar Strait. Another characteristic of this region is its particular morphology, defined by many basins, gulfs, and peninsulas. The presence of mountain ranges surrounding the Mediterranean Sea on most sides leads to the existence of particular climatic features that wouldn't exist in their absence. These characteristics lead to a high degree of spatial variability and the existence of several sub-regional and mesoscale features [5]. Within the basin, its north-eastern and north-western regions can be classified as colder and wetter, while its south-eastern and south-western regions can be classified as hotter and more arid.

Annual precipitation may range from 400 to 800 mm and mean annual temperatures may range from 16 to 19 °C. Besides the characteristic dry summers, the occurrence of dry periods during the colder months is common due to the formation of thermal anticyclones and may persist for long periods under specific circumstances. Fronts of humid air from the Atlantic make the Spring the wettest season in most areas, especially the Iberian Peninsula [6]. In Morocco, the Atlas mountain range may experience annual rainfall above 200 cm and in the mountains of north-west Tunisia may experience more than 150 cm of annual rainfall [7].

The development of arid climates in the Mediterranean started during the Early Paleogene, marked by the expansion of evergreen sclerophyllous shrubs, and continued through the Late Paleogene and the Neogene. These changes in climate led to an increased frequency of xerophytic (drought-resistant) taxa showing higher resistance to this type of climate. Nevertheless, the Mediterranean climate as we know it today, characterized by periods of summer drought, was only established in the northern hemisphere at the end of the Neogene and beginning of the Quaternary [6].

Although the last 8000 years have observed a relatively stable climate, the Quaternary has been marked by several climatic oscillations which have shaped present-day biogeographical distributions [8]. The Pleistocene (early Quaternary) observed a series of climate fluctuations [9]. Major glaciation events in the northern hemisphere started around 2.8 Myr BP. This period was characterized by intense glaciations and long glacial-interglacial cycles [8]. During this period, refugia were generally areas that experienced warmer, more humid climates than the surrounding regions, which made them crucial for the persistence of biodiversity [8]. Although major Mediterranean peninsulas (Iberian, Italic, and Balkan) are normally pointed as the main hosts of glacial refugia, Medařl & Diadema (2009) [8] have pointed to several other locations outside of the peninsulas that could have served this purpose, such as Mediterranean islands, and several locations in North Africa, normally associated to mountain ranges. These glacial refugia shaped current patterns of genetic diversity by inducing diversification events and being areas favorable to plant speciation. These events happened through vicariance events followed by allopatric differentiation, where gene flow was majorly reduced. A lot of these populations were well established before the Quaternary and were able to maintain high levels of genetic variation. A wide diversity of ecological niches and the ability of species to migrate in altitude through glacial and interglacial cycles made mountain refugia, in particular, very adequate locations for the long-term survival of plant species and the occurrence of speciation events [8]. During interglacial periods, and after the Last Glacial Maximum in particular, with rising global temperatures, populations expanded their range as favorable habitats became available [8, 9]. The mode and rate at which this range expansion took place varied by species, mainly because of differences in dispersal capacity, location of the refugia, and barriers to expansion [8].

Given the expected changes in climate it is expected to experience in the coming decades, the Mediterranean region has been described as a "climate change hotspot". Evidence points to significant warming of the entire region [10, 11]. This warming is expected to take place for all seasons [11], with higher rates of warming in summer and autumn, and lower rates of warming for spring and winter [12]. The highest increase in temperature is expected to take place in the interior of the Iberian and Balkan Peninsulas [11], particularly during summer [10]. Declines in precipitation are also expected across most parts of the Mediterranean during the summer, and in the south specifically during the winter [10, 11]. An increase in the length of dry spells is projected [11, 12], leading to an increase in the risk of droughts, with a significant decrease in soil moisture, especially in the summer [11].

Taking into account the rate and intensity at which climate change is predicted to take effect, especially in the Mediterranean region, it is important to understand if species have the capacity to respond to the changes in their environment, and if so, predict how they might respond. This might allow us to respond proactively through conservation measures aimed at lessening the great losses in biodiversity that are expected as a result of climate change [13].

1.3 How species may adapt to climate change

In general, species may adapt to climate change, either by adjusting their range through dispersal, accommodating to the changing climate through phenotypic plasticity or through epigenetic modifications, or responding through genetic adaptation.

Range shift consists of changes in species distribution in response to changes in their environment [2]. This allows species to persist by tracking their ecological niches spatially [14]. The ability of a species to shift its range will depend on (i) its movement ability (either through migration or seed/pollen dispersal) and (ii) its capacity to establish itself at a new site [2]. Its movement ability is influenced by

intrinsic capabilities, such as dispersal rate or potential for long-distance dispersal [2], and by extrinsic factors, such as the level of fragmentation of their current habitat [13] the permeability or existence of physical barriers along their dispersal routes [2]. Its establishment capacity will depend on the available resources at the new site [2].

Phenotypic plasticity consists of modifications in fitness traits that allow an organism to persist changes in its environment, without underlying genetic change [4]. It can be advantageous as a relatively rapid first line of defense against environmental variation at an individual level [4], especially when faced with strong, abrupt episodes of directional selection [15]. Adaptive plastic responses may be a first step in the process of adaptation [15, 16], by allowing individuals sufficient time to survive/reproduce so that they can then respond through genetic adaptation [15]. By contrast, adaptive plasticity may also hinder adaptation by reducing selective pressures and slowing the rate of evolution [15, 17].

Another mechanism that may also play an important role in adaptation to climate change is epigenetics. Epigenetic processes consist of chemical modifications to chromatin that modulate gene expression without altering the genetic code. Epigenetic modifications to DNA happen in response to developmental or environmental triggers and may act as a source of phenotypic variation. Similar to phenotypic plasticity, these processes also allow for a fast response of individuals to changes in their environment, acting as a buffer until organisms are able to respond through genetic adaptation. A characteristic of epigenetics that awards it adaptive value is the transgenerational inheritance of epigenetic modifications, given that parents and offspring share similar environments and environmental fluctuation is slow and longer than a species generation time [18].

Genetic adaptation consists of changes in allele frequencies over many generations driven by natural selection [4]. Genetic adaptation may draw from two sources: standing genetic variation or novel mutations. Adaptation from standing genetic variation tends to be the faster route mainly because: (i) the alleles are already present in the population, leading to a faster response; (ii) alleles are present in higher frequencies, reducing fixation time; and (iii) alleles may have been previously selected in similar environments, be it in the past, or in another part of the species' range [19]. The amount of genetic variation present in a species will determine its capacity to respond to new climate conditions and determine if the rate of response is sufficient to keep up with climate change [2]. The occurrence of gene flow may affect the process of genetic adaptation in several ways, either positive or negative. For example, it might hinder local adaptation by homogenizing allele frequencies, constraining adaptive divergence [20], but it might also enhance the amount of genetic variation in a population [19, 20], and facilitate the tracking of environmental changes [20].

In sum, the magnitude of plasticity in a population may determine its short-term response potential, while the speed at which genetic adaptation can occur in a population may determine its long-term response potential [13].

Tree species are characterized by long lifespans, long generation times, sedentarism, large distribution ranges in general, and capacity for long-distance dispersal through pollen and seeds [20]. A long generation time can mean a lower response capacity to new selective pressures, especially in the presence of rapid climate change, but might also increase the probability of a population to persist over short timescales [13]. For species with limited dispersal ability, their capacity to migrate might be outpaced by the accelerated changes in their environment [13]. Fragmentation and modification of current habitats might also lead to a decrease in dispersal ability across landscapes for most plant species [13]. These characteristics of tree species make them a case study of particular interest regarding their response to climate change.

1.4 General characteristics of Holm oaks

Distribution, habitat, and ecology

Holm oaks are evergreen trees of the genus *Quercus*, sub-genus *Cerris*, section *Ilex* [21]. Native to the central-western Mediterranean region, they're among the dominating species in Mediterranean woodlands and maquis. Their distribution ranges from Portugal and Morocco in the west to the Aegean Islands and western Turkey in the east, stretching northward to northern Italy and to the Atlantic coast of France, and southward through the north of Africa. They grow in variate climate conditions, from semi-arid to very humid in terms of precipitation, and from warm to cold (mainly at high altitudes, associated with low precipitation) in terms of temperature [22]. Their distribution seems to be limited mainly by competition in terms of water availability [22].

In the Atlas mountains of Morocco, where forests are simpler and competition is reduced, Holm oak forms dense, high-altitude forests [22] reaching 2700 m [23]. In Tunisia, Holm oak can be found in the north in the region of High-Tell, which experiences a wetter climate due to the proximity to the Mediterranean Sea, and the Tunisian Ridge, which can be characterized by a drier climate [24]. In eastern Algeria, Holm oak is one of the main components in the forest landscapes of the Belezma and Aures mountains [25].

In the Iberian Peninsula, Holm oaks can be found in multi-purpose agroforestry systems characterized by grasslands with scattered oak trees planted by man. These are known as *dehesas* in Spain and *montados* in Portugal. Besides Holm oaks, Cork oaks (*Quercus suber*) are also characteristic of these systems. The grasslands are composed mainly of herbaceous native species used for grazing. The oaks provide shade for livestock, and acorn production goes towards feeding pigs. [22, 26]. While Cork oaks have been subjected to intense artificial selection in order to improve cork production throughout their distribution, the selection of specific traits in Holm oaks is restricted to acorn and firewood/coal production in the Iberian Peninsula. This led to a reduction in genetic variability in Cork oak that is not observed in Holm oak [9].

1.4.1 Taxonomic classification of Holm oaks

In the Holm oak group, two species are commonly distinguished: *Quercus ilex* and *Quercus rotundifolia*. Despite this, the taxonomic classification of these two taxa has not always been consistent, and they may appear classified as two subspecies, *Quercus ilex* subsp. *ilex* and *Quercus ilex* subsp. *rotundifolia* (also referred to as *Quercus ilex* subsp. *ballota*), or simply two varieties of the same species [27]. From this point forward, these taxa will be referred to as distinct species.

Several criteria have been used to distinguish between the two taxa. The first is their distribution areas and the associated bioclimatic conditions: *Q. ilex* inhabits humid or sub-humid areas, from the east of the Mediterranean to the Atlantic coast of France; *Q. rotundifolia* can be found in inland areas marked by dry Mediterranean climates, such as North Africa and the Iberian Peninsula [28, 29].

A second criterion, being the most widely used, is based on leaf morphological traits. Here, *Q. ilex* and *Q. rotundifolia* can be separated into two different morphotypes. The former is characterized by large, elongated leaves with an entire and serrate margin, with teeth that are not sharply pointed, and weak brochidodromous or mixed craspedodromous secondary venation [28]. The latter is characterized by small, rounded, thick leaves with a serrate margin, sharply pointed teeth, and craspedodromous secondary venation [28, 29]. An intermediate morphotype has been described, inhabiting coastal areas of eastern Spain and south-eastern France, morphologically closer to the *Q. rotundifolia* morphotype, but with more

elongated leaves [29]. Lebreton et al. (2001) [30] proposed a similar classification, with the additional parameter of leaf biochemical composition: *Q. ilex*, characterized by elongated leaves, high amount of veins, and high concentration of kaempferol and prodelphinidine; *Q. rotundifolia*, characterized by short leaves, reduced number of veins, and low concentration of kaempferol and prodelphinidine, present in the Iberian Peninsula; and *Q. rotundifolia* subsp. *maghrebiana*, characterized by rounded leaves, reduced number of veins, and high concentration of kaempferol and prodelphinidine, present in North Africa.

These morphotypes have also been shown to be genetically different, with *Q. ilex* and *Q. rotundifolia* displaying different chlorotypes, the latter species being divided into a North African and an Iberian group [28]. Both morphotypes have also been shown to correspond to distinct ecotypes, which contrast in terms of functional traits related to drought stress response. This functional classification points to *Q. ilex* as the less tolerant and *Q. rotundifolia* as the most tolerant morphotype in terms of drought stress [29].

Origin of section *Ilex* and Holm oak presence in the Mediterranean

Evergreen oaks are considered by some authors as older, more "primitive" taxa displaying tropical/sub-tropical affinities when compared to their deciduous counterparts [9]. Mediterranean species of section *Ilex* might have their origin in the East Asian/Himalayan subtropical-tropical humid forests of the Eocene. Fossil evidence supports a wide establishment of this section in Eurasia by, at least, the Oligocene. Morphological similarities between extant Mediterranean species and species of the Himalayan sub-Alpine clade support a corridor between these two groups. Westward dispersion of this clade from East Asia to Europe likely took place during the late Oligocene through a warm, humid forest zone along the margin of the Himalayas and across Tibet. This Tibet-Himalaya corridor likely owes its origin to climatic changes brought on by major geological events, such as the disappearance of the Tethys Sea and the evolution of the Paratethys Sea [31].

Floral records from several Mediterranean regions dating to the Early and Middle Miocene show the presence of probable sclerophyllous leaves belonging to sect. *Ilex* [32]. Floral records from the *Tejo* basin in Portugal dating from the Late Miocene show the presence of *Q. rotundifolia*, as well as other modern taxa [32]. Sect. *Ilex* was a common component of the Mediterranean vegetation during the Pliocene [32]. During the Quaternary, Holm oak's distribution went through several fluctuations [33]. It was abundant in central and southern Spain 9,500 years ago, and in north-eastern Spain 5,000 years ago. In the northwest of the Mediterranean basin, the presence of Holm oak is more recent, probably driven by human influence [23]. In the eastern Mediterranean basin, an increase in Holm oak abundance around 4,700 years ago is also thought to be associated with human activity. In northern Africa, Holm oak was probably present even earlier in extensive forests [23].

Lumaret et al. (2002) [28] laid out a hypothesis for the colonization of the Mediterranean by Holm oak, based on the phylogeographical variation of chloroplast DNA. They argue that migration and differentiation occurred initially from east to west and that current *Q. rotundifolia* populations descend from a lineage that originated in the central-eastern Mediterranean, migrating east to west during the Neogene, colonizing northern Africa and then the Iberian Peninsula [28]. This would be congruent with the Neogene origin of *Q. rotundifolia* in the Rif pointed out by Muller et al. (2015) [34]. Furthermore, Lumaret et al. (2002) [28] argue that, during the Quaternary glaciations, several populations of Holm oak contracted into glacial refugia, probably located in the south of the Balkan, Italian, and Iberian Peninsulas and in North Africa. In southern Iberia, these refugia were likely associated with the Baetic mountain range [8].

In Morocco, mountain ranges such as the Rif Mountains and the High and Middle Atlas have been pointed as putative glacial refugia [8]. In these cases, populations might have been able to survive in areas associated with these mountain ranges, such as valleys or plains, with their distribution fluctuating during glacial and inter-glacial periods, as conditions changed. Also in northern Africa, locations in present-day Algeria and Tunisia [8] may have served as refugia. Other locations in the Mediterranean basin that might have served as putative refugia include Mediterranean islands such as Corsica, Sardinia, or Sicily, and even some regions in southern France [8].

Finally, Lumaret et al. (2002) [28] argue that the subsequent post-glacial colonization from distinct refugia originated the current populations of Holm oak, and that distinct post-glacial migration routes gave rise to the observed phylogeographical distribution.

Functional traits of Holm oak related to temperature and drought-resistance

In geographical regions that display a Mediterranean climate, the two main stressors tend to be high temperatures and periods of low water availability. Because of this, species living under these conditions tend to display a series of adaptations to deal with heat and water stress, and Holm oak is not an exception.

As mentioned before, Mediterranean species of section *Ilex* might originate from East Asian / Himalayan subtropical-tropical humid forests of the Eocene, colonizing Europe through a warm, humid forest corridor. The same climatic conditions that gave way to the creation of this corridor might also have contributed to the early and rapid ecological differentiation of this clade [31]. This diversification might have been driven by periods of low rainfall and increased temperatures inducing water stress [35]. Furthermore, the evidence of strong phylogenetic inertia in terms of climate specialization in this section [31], strengthens the hypothesis that this ancestral Asian climate might have led to the pre-adaptation of section *Ilex* to the present European climate, specifically regarding drought-coping mechanisms. [31, 35].

Reduction in leaf size is usually associated with adaptation to dry habitats. Evergreen Mediterranean oaks display the smallest leaf size when compared to oaks from other climates. This has several advantages, such as resistance to cavitation in water-stress conditions, facilitation of heat loss in warm climates [6] and maintenance of leaf water supply in dry environments [6, 29]. Lower leaf size has been shown to lead to a lower degree of overheating in *Q. ilex* [6] and better hydraulic performance in the more xeric areas *Q. rotundifolia* inhabits [29]. *Q. rotundifolia* shows a high degree of resistance to leaf xylem cavitation, which may allow maintenance of foliage through high water-stress periods such as summer [36]. *Q. ilex* has also been shown to have higher cavitation resistance and lower xylem vulnerability to drought-induced embolism [37, 38]. The lower vulnerability of *Q. ilex* and *Q. rotundifolia* may be associated with their diffuse-porous wood and higher wood density, which works by reinforcing vessel walls and allowing for the maintenance of a functional hydraulic system even in periods of drought. Additionally, *Q. rotundifolia* has been reported to display plasticity in terms of vessel diameter in response to seasonal drought [39].

Evergreen oaks in general, and *Q. rotundifolia* in particular, have been shown to display a lower maximum stomatal conductance and higher stomatal sensitivity to drought [40, 41], translating into more conservative water-saving strategies which may delay desiccation and contribute to leaf longevity [40]. Despite this, evidence shows that this water-saving strategy is not present in earlier life stages of the species, indicating that seedlings may be especially vulnerable to drought periods. *Q. ilex* has also shown drought-responsive rooting systems, characterized by deeper root systems with higher production of fine roots in deeper soil layers, leading to better performance during drought periods [42]. The exis-

tence of trichomes in the leaf surface of *Q. rotundifolia* also seems to constitute an advantage in the more xeric regions it inhabits. They've been shown to protect photosynthetic machinery from photodamage, especially important in areas with elevated levels of irradiance such as the Mediterranean [43]. In *Q. ilex*, they've been shown to contribute to leaf re-hydration, important in xeric conditions with low precipitation, and maintenance of adequate gas exchange [44], important in foggy mountainous areas [6], for example.

1.5 Obtaining genome-wide SNP data using Genotyping by Sequencing

The advent of high-throughput Next-Generation Sequencing (NGS) technologies has allowed for the production of high amounts of sequencing information in a highly affordable manner. This has allowed scientists to expand not only the number of samples they can sequence but also the number of sequences they can get from the process [45]. The use of Genotyping by Sequencing (GBS) approaches, in particular, allows for the sequencing of a targeted fraction of the genome using restriction enzymes and the production of reduced representation libraries using high-throughput NGS technologies. This reduction of genome complexity allows for the discovery and genotyping of a high number of variant positions throughout the genome (such as single nucleotide polymorphisms - SNPs), even for species with reduced genomic resources available or with large genomes [46, 47]. This type of approach, commonly referred to as reduced representation methods, is very valuable for population genomics, as it allows for the genome to be scanned for neutral and adaptive genetic variation, which, historically, had been done using a reduced number of molecular markers, such as allozymes or microsatellites, that covered a very reduced extent of the genome. The use of GBS has contributed to several works involving analyses such as phylogenetic reconstruction, phylogeography, population structure and differentiation, QTL mapping, and identification of outlier loci, among others [46]. However, despite its advantages, the use of GBS methods involving restriction enzymes may involve biases. An example is restriction fragment length bias, where shearing may lead some fragments to be discarded from sequencing based on size. This may lead to read depth variation, as some loci end up being sequenced at very low depths. Another example is restriction-site polymorphism, which may lead to loci being wrongly genotyped as homozygous because of alleles that were not sequenced [46, 48]. These biases may in turn have an effect on downstream analyses (for example when calculating population genetics statistics) and must be addressed by investigators when filtering their data [46].

The processing of GBS raw data into usable SNP datasets is done through SNP calling, which aims to determine the positions in which polymorphisms, in this case SNPs, exist [49]. This can be done with the use of several different software pipelines. Two such pipelines are *ipyrad* [50] and *Stacks* [51]. *ipyrad* is an open-source tool for genome assemblage and analysis of restriction site-associated DNA sequence datasets, such as those generated by GBS. It is a powerful tool given that it integrates a branching design that allows for the creation of multiple datasets. In this way, *ipyrad* allows the exploration of a range of settings, without the need to re-run the entire assembly [50]. It clusters/maps loci within samples, either *de novo* or using a reference genome, estimate error rates and heterozygosity, which it uses during consensus base calling, and then clusters/maps consensus sequences among samples. *Stacks* is a software pipeline for building loci from short-read sequences, developed to work with restriction enzyme-based data (such as those generated by GBS), for conducting population genomics studies. The software is designed as a modular pipeline to efficiently curate and assemble large numbers of short-read sequences from multiple samples. It identifies loci in a set of individuals, *de novo* or using a reference genome, and then genotypes each locus. *Stacks* incorporates a maximum likelihood statistical model to identify

sequence polymorphisms and distinguish them from sequencing errors and employs a catalog to record all loci identified in a population, which it then matches individuals to, to determine which haplotype alleles are present at every locus in each individual [51, 52].

Several studies have shown that different SNP calling pipelines present particular advantages and disadvantages [53–56], and might even produce different results [57, 58]. For example, the analysis of the same Cork oak (*Q. suber*) GBS dataset using two different SNP calling pipelines, *ipyrad* [59] and *Stacks* [60], led to slightly different results population structure results. Consequently, to avoid this problem and ensure consistent results, the use of more than one SNP calling pipeline might be advised.

1.6 Complementary approaches for the inference of population structure

Population structure may be defined as the organization of genetic variability within and between populations, where genetic similarity tends to be structured by the existence of subgroups or geographic isolation [61, 62]. Structuring may arise from several different factors, which may be intrinsic to an organism's life history, or be the result of the past demographic history and biogeography of populations [62]. Understanding population structure is then an essential step when studying the population genetics and genomics of a particular taxon [61], as it not only informs us about its demographic history and the factors that may have led to the observed structuring but since it may also act as a confounding factor in several analyses, such as detection of selection [63].

There are two widely used approaches for the study of population structure: the Principal Components Analysis (PCA) [64], a non-parametric approach in which the principal components of a similarity matrix are used to maximally explain observed patterns of genetic similarity among samples [65]; and *STRUCTURE* [66], a parametric, model-based Bayesian clustering method for using multilocus genotype data to infer population structure and assign individuals to populations. The latter was introduced by Pritchard et al. (2000) [66], and has since seen wide use in the fields of human genetics, evolutionary genetics, and molecular ecology. In this approach, a model in which there are K populations is assumed, and individuals are assigned probabilistically to one of K populations, or more in the case they are admixed [66]. Since its introduction, this method has not only been repeatedly expanded by its authors but also by others [61]. Several authors have implemented *STRUCTURE*'s underlying inference method, making several refinements allowing, for example, for larger datasets to be analyzed in a faster way [67] (*ADMIXTURE* [68], for example).

Although widely used, both of the above approaches analyze single mutations individually, and lack integration of other information, such as the relative positions of these mutations in the genome [69]. Lawson et al. (2012) [69] proposed a population structure inference framework based on patterns of haplotype similarity, which they implemented in *fineSTRUCTURE* and *chromoPainter*. In this method, each individual is considered as a recipient, whose chromosomes are reconstructed using chunks of DNA donated by the other individuals [69]. This type of analysis allows the haplotype similarity between individuals to be summarized in a 'coancestry matrix', which holds information regarding the ancestral relationships among individuals and allows for the identification of discrete populations [69]. *fineSTRUCTURE* is a high-resolution method, which owes to the use of haplotype linkage information and its focus on recent common ancestry between samples [65]. Malinsky et al. (2018) [65] later developed *fineRADstructure*, a population inference package for restriction-site-associated DNA sequencing (GBS or RAD-seq) data. It combines *fineSTRUCTURE* [69], with its powerful, high-resolution, model-based approach to investigate population structure and recent shared ancestry, and *RADpainter*, which allows for the coancestry matrix to be inferred from GBS or RADseq data.

Recognizing the merits and drawbacks of both likelihood-based approaches (such as *STRUCTURE* or *STRUCTURE*-based methods) and PCA-based approaches, Cabrer0s et al (2019) [70] attempted to 'bridge the gap' between these two approaches. As so, they introduced *ALStructure*, which they described as a computationally efficient and statistically accurate method for modeling admixture and decomposing systematic variation due to population structure [70]. *ALStructure* is a PCA-based method providing a substitute for popular likelihood-based methods, eliminating the main shortcomings of PCA-based methods, specifically their inability to provide estimates of admixture proportions, and retaining their main advantages over likelihood-based methods, such as computational efficiency, accuracy, and the minimal assumption made regarding the data-generative model [70].

The use of complementary, methodologically different approaches for the inference of population structure, such as those described above, ensures that the results of the analysis are sound and consistent.

1.7 Understanding the adaptive process through Landscape Genomics

The large amount of data that can be obtained with approaches such as GBS allows for genome-wide scans for signs of natural selection. The first methods developed for detecting natural selection were based on the detection of outlier F_{st} values. These were based on the idea that a locus under the effect of natural selection would show a value of differentiation beyond what is expected under neutral conditions [71]. These values might be unusually high in the case of divergent selection, or unusually low in the case of balancing selection [72]. This allows for the detection of outlier loci, possibly under the effect of natural selection, and with putative roles in adaptation [71]. Although these types of tests are widely used, they present a series of drawbacks. The first is that they are not able to identify the specific environmental selective forces driving evolution at the outlier loci. The second is their tendency to display high false-positive rates if they're not adjusted for evolutionary non-independence among samples. Because of population structure, demographic history, or dispersal patterns (bottlenecks, range expansions, etc.), neutral regions of the genome may vary substantially in their F_{st} values, creating signatures that may be wrongly identified as being caused by natural selection [72]. A third drawback of these methods stems from the fact that adaptation, in a lot of cases, happens through the additive effect of many small effect loci. In such cases, loci may only experience subtle changes in terms of allele frequencies between populations. Therefore, F_{st} outlier tests will not be able to detect these loci, even though they play a part in the adaptation process [73].

Besides genetic and genomic data, the availability of environmental data for several climatic variables such as temperature and precipitation has allowed for the development of methods that take into account the role of the environment as a driving force behind natural selection. Methods that take into account the relationship between genotype and environment are collectively referred to as genetic-environment association [74]. These methods aim to identify alleles showing unusually strong correlations with environmental variables, suggesting their involvement in the process of local adaptation [63]. This can be done through what is known as a top-down approach, where we start with genome-wide genetic marker data and try to identify adaptive genetic variation among our set of markers. The use of markers such as SNPs is advantageous, as it allows for the identification of functional genes and regulatory regions under selection. Genetic-environmental association presents an advantage in relation to F_{st} -outlier-based methods when it comes to dealing with polygenic traits. In these cases, where loci may only show subtle changes in allele frequencies between populations, genetic-environmental association may still be able to detect the effects of natural selection if numerous populations have been sampled along an environmental gradient [75].

Despite these advantages, as with F_{st} outlier tests, early implementation of genetic-environmental association methods failed to control for population structure, leading to the occurrence of high false-positive rates [74]. Multiple approaches arose, aiming to control for population structure in outlier tests and genetic-environmental association. One of these is implemented in *BayPass*, a software that includes several methods to characterize adaptive genetic differentiation across populations [76]. The program builds upon the Bayesian hierarchical model put forward by Coop et al. (2010)[77] and allows not only for the detection of outliers but also for the quantification of their association with population-specific covariables (such as environmental variables, quantitative or categorical phenotypic characteristics, etc.). An important characteristic of this model is its ability to correct for population structure. This is done through a population covariance matrix, which incorporates the populations' neutral correlation structure and allows for their demographic history to be taken into account [78].

In sum, the use of large sample sizes and sequencing approaches that allow for the production of large amounts of genome-wide SNP data, in conjunction with top-down landscape genomics approaches may consist in a first step in understanding the effect of selection on natural populations [75], which is especially important given the high selective pressures expected to be brought on by climate change.

Even though reduced-representation methods such as GBS may be considered a great advancement in terms of the amount of data they yield when compared to previous approaches (use of microsatellites, for example), they may still be considered too sparse to detect the majority of adaptive loci. Such is the case because the SNPs captured by these methods remain a small representation of the whole genome. This means that, unless a gene is tightly linked to a marker, it will be missed by our analysis, which tends to be more problematic the higher the genome size. Some strategies may be used to minimize these problems. One of them is the use of reference genomes, which would improve genotyping accuracy, allow the assessment of haplotype structure and, if annotated, allow for the investigation of the putative function of genes and regulatory regions related to adaptive variants [63]. Still, until the accessibility of whole genome sequencing extends to non-model organisms, which would improve the detection of loci under selection [63], the use of genome-wide SNP data resulting from reduced representation methods such as GBS, in conjunction with top-down landscape genomics approaches may consist in useful and worthwhile step in understanding the effect of natural selection on natural populations [75].

1.8 Objectives of this thesis

In this thesis, our aim is to investigate population structure and the existence of local adaptation for samples of Holm oak collected along the West Mediterranean basin. The specific objectives are:

1. Analyze the genetic diversity and the population structure of Holm oak using several complementary approaches and based on different SNP calling and filtering strategies;
2. Implement outlier detection methods and genetic-environmental association analyses to detect SNPs putatively under the effect of natural selection and annotate them.

This is, to our knowledge, the first attempt to use nuclear, genome-wide genetic markers to uncover the existence of population structure and signatures of local adaptation in Holm oak.

Chapter 2

Materials and Methods

2.1 Study species and sampling sites

This work used a total of 94 Holm oak samples, collected from a total of 18 sampling locations along the Mediterranean basin. The sampling locations were based on the known species distributions at the time of sampling [22] (Figure 2.1). The following samples were collected: 69 *Q. rotundifolia* samples, 49 from the Iberian Peninsula (Portugal and Spain) and 20 from North Africa (Morocco and Tunisia); 25 *Q. ilex* samples (France, Italy, and Catalonia, in Spain) (Table 2.1).

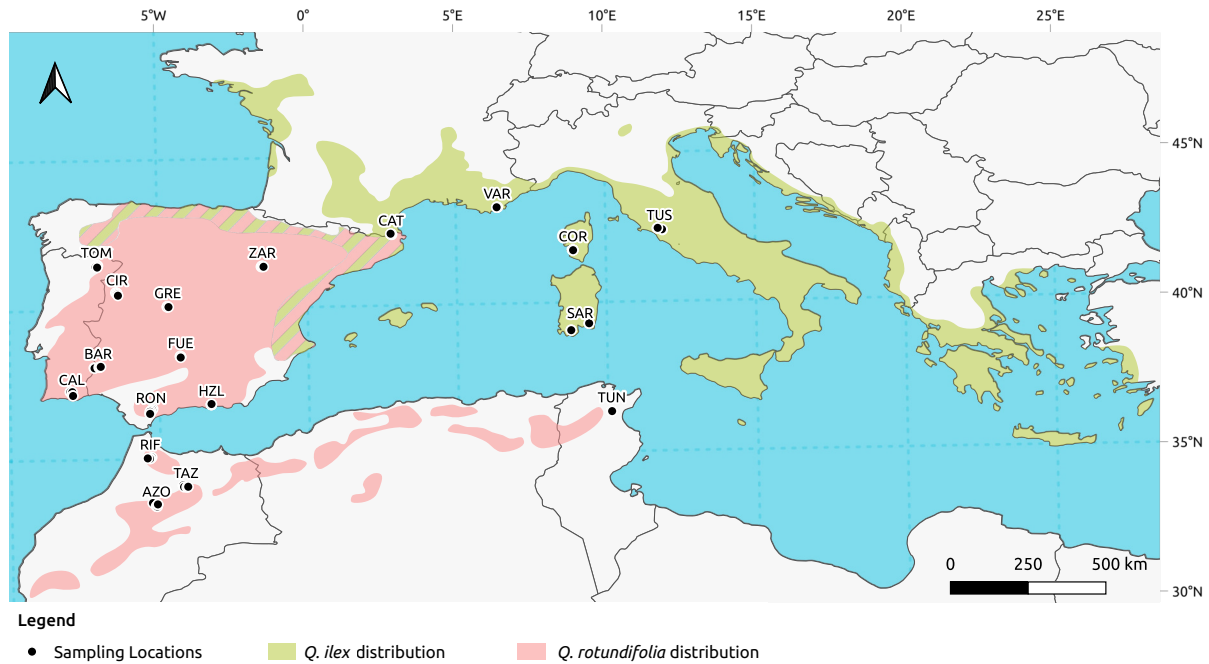


Figure 2.1: Sampling locations and species distribution range of *Q. rotundifolia* and *Q. ilex* (as presented in Cuadullo et al. (2017) [22]). Sampling location codes as in Table 2.1.

2.2 DNA extraction, quantification and quality assessment

For each sample, about five leaves were selected. These were then frozen in liquid nitrogen and ground into a fine powder using mortar and pestle. DNA was extracted from the grounded samples using the ‘innuPREP Plant DNA’ kit by Analytik Jena AG, following the manufacturer’s standard protocol with some minor modifications. DNA concentration of the extracted samples was measured using a Qubit 2.0 Fluorometer (Invitrogen). The integrity of the genetic material was verified through gel electrophoresis. Trial digestions were performed on 10% of samples using the EcoRI restriction enzyme. The 94 samples were then plated, dried using a thermocycler, packed, and sent to *The Elshire Group Ltd.* (New Zealand) for Genotyping by Sequencing (GBS).

2.3 Sequencing, raw data filtering, and SNP calling

The GBS data was generated following Elshire et al. (2011) [47] with the following changes: 100 ng of genomic DNA were used, 1.44 ng of total adapters were used, the genomic DNAs were restricted with EcoT22i enzyme, and the library was amplified with 18 PCR cycles.

Table 2.1: Sampling information. Additional information for each sample, including geographical coordinates, can be found in Supplementary Table S1

Sampling location code	Sampling location name	Country	Species	Number of samples
BAR	Barrancos	Portugal	<i>Q. rotundifolia</i>	6
CAL	Caldeirão	Portugal	<i>Q. rotundifolia</i>	5
TOM	Trás-os-montes	Portugal	<i>Q. rotundifolia</i>	6
CIR	Ciudad Rodrigo	Spain	<i>Q. rotundifolia</i>	5
RON	Ronda	Spain	<i>Q. rotundifolia</i>	6
HZL	Haza del Lino	Spain	<i>Q. rotundifolia</i>	5
FUE	Fuente Caliente	Spain	<i>Q. rotundifolia</i>	5
GRE	Gredos	Spain	<i>Q. rotundifolia</i>	5
ZAR	Zaragoza	Spain	<i>Q. rotundifolia</i>	6
CAT	Catalonia	Spain	<i>Q. ilex</i> (distribution overlap)	5
VAR	Var	France	<i>Q. ilex</i>	5
COR	Corsica	France	<i>Q. ilex</i>	5
TUS	Tuscany	Italy	<i>Q. ilex</i>	5
SAR	Sardinia	Italy	<i>Q. ilex</i>	5
AZO	Azrou	Morocco	<i>Q. rotundifolia</i>	6
RIF	Rif	Morocco	<i>Q. rotundifolia</i>	4
TAZ	Taza	Morocco	<i>Q. rotundifolia</i>	5
TUN	Tunisia	Tunisia	<i>Q. rotundifolia</i>	5

A general, simplified workflow for the raw data filtering and SNP calling process, which will be discussed in the following sections, is illustrated in Figure 2.2.

2.3.1 Demultiplexing and removal of mitochondrial and chloroplastidial reads

Demultiplexing of the sequencing library was performed using the *axe-demux* software [79], resulting in two paired-end FASTA files for each of the 94 samples sequenced. To filter mitochondrial DNA (mtDNA) and chloroplastidial DNA (cpDNA) reads from the nuclear DNA (nDNA), *BBMap* v38.18's [80] *BBSplit* function was used. For this, the complete mitochondrial genomes of *Populus davidiana* (KY216145.1) [81], *Pyrus pyrifolia* (KY563267.1) [82] and *Rosa chinensis* (CM009589.1) [83], and chloroplastidial genomes of *Quercus rubra* (JX970937.1) [84], *Quercus aliena* (KU240007.1) and *Quercus variabilis* (KU240009.1) [85] were used. The reads that matched either the mitochondrial or chloroplastidial genomes were removed from the original datasets. This resulted in one raw dataset for each sample, containing only nDNA reads. *FastQC* v0.11.9 [86] was used to perform a quality control check on this dataset.

2.3.2 SNP calling and filtering

The nDNA raw data were processed by two SNP (single nucleotide polymorphism) calling software programs: *ipyrad* v0.9.82 [50] and *Stacks* v2.55 [51, 70]. This was done to ensure that the results of this work were consistent and would not be skewed by the use of single SNP calling software. Two types of assemblies were run for each software: full assemblies, containing all 90 samples, to be used in the population structure analyses and calculation of summary statistics; and subset assemblies, containing

only *Q. rotundifolia* samples from Iberia or from Iberia and Morocco, to be used in the outlier detection and genetic-environmental association analyses using *BayPass* v.2.31 [76]. For the full assemblies, several runs were made with different parameter combinations. These were then compared in terms of the final number of variable sites and the amount of missing data. The combination of parameters found to be more adequate was then adopted for both full and subset assemblies. The remainder of this section describes this process:

ipyrad

First, an *ipyrad* assembly was created and steps 1 (importing demultiplexed sequence reads) and 2 (quality control, filtering, and adapter trimming) were run. Filtered reads were checked using *FastQC* to ensure their quality. This assembly was then branched into five: three were full assemblies and two were sub-assemblies, one containing only Iberian *Q. rotundifolia* samples, and the other having only Iberian and Moroccan *Q. rotundifolia* samples. Steps 3 through 7 (de novo clustering of reads from the same locus, estimation of sequencing error and sample heterozygosity, consensus base and haplotype calling within samples, orthology identification across samples, filtering and trimming of the assembled loci/writing of output files) were run for each of them.

The following parameters were set for all assemblies: *data type* was set as *pairgbs*; no population assignment file was given (all samples were considered as being part of a panmictic population); minimum depths for statistical and majority-rule base-calling (*mindepth_statistical*, *mindepth_majrule*) were set to 8; maximum cluster depth within samples (*maxdepth*) was set to 30,000; adapter filtering (*filter_adapters*) was set to strict; the maximum fraction of heterozygous bases allowed in consensus sequences (*max_Hs_consens*) was set to 0.08; the minimum number of samples where a SNP had to be present for it to be called (*min_samples_locus*) was set to 8; the maximum proportion of shared polymorphic sites in a locus (*max_shared_Hs_locus*) was set to 0.8; reads had the first six bases trimmed and maximum length set to 130 base-pairs (*trim_reads*).

For each of the full assemblies, the clustering threshold parameter (*clust_threshold*), which defines the level of sequence similarity at which two sequences are identified as being homologous, and thus cluster together, was set as 0.85, 0.90, and 0.95. The tested parameter combinations can be found in Supplementary Table S2. For the sub-assemblies, the clustering threshold was set as 0.85.

Two datasets were generated from each assembly, in the *VCF* format: an all-SNPs dataset containing every SNP that passed the filtering process (outputted by *ipyrad*) and a sub-dataset containing a single SNP per locus. The latter was obtained using the Python script *vcf_parser.py* (https://github.com/CoBiG2/RAD_Tools/blob/master/vcf_parser.py, as of commit “0e96c2c”). This was done by taking the all-SNPs dataset and excluding every SNP but the one closest to the center of each locus. This was done to avoid confounding signals of linkage disequilibrium in some downstream analyses.

Stacks

First, the *Stacks*’ *process_radtags* module, which is responsible for importing and cleaning raw sequence reads, was run with the following parameters: reads with uncalled bases (*-c*) or low-quality scores (*-q*) were removed and RAD-tag cut sites were rescued (*-r*). The restriction enzyme used was provided (*-e ecoT22I*). The Illumina Universal Adapter sequence was inputted into the program in order to remove any reads containing it (*-adapter_1 AGATCGGAAGAG*). No mismatches were allowed in the adapter sequence (*-adapter-mm 0*). The final read length was truncated to 130 base-pairs (*-t 130*). After this, the filtered reads were checked using *FastQC* to ensure their quality.

The reads were then processed by *Stack*'s *denovo_map.pl* pipeline, which wraps a series of modules responsible for the grouping of data from each individual and identification of polymorphic sites (*ustacks* module), the grouping of loci across individuals and writing of a loci catalog (*cstacks* module), the matching of individual's loci against the catalog and allele calling of each individual (*sstacks* module), and calculation of population genetic statistics from allelic states and writing of output files (*populations* module). The pipeline was run multiple times. Several full assemblies were run with different parameter combinations. These parameters were: the minimum number of raw reads required to form an exactly matching stack (*ustacks*: *-m*) which was set to 3, 5, and 8; the number of mismatches allowed between stacks within individuals (*-M*), which was set to 2, 5 and 10; the number of mismatches allowed between stacks between individuals (*-n*), which was set to 1, 5 and 10; the minimum percentage of individuals across populations required to process a locus was always set to 9% (*populations*: *-R 0.09*; all samples were considered as being part of a panmictic population). The parameter combinations can be found in Supplementary Table S3. Two sub-assemblies were run, one containing only Iberian *Q. rotundifolia* samples, and the other containing only Iberian and Moroccan *Q. rotundifolia* samples. The parameters used for these were: *-m 3*, *-M 2*, *-n 1*, and *-R 0.09*.

Two data sets were generated for each assembly, in the *VCF* format: an all-SNPs dataset containing every SNP that passed the filtering process and a sub-dataset containing only one SNP per locus. This was done to avoid confounding signals of linkage disequilibrium in some downstream analyses. For comparison purposes between full assemblies run with different parameter combinations, these sub-datasets were generated by re-running the *populations* module on its own with the additional parameter *--write-random-snp*, which keeps only a random SNP per locus. The final sub-datasets, to be used in subsequent analyses, were obtained using the Python script *vcf_parser.py* (https://github.com/CoBiG2/RAD_Tools/blob/master/vcf_parser.py, as of commit "0e96c2c"), which excludes every SNP but the one closest to the center of each locus. The use of this script instead of the built-in function of *Stacks* was done so that in subsequent analyses, any comparison between *ipyrad*'s and *Stacks*' results would be more robust.

VCF filtering

All *VCF* files resulting from the SNP calling process were filtered using *VCFTools* v0.1.16 [87]. Sites with a minor allele count less than 2 (*--mac 2*) and/or more than 80% of missing data (*--max-missing 0.80*) were removed. The percentage of missing data per sample was calculated using *VCFTools* and samples with missing data above 65% were removed. For the *VCF* files resulting from the sub-assemblies, the same parameters were used, with the addition of the parameter *--max-alleles 2* in order to restrict to two the number of alleles per locus, as *BayPass* v.2.31 [76] cannot handle loci with more than two alleles.

The several *ipyrad* and *Stacks* full assemblies run with different parameters were compared in terms of the number of variable sites and the amount of missing data. The datasets with parameter combinations that showed more adequate results were adopted for subsequent analyses.

In sum, the filtering and SNP calling process resulted in 8 final datasets, that were then used for subsequent analyses. Dataset names and assembly parameter information can be found in Table 2.2.

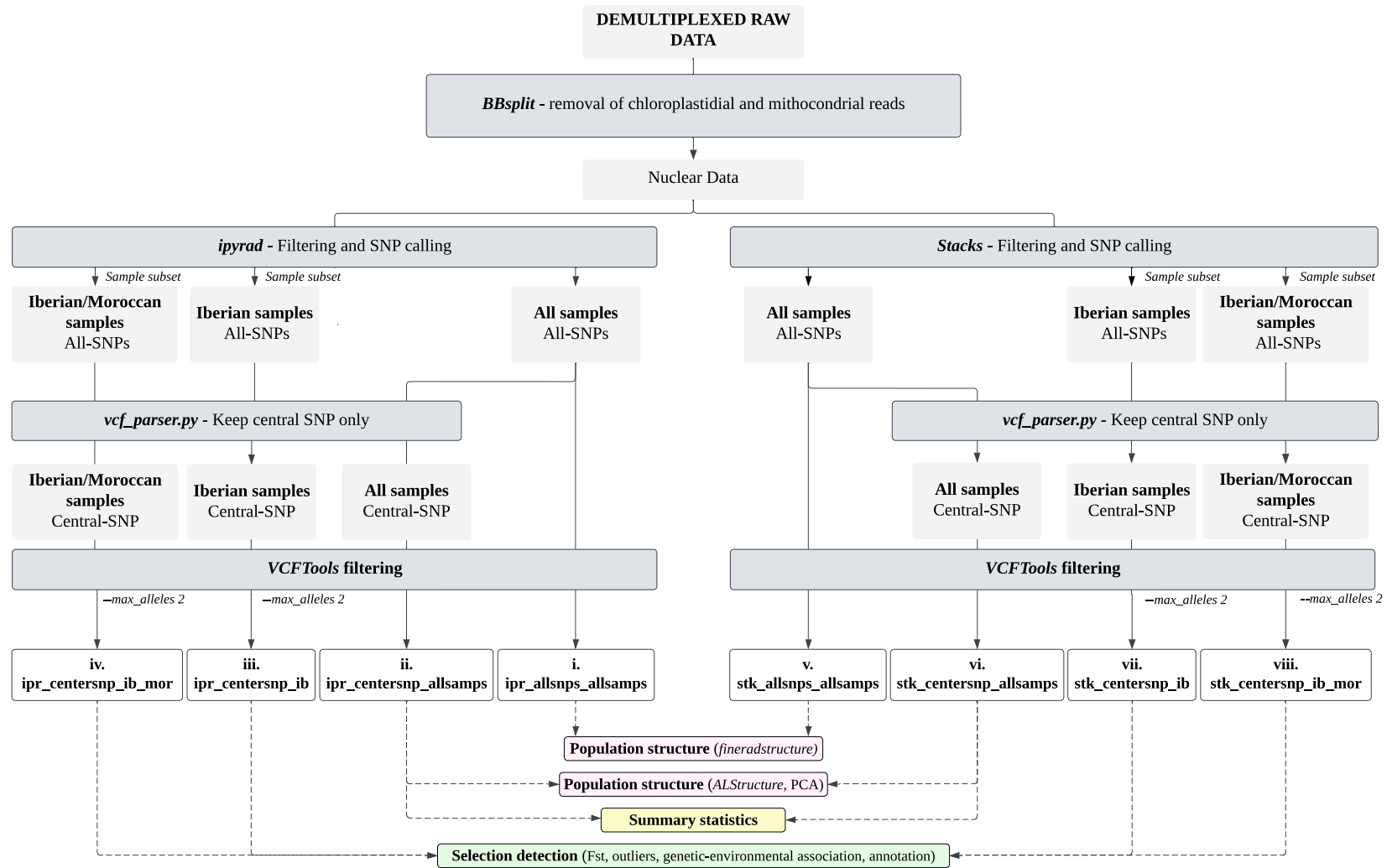


Figure 2.2: Simplified workflow of the filtering and SNP calling process using *ipyrad* and *Stacks*. Represented are the main steps in the filtering and SNP calling process for different assemblies, including the intermediate datasets. Only the eight final datasets are shown (Table 2.2), as well as the analyses for which they were used. The remaining assemblies run with different parameter combinations are not shown (Supplementary Tables S2 and S3).

Table 2.2: Summary table of all datasets used in the data analyses, showing the dataset names, included samples, SNPs used in final datasets, SNP calling software parameters, and *VCFTools* filtering parameters for (a) *ipyrad* and (b) *Stacks*. Datasets iii, iv, vii, and viii contain only a subset of the total samples, corresponding to either Iberian *Q. rotundifolia* or Iberian and Moroccan *Q. rotundifolia* samples.

(a) *ipyrad*

Dataset name	Samples	SNPs	<i>ipyrad</i> parameters							<i>VCFTools</i> parameters		
			<i>mindepth</i> <i>statistical</i>	<i>mindepth</i> <i>majority</i>	<i>maxdepth</i>	<i>max</i> <i>Hs</i> <i>consensus</i>	<i>min</i> <i>sam-</i> <i>ples</i> <i>locus</i>	<i>max</i> <i>shared</i> <i>Hs</i> <i>locus</i>	<i>clust</i> <i>thresh-</i> <i>old</i>	<i>--mac</i>	<i>--max</i> <i>miss-</i> <i>ing</i>	<i>--max</i> <i>alleles</i>
(i) <i>ipr_allsnps_allsamp</i> s	all	all	8	8	30,000	0.08	8	0.8	0.85	2	0.80	–
(ii) <i>ipr_centersnp_allsamp</i> s	all	center	8	8	30,000	0.08	8	0.8	0.85	2	0.80	–
(iii) <i>ipr_centersnp_ib</i>	Iberian	center	8	8	30,000	0.08	8	0.8	0.85	2	0.80	2
(iv) <i>ipr_centersnp_ib_mor</i>	Iberian/Moroccan	center	8	8	30,000	0.08	8	0.8	0.85	2	0.80	2

(b) *Stacks*

Dataset name	Samples	SNPs	<i>Stacks</i> parameters				<i>VCFTools</i> parameters		
			<i>-m</i>	<i>-M</i>	<i>-n</i>	<i>-R</i>	<i>--mac</i>	<i>--max</i> <i>missing</i>	<i>--max</i> <i>alleles</i>
(v) <i>stk_allsnps_allsamp</i> s	all samples	all	3	10	10	0.09	2	0.80	–
(vi) <i>stk_centersnp_allsamp</i> s	all samples	center	3	10	10	0.09	2	0.80	–
(vii) <i>stk_centersnp_ib</i>	Iberian	center	3	10	10	0.09	2	0.80	2
(viii) <i>stk_centersnp_ib_mor</i>	Iberian/Moroccan	center	3	10	10	0.09	2	0.80	2

2.4 GBS data analysis

2.4.1 Population Structure

fineRADstructure

Datasets 'ipr_allsnps_allsamps' and 'stk_allsnps_allsamps' were analyzed using *fineRADstructure*. These datasets contained every SNP that passed the filtering and SNP calling process and were created solely for this analysis, given that *fineRADstructure* is prepared to work with haplotype data, and takes advantage of haplotype linkage information. The results of the analyses were plotted in R v4.2.1 [88] using the script made available by *fineRADstructure*'s authors (<https://github.com/millanek/fineRADstructure/blob/4c9b385697f65be03097edf88f0d6647a3f5520a/fineRADstructurePlot.R>, as of commit "13036d3").

ALStructure and PCA analyses

Additionally, datasets 'ipr_centersnp_allsamps' and 'stk_centersnp_allsamps' were analyzed by *ALStructure* [70]. It was run wrapped under *Structure.threader* [89], a program to parallelize and automate the runs of *Structure*, *fastStructure*, *MavericK*, and *ALStructure*.

PCA analyses were also performed on datasets 'ipr_centersnp_allsamps' and 'stk_centersnp_allsamps' in R v4.2.1 using the package *SNPRelate* v1.28.0 [90], implemented in the Rscript *snp_pca_static.R* (https://github.com/CoBiG2/RAD_Tools/blob/master/snp_pca_static.R, as of commit "a69e9bd").

2.4.2 Summary Statistics

To characterize genetic variation, the following statistics were calculated for each population and overall: expected heterozygosity (H_e) and mean observed heterozygosity (H_o), inbreeding coefficient (F_{is}) and allelic richness (A_R). F_{st}/F_{st}^{max} was calculated for each population. This statistic takes a Q-matrix (common output of population structure inference software, that represents the membership coefficient of each individual for each cluster) and quantifies the variability in membership coefficients among a collection of individuals. Pairwise F_{st} was calculated between all sampling locations and between the populations obtained from the population structure analyses.

All statistics were calculated in R v4.2.1 for *ipyrad*'s and *Stacks*' center-SNP datasets. A_R and H_e were calculated using package *adegenet* v2.1.7 [91]. H_o , F_{is} , overall and pairwise F_{st} were calculated using package *hierFstat* v0.5-11 [92]. F_{st}/F_{st}^{max} was calculated using the package *FSTruct* v1.0.0 [93].

2.4.3 Outlier detection and genetic-environmental association

In order to detect putative SNPs under selection and to test possible associations between the environment and the observed patterns of differentiation, *BayPass* v.2.31 [76] was run under the script *BayPass_workflow.R* (https://gitlab.com/StuntsPT/pyRona/blob/master/pyRona/R/BayPass_workflow.R, as of commit "6e5908d8"). This was done for two geographical areas: the Iberian Peninsula, and the Iberian Peninsula and Morocco.

Datasets 'ipr_centersnp_ib', 'ipr_centersnp_ib_mor', 'stk_centersnp_ib', and 'stk_centersnp_ib_mor' were converted from VCF to the *.geste* format using *PGDSpider* v2.1.1.5. The software was given a population definition file, in which samples were grouped by sampling location. The resulting *.geste* files were then run through the *correct_geste.py* script (https://github.com/CoBiG2/RAD_Tools/blob/master/correct_geste.py, as of commit "dd753fc"), to correct formatting problems that might occur during file conversion. The corrected *.geste* files were then converted to the *.baypass* format using the *geste2BayPass.py* script (https://github.com/CoBiG2/RAD_Tools/blob/master/geste2BayPass.py, as of commit "b99636e").

Outlier detection under *BayPass*'s core model mode

Datasets 'ipr_centersnp_ib', 'ipr_centersnp_ib_mor', 'stk_centersnp_ib' and 'stk_centersnp_ib_mor' were analyzed under *BayPass* v2.31's core model mode. For each dataset the software was given the *.baypass* genotyping data file (*-gfile*), containing allele counts for each sampling location. The analyses involved the calibration of the XtX statistic with pseudo-observed datasets (PODs) generated with the function *simulate.baypass()*. After calibration, only the outliers above the defined 1% significance level threshold were accepted as putative loci under selection.

Environmental datasets and selection of bioclimatic variables

The environmental data for 19 bioclimatic variables for the period 1981-2010 was obtained from CHELSA V2.1 [94–96]. Spearman's correlation coefficient was calculated pairwise for all variables, in R v.4.2.1 (see Supplementary Figure S3). A correlation threshold was defined as Spearman's correlation coefficient (ρ) of ± 0.90 . For correlations where $\rho > 0.90$ or $\rho < -0.90$, variables were removed to eliminate the high correlation. In these cases, the variables to remove/keep were chosen based on the results of works where a similar analysis in the same geographic region was employed (Pina-Martins et al. (2019) [59] and Vanhove et al. (2021) [60]), as well as on the climatic context of the Mediterranean region. Since the datasets correspond to two different geographic areas (Iberia, and Iberia and Morocco), this was done for each of them. In both cases, 8 environmental variables were removed, and 11 environmental variables were kept, 10 of which are common to both areas. These variables, as well as their name and explanation, can be found in Supplementary Table S4).

Genetic-environmental association under *BayPass*'s AUX covariate mode

The environmental variable values for each sampling site were retrieved using *QGIS* v.3.22.11 [97]. For each sampling location, the average value of all sampling sites was calculated (average values for all 19 bioclimatic variables for each Iberian and Moroccan sampling location can be consulted in Supplementary Figure S4). Two covariate files were created: one for Iberia, containing 9 sampling locations, and one for Iberia and Morocco, containing 12 sampling locations. The datasets 'ipr_centersnp_ib', 'ipr_centersnp_ib_mor', 'stk_centersnp_ib', and 'stk_centersnp_ib_mor' were analyzed under *BayPass* v2.31's auxiliary covariate mode. For each dataset, the software was given the *.baypass* genotyping data file (*-gfile*) and the corresponding covariate file (*-efile*). The remaining parameters were kept as default.

The strength of the identified associations is assessed in BF_s, normally expressed in deciban (dB) units (via the transformation $10 \log_{10} \text{BF}$). The strength of the associations was classified according to

Jeffreys' rule following the db unit scale presented in Gautier et al. (2015): “strong evidence” when $10 \leq BF < 15$, “very strong evidence” when $15 \leq BF < 20$, and “decisive evidence” when $BF > 20$ [76].

Allele frequency analysis

The outlier and associated SNP data for each dataset ('ipr.centersnp_ib', 'ipr.centersnp_ib_mor', 'stk.centersnp_ib' and 'stk.centersnp_ib_mor') was retrieved from the original VCF files using the script *outlier_removal.py* (https://github.com/CoBiG2/RAD_Tools/blob/master/outlier_removal.py, as of commit "ba731f2").

Allele frequencies were calculated for the outlier and associated SNPs using *VCFTools*. Outlier SNP allele frequencies were plotted using R v.4.2.1. For the associated loci, to illustrate the pattern of allele turnover along the environmental gradients, the top three strongest associations for each of the common associated variables obtained for datasets 'ipr.centersnp_ib' / 'stk.centersnp_ib' and 'ipr.centersnp_ib_mor' / 'stk.centersnp_ib_mor' were retrieved. The major allele frequencies for the associated SNPs were plotted against the mean environmental variable values for each sampling location using R v.4.2.1. The common bioclimatic variables were also represented geographically for the respective geographical areas, using QGIS v.3.22.11.

Additionally, associated SNPs were classified into two groups according to their allele frequency turnover pattern along the bioclimatic variable they were associated with. The classification was based on a tendency for the major allele frequency to increase or decrease along the bioclimatic variable. This was done by fitting a linear regression to each SNP's allele frequencies along the bioclimatic variable and classifying them according to the slope of the regression. Then, for each variable, the average allele frequency and standard deviation at each sampling location were calculated for each of the two groups of associated SNPs, and plotted. This was done to all SNPs regardless of association strength, and only to the SNPs with the highest level of association strength.

Loci annotation and enrichment analyses

The sequences for the loci containing the outlier and associated SNPs (henceforth referred to as outlier and associated loci) with a $BF > 10$ were retrieved. For datasets 'ipr.centersnp_ib' and 'ipr.centersnp_ib_mor', this was done using the script *loci_consensus.py* (https://github.com/CoBiG2/RAD_Tools/blob/master/loci_consensus.py as of commit "d277ed3"). For each dataset, a *blastn* query was performed against the reference genome of *Q. lobata* [98]. This was done using *BLAST*+v.2.13.0 [99] via the R package *rBLAST* v.0.99.2 [100]. Blast results were filtered and, for each locus, only the match with the highest percentage of similarity was kept. The list of similar *Q. lobata* genes obtained for each dataset was compared in order to find common genes appearing as outliers or as associated with environmental variables.

Additionally, in a simple and exploratory way, enrichment analyses were performed for each dataset. The list of *Q. lobata* genes with a high degree of similarity to the outlier and associated loci found in our datasets was run through *g:Profiler* [101]. The organism was set as *Quercus lobata* and the remaining parameters were left as default. The results were plotted using R v.4.2.1.

Finally, the biological function of genes similar to the outlier loci and a subset of the associated loci was explored in order to find potential genes related to the adaptive process to heat and water stress.

Chapter 3

Results

3.1 Filtering and SNP calling

Paired-end read count before and after filtering

After the sequencing of the GBS libraries, we obtained a total of 293,145,449 paired-end reads, with an average of 3,118,569 paired-end reads per sample. Four samples showed exceptionally low coverage. Preliminary data filtering showed that this low coverage would later translate into high values of missing data after the filtering process. Therefore, the following samples were excluded from any further analyses: COR_686, GRE_111, SAR_617, and VAR_015 (Supplemental Table S1). After the removal of mtDNA and cpDNA reads, the nDNA dataset contained a total of 289,107,126 paired-end reads, with an average of 3,212,301 paired-end reads per sample (Supplementary Figure S2 (a)). After *ipyrad*'s quality filtering steps (steps 1 and 2), 263,590,360 paired-end reads were kept, with an average of 2,928,782 paired-end reads per sample. After *Stacks*' filtering step (*process_radtags* module) 198,500,534 paired-end reads were kept, with an average of 2,205,561 paired-end reads per sample (Supplementary Figure S2 (b)).

ipyrad and *Stacks* parameter testing results

The results for the *ipyrad* assemblies run with different *clustering_threshold* values can be found in Supplementary Table S2. Raising the clustering threshold above the default value of 0.85 led to a higher number of SNPs in both all-SNPs and center-SNP datasets. Moreover, it increased the missing data per individual (past the defined 65% threshold in some cases), which meant that samples would have to be removed. Because of this, the clustering threshold was maintained at 0.85, and the resulting assembly was kept and used in further analyses.

The results for *Stacks*' assemblies run with different parameter combinations can be found in Supplementary Table S3. Raising the minimum number of raw reads required to form an exactly matching stack (*-m*) above the default value of 3 seemed to drastically reduce the number of SNPs in the final datasets, especially in the random-SNP datasets. It also led to an increase in the amount of missing data per individual above the 65% threshold. Raising the number of mismatches allowed between stacks within individuals (*-M*) and the number of mismatches allowed between stacks between individuals (*-n*) above the default values (3 and 2, respectively), lead to an increase in the number of SNPs in the final datasets. This also slightly raised the amount of missing data per individual, although not above the 65 % threshold. Because of this, the parameter combination *-m* 3 *-M* 10 *-n* 10 was adopted and used in datasets for subsequent analyses.

Final *ipyrad* and *Stacks* datasets

The number of variable sites at the end of the SNP calling process for the final datasets can be found in Table 3.1. All datasets created for the population structure and genetic diversity analyses retained 90 samples. Datasets created for selection detection analyses retained 48 samples for the area of Iberia and 63 samples for the area of Iberia and Morocco.

Stacks consistently outputted more SNPs than *ipyrad*. Datasets 'ipr_allsnps_allsamp' and 'ipr_centersnp_allsamp' contained 19,652 and 1,921 SNPs, respectively (3.1). This corresponds to about 26% and 45% of the number of SNPs in datasets 'stk_allsnps_allsamp' and 'stk_centersnp_alsamp', respectively. This pattern was also observed for the datasets created for the outlier detection and genetic-environmental association analyses (Table 3.1 (b)). These also showed a

Table 3.1: Summary of final datasets used for (a) the population structure and genetic diversity analyses and (b) outlier detection and genetic-environmental association. Presented is the number of SNPs after *ipyrad/Stacks* SNP calling process (and *vcf_parser.py* for Center-SNP datasets), and the number of SNPs and samples in the final datasets, after *VCFTools* filtering.

(a)

Dataset	Nr of SNPs after SNP calling	Nr of SNPs after VCFTools filtering	Nr of samples
(i) ipr_allsnps_allamps	235,964	19,652	90
(ii) ipr_centersnp_allamps	22,686	1,921	90
(v) stk_allsnps_allamps	1,215,862	77,050	90
(vi) stk_centersnp_allamps	71,350	4,227	90

(b)

Dataset	Nr of SNPs after SNP calling	Nr of SNPs after VCFTools filtering	Nr of samples
(iii) ipr_centersnp_ib	19,316	2,981	48
(iv) ipr_centersnp_ib_mor	20,851	2,516	63
(vii) stk_centersnp_ib	74,908	4,872	48
(viii) stk_centersnp_ib_mor	78,726	4,736	63

higher number of SNPs when compared to the general datasets.

Regarding the use of the *vcf_parser.py* script instead of the *--write-random-snp* parameter, results showed it lead to an increase in the number of SNPs in the final dataset: 3,202 SNPs in the Random-SNP *-m 3 -M 10 -n 10* dataset created for parameter testing (Supplemental Table S3), versus 4,227 SNPs in the final Center-SNP dataset ('stk_centersnp_allamps'; Table 3.1).

3.2 Population Structure

Population structure results were consistent between *ipyrad*'s and *Stacks*' datasets ('ipr_allsnps_allamps' and 'stk_allsnps_allamps'), despite their considerable difference in the number of SNPs. Results point to the existence of three clear genetic populations among our samples. *fineRADstructure* seems to separate the samples into two main groups (Figures 3.1 A and 3.2 A): a group comprised of *Q. rotundifolia* samples from the Iberian Peninsula and Morocco, which further splits into separate Iberian and Moroccan groups (**I** and **II**, respectively), and a group comprised of *Q. ilex* samples from France, Italy, and Catalonia, and *Q. rotundifolia* samples from Tunisia (**III**).

Alstructure results corroborate these findings (Figures 3.1 B and 3.2 B): considering K2 populations, Iberian and Moroccan *Q. rotundifolia* samples seem to be assigned to one cluster, while *Q. ilex* samples from France, Italy, and Catalonia, and *Q. rotundifolia* samples from Tunisia seem to be assigned to another; considering K3 populations, *Q. rotundifolia* samples seem to further split into different Iberian and Moroccan clusters. Additionally, for K4 (*Stacks*) and K5 (*ipyrad*), Tunisian samples appear to form a cluster of their own. Full *ALStructure* results organized by sampling location can be found in Supplemental Figure S5.

The PCA results confirm the same pattern, separating samples into the same three clusters (Figure 3.3). PC1 explains the largest amount of variation among samples (5.8% and 5.38% for datasets 'ipr_centersnp_allamps' and 'stk_centersnp_allamps', respectively), and seems to separate the *Q. rotun-*

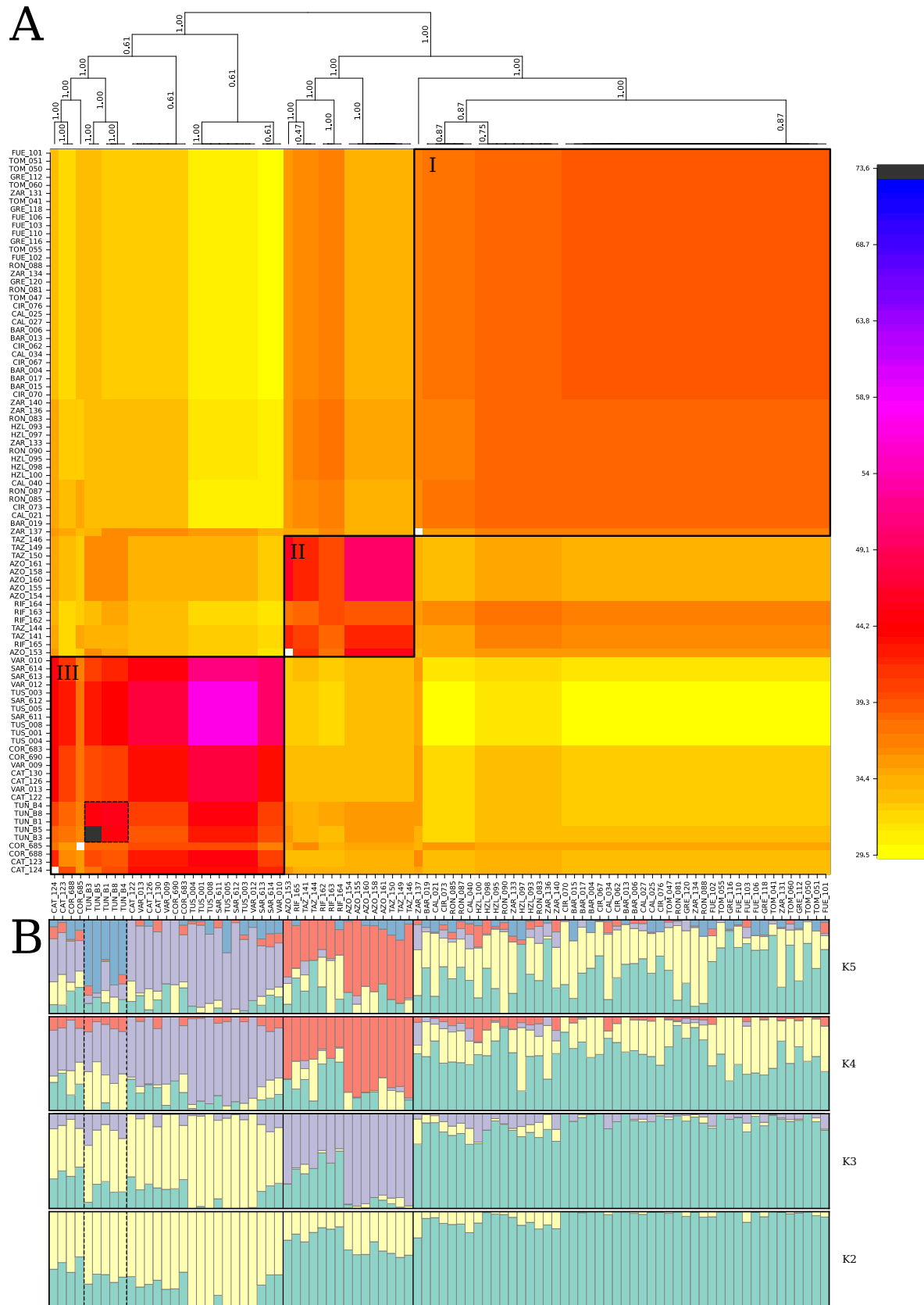


Figure 3.1: Population structure results for *ipyrad* datasets, highlighting the three genetic groups found. (A) *fineRADstructure* results for dataset 'ipr_allsnps_allsampls'. Genetic groups are identified in the following way: I - Iberian *Q. rotundifolia*; II - Moroccan *Q. rotundifolia*; and III - *Q. ilex*. (B) *ALStructure* results for dataset 'ipr_centersnp_allsampls', from K2 to K5.

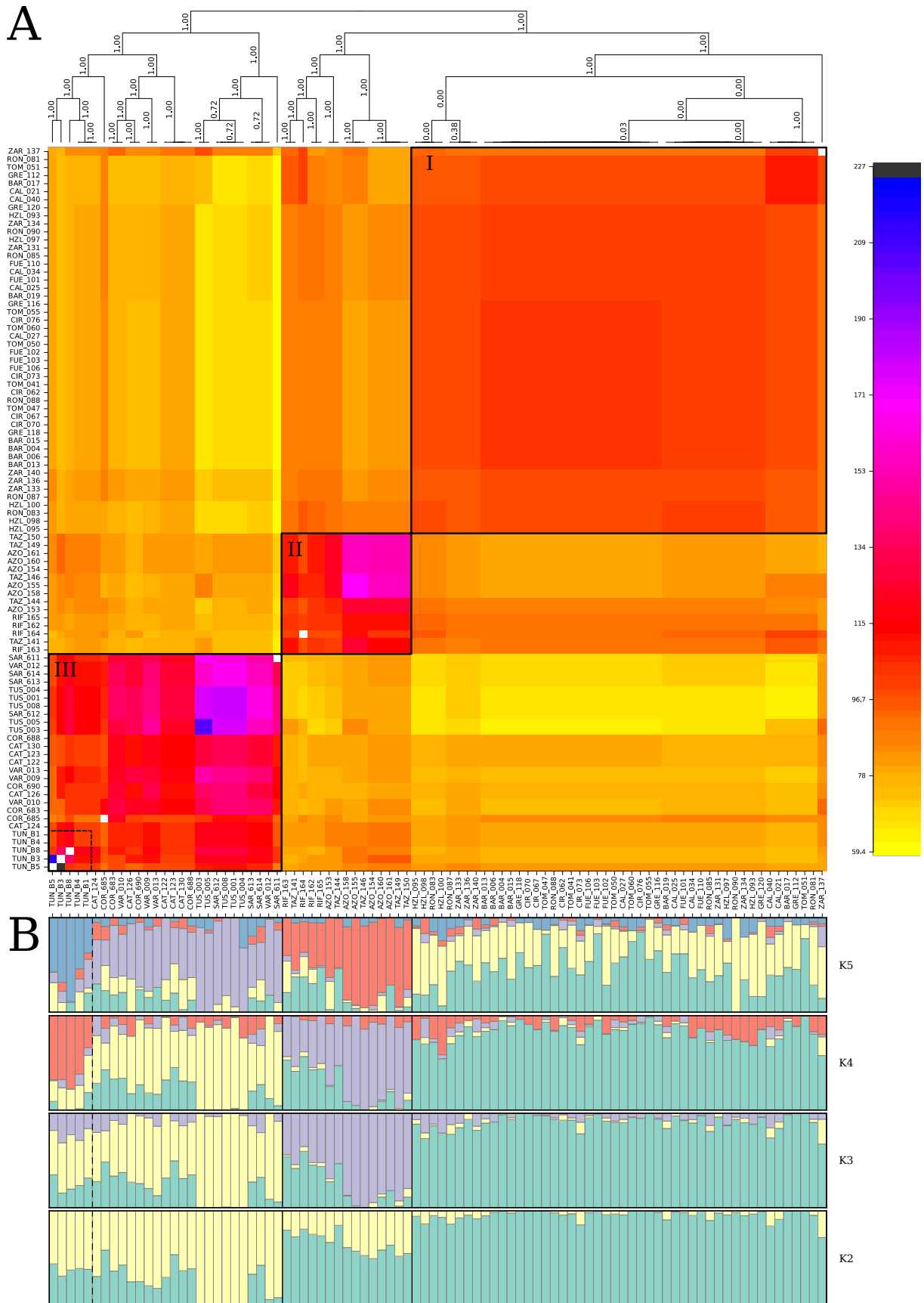


Figure 3.2: Population structure results for *Stacks* datasets, highlighting the three genetic groups found. (A) *fineRADstructure* results for dataset 'stk_allsnps_allsamp'. Genetic groups are identified in the following way: I - Iberian *Q. rotundifolia*; II - Moroccan *Q. rotundifolia*; and III - *Q. ilex*. (B) *ALStructure* results for dataset 'stk_centersnp_allsamp', from K2 to K5.

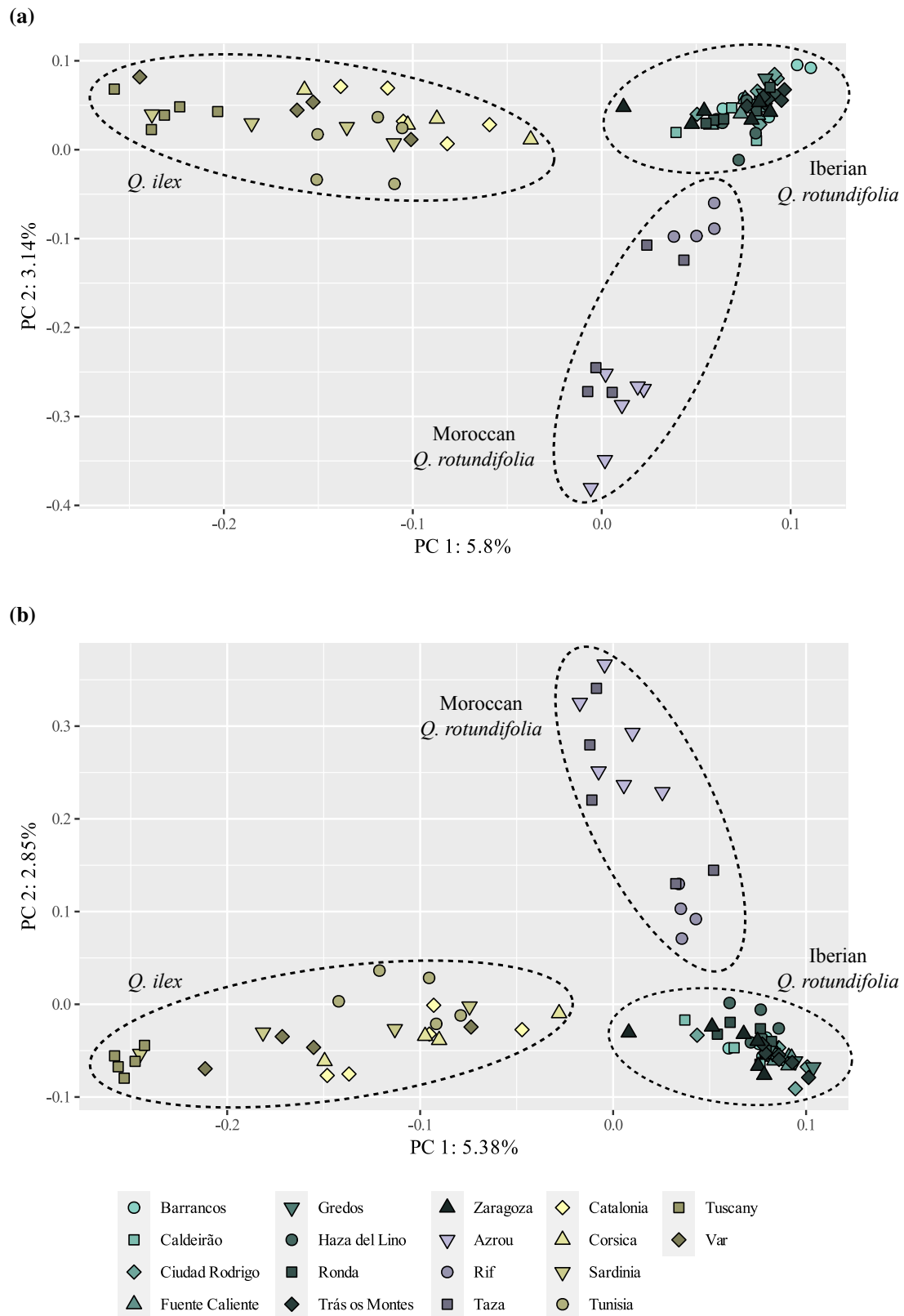


Figure 3.3: PCA results for datasets (a) 'ipr.centersnp.allsamp' and (b) 'stk.centersnp.allsamp'. Points represent the samples. Each sampling location is represented by a different point color/shape. Samples belonging to each of the genetic groups revealed in the *fineRADstructure* and *ALStructure* results are highlighted.

difolia samples (Iberia and Morocco) from the *Q. ilex* samples (France, Italy, Catalonia) and Tunisian *Q. rotundifolia* samples. PC2 explains slightly less variation among samples (3.14% and 2.85% for datasets 'ipr_centersnp_allsamp' and 'stk_centersnp_allsamp', respectively), and seems to separate the Moroccan *Q. rotundifolia* samples from the remaining.

Inside the Iberian population, individuals seem to be relatively homogeneous. This is visible in the *fineRADstructure* results, which show that in other populations there are groups of samples that show a higher degree of differentiation from the rest (Azrou - AZO and Taza - TAZ in Moroccan *Q. rotundifolia*, Tuscany - TUS and Sardinia - SAR in *Q. ilex*), which does not seem to happen in Iberian *Q. rotundifolia* (Figures 3.1 A and 3.2 A). This is especially evident in the PCA results, where samples from Iberia tend to cluster tightly together, while samples from the other populations appear to be spread along the axes (Figure 3.3). According to *ALStructure* results, the Iberian population shows the least amount of shared ancestry with the other groups. A small degree of shared ancestry with the Moroccan group is visible, mainly in samples closer to the southern limits of the Iberian Peninsula (Caldeirão - CAL, Ronda - RON and Haza del Lino - HZL) (Figures 3.1 B and 3.2 B).

As mentioned, inside the Moroccan population, some samples appear more differentiated. These are mainly samples from Azrou and Taza, which can be found deeper in the species distribution in that geographical area. This is evident in *fineRADstructure* results, and is especially visible for dataset 'stk_allsnps_allsamp' (Figures 3.1A and 3.2A). Furthermore, samples closer to Iberia (mainly from Rif - RIF), and a few samples from Taza and Azrou, show a considerable amount of shared ancestry with the Iberian population, visible in *ALStructure* results (Figures 3.1B and 3.2B). *fineRADstructure* results also point to the higher similarity between these samples and those of Iberia. PCA results corroborate this, showing a group of Moroccan samples clustering closer to the Iberian samples (mainly from Rif, with some samples from Taza) and a more isolated group of samples (Azrou and Taza). The pattern of shared ancestry in this area (southern Iberia - northern Morocco) seems to be asymmetrical: Moroccan samples show a much higher proportion of their genome assigned to the Iberian cluster than vice-versa.

As with Moroccan *Q. rotundifolia*, some samples in the *Q. ilex* group appear to be more differentiated. This is evident in *fineRADstructure* results, where samples from Tuscany and some from Sardinia seem to form a more differentiated group (Figures 3.1A and 3.2A). In general, these samples also show the least amount of shared ancestry with other populations, as evidenced by *ALStructure* results (Figures 3.1B and 3.2B). The proportion of the genome assigned to other clusters seems to be relatively higher in the remaining samples in the *Q. ilex* population, mainly from the Iberian cluster in individuals from Catalonia - CAT, Var - VAR, Corsica - COR, Sardinia, and Tunisia, and from the Moroccan subgroup in individuals from Catalonia, Corsica, and especially Tunisia. Once again, PCA results corroborate this: samples from the *Q. ilex* population appear spread along PC1, with samples from Tuscany and some from Sardinia clustering farther away from *Q. rotundifolia* samples (Figure 3.3). A slight similarity of Tunisian samples to those of the Moroccan population is also apparent in *fineRADstructure* and PCA results. In the latter, it's visible that some Tunisian samples appear to cluster slightly closer to Moroccan samples along PC2.

3.3 Genetic diversity and differentiation

Observed and expected heterozygosity were calculated for each population and overall for datasets 'ipr_centersnp_allsamp' and 'stk_centersnp_allsamp'. For *ipyrad*'s dataset, Ho and He values were, respectively, 0.093 and 0.110 for the Iberian *Q. rotundifolia* population, 0.096 and 0.112 for the Moroccan *Q. rotundifolia* population, 0.095 and 0.115 for the *Q. ilex* population and 0.095 and 0.120 overall.

Table 3.2: Summary statistics, for each population and overall, and global F_{st} values calculated for datasets (a) 'ipr.centersnp_allamps' and (b) 'stk.centersnp_allamps'. Represented is the number of individuals (N), number of alleles in each population (N_A), allelic richness A_R , observed heterozygosity (H_o), expected heterozygosity (H_e), inbreeding coefficient (F_{is}), F_{st}/F_{st}^{max} for each population, and overall F_{st} .

(a)

	N	N_A	A_R	H_o	H_e	F_{is}	F_{st}/F_{st}^{max}
<i>Q. rotundifolia</i> Iberia	48	3431	1.420	0.093	0.110	0.149	0.079
<i>Q. rotundifolia</i> Morocco	15	2972	1.417	0.096	0.112	0.111	0.242
<i>Q. ilex</i>	27	3096	1.413	0.095	0.115	0.135	0.131
Global	90	–	1.417	0.095	0.120	0.155	F_{st} : 0.122

(b)

	N	N_A	A_R	H_o	H_e	F_{is}	F_{st}/F_{st}^{max}
<i>Q. rotundifolia</i> Iberia	48	7488	1.364	0.084	0.100	0.146	0.066
<i>Q. rotundifolia</i> Morocco	15	6364	1.359	0.084	0.102	0.138	0.218
<i>Q. ilex</i>	27	6723	1.360	0.080	0.103	0.167	0.156
Global	90	–	1.361	0.083	0.108	0.179	F_{st} : 0.123

Table 3.3: Pairwise F_{st} values calculated for datasets (a) 'ipr.centersnp_allamps' and (b) 'stk.centersnp_allamps'. F_{st} values were calculated pairwise considering two populations (*Q. rotundifolia* and *Q. ilex*) and three populations (Iberian *Q. rotundifolia* (*Q. rotund. Ib.*), Moroccan *Q. rotundifolia* (*Q. rotund. Mor.*), and *Q. ilex*).

(a)

	<i>Q. ilex</i>	<i>Q. rotund. Ib.</i>
<i>Q. rotund. Mor.</i>	0.123	0.077
<i>Q. rotund. Ib.</i>	0.148	–
<i>Q. rotundifolia</i>	0.129	–

(b)

	<i>Q. ilex</i>	<i>Q. rotund. Ib.</i>
<i>Q. rotund. Mor.</i>	0.113	0.080
<i>Q. rotund. Ib.</i>	0.151	–
<i>Q. rotundifolia</i>	0.128	–

For *Stacks*' dataset, H_o and H_e values were, respectively, 0.084 and 0.100 for the Iberian population, 0.084 and 0.102 for the Moroccan population, 0.081 and 0.103 for the *Q. ilex* population and 0.083 and 0.108 overall. Values are similar between both datasets, with *Stacks*' returning slightly lower values.

F_{is} values were calculated for each population and overall for datasets 'ipr.centersnp_allamps' and 'stk.centersnp_allamps'. Values are in line with what was observed in terms of expected and observed heterozygosity. Populations with the lowest H_o values show the highest F_{is} values (Iberian population for 'ipr.centersnp_allamps' and *Q. ilex* for 'stk.centersnp_allamps'). In turn, the lowest F_{is} values were 0.111 and 0.138 for the Moroccan population (in 'ipr.centersnp_allamps' and 'stk.centersnp_allamps', respectively). This same population shows the highest H_o value of 0.09600442 and 0.08410277 (in 'ipr.centersnp_allamps' and 'stk.centersnp_allamps', respectively). Values are similar between datasets, with F_{is} values tending to be higher in *Stacks*' dataset. Inside and between datasets, the difference between populations does not seem considerable. Allelic richness was calculated for each population and overall for both datasets. Values are similar for all populations inside both datasets, being slightly lower for *Stacks* dataset.

F_{st}/F_{st}^{max} was calculated for each population. Values corroborate the population structure results. The Iberian population, which population structure results showed to be more homogeneous than the others, shows the lowest value of F_{st}/F_{st}^{max} (0.079 and 0.066, for 'ipr.centersnp_allamps' and

'stk_centersnp_allamps', respectively). The remaining populations show higher values, especially the Moroccan population (0.242 and 0.218 for the Moroccan population, 0.131 and 0.156 for *Q. ilex*, for 'ipr_centersnp_allamps' and 'stk_centersnp_allamps', respectively). Pairwise comparisons using a Kruskal-Wallis rank sum test and a Wilcoxon rank sum test on generated bootstrap distributions showed the difference between the populations' F_{st}/F_{st}^{max} distributions to be statistically significant ($p < 2e-16$).

The overall F_{st} was 0.122 and 0.123 for datasets 'ipr_centersnp_allamps' and 'stk_centersnp_allamps', respectively (Table 3.2). Pairwise F_{st} values are consistent with the observed population structure (Table 3.3). *Q. rotundifolia* and *Q. ilex* populations showed pairwise F_{st} values of 0.129 and 0.128 for 'ipr_centersnp_allamps' and 'stk_centersnp_allamps', respectively. Considering the Iberian and Moroccan *Q. rotundifolia* populations, the pairwise F_{st} values between these and *Q. ilex* were 0.148 and 0.151 for Iberia vs. *Q. ilex*, and 0.123 and 0.113 for Morocco vs. *Q. ilex*, for 'ipr_centersnp_allamps' and 'stk_centersnp_allamps', respectively. Within *Q. rotundifolia*, Iberia and Morocco showed pairwise F_{st} values of 0.077 and =0.080 for 'ipr_centersnp_allamps' and 'stk_centersnp_allamps', respectively. Pairwise F_{st} values calculated between every sampling location for datasets 'ipr_centersnp_allamps' and 'stk_centersnp_allamps' can be found in Supplemental Table S5. Pairwise F_{st} values tend to be relatively low between samples from the same genetic population. Inside Moroccan *Q. rotundifolia*, samples from the Rif share lower values of pairwise F_{st} with Iberian sampling locations than samples from Taza or Azrou. Tunisian samples share higher pairwise F_{st} values with Iberian *Q. rotundifolia* sampling locations than with Moroccan sampling locations. The latter are only slightly higher than the values shared between Tunisian samples and other *Q. ilex* sampling locations.

3.4 Outlier detection

BayPass, under the core model mode, identified 4 outlier SNPs for 'ipr_centersnp_ib' (0.1% of total), 16 outliers for 'ipr_centersnp_ib_mor' (0.6% of total), 6 outliers for 'stk_centersnp_ib' (0.1% of total) and 34 outliers for 'stk_centersnp_in_mor' (0.7% of total) (Table 3.4). In general, datasets containing only Iberian samples returned fewer outliers than those containing Iberian and Moroccan samples. Major allele frequencies for the outlier SNPs for datasets 'ipr_centersnp_ib_mor' and 'stk_centersnp_ib' are plotted in Supplemental Figure S7 and for datasets 'ipr_centersnp_ib_mor' and 'stk_centersnp_ib_mor' in Figure 3.4. For 'ipr_centersnp_ib_mor' and 'stk_centersnp_ib_mor', most outlier loci seem to follow the same pattern of allele frequency distribution, with the major allele showing a higher frequency or being fixed in one of the populations (Iberian or Moroccan) while showing lower frequency in the other.

3.5 Genetic-environmental association

3.5.1 Associated SNPs and bioclimatic variables

The genetic-environmental associations identified by *BayPass* under the AUX covariate mode for datasets 'ipr_centersnp_ib', 'ipr_centersnp_ib_mor', 'stk_centersnp_ib' and 'stk_centersnp_ib_mor' are expressed in Table 3.5. The number of associated SNPs in each dataset is expressed in Table 3.4 (b). In general, *Ipyrad*'s datasets ('ipr_centersnp_ib', 'ipr_centersnp_ib_mor') returned fewer associations than their *Stacks* counterparts ('stk_centersnp_ib', 'stk_centersnp_ib_mor').

Datasets for the Iberia region ('ipr_centersnp_ib' and 'stk_centersnp_ib') returned by far the most associations with $BF > 10$. 'ipr_centersnp_ib' returned a total of 1000 associations, with 853 associated

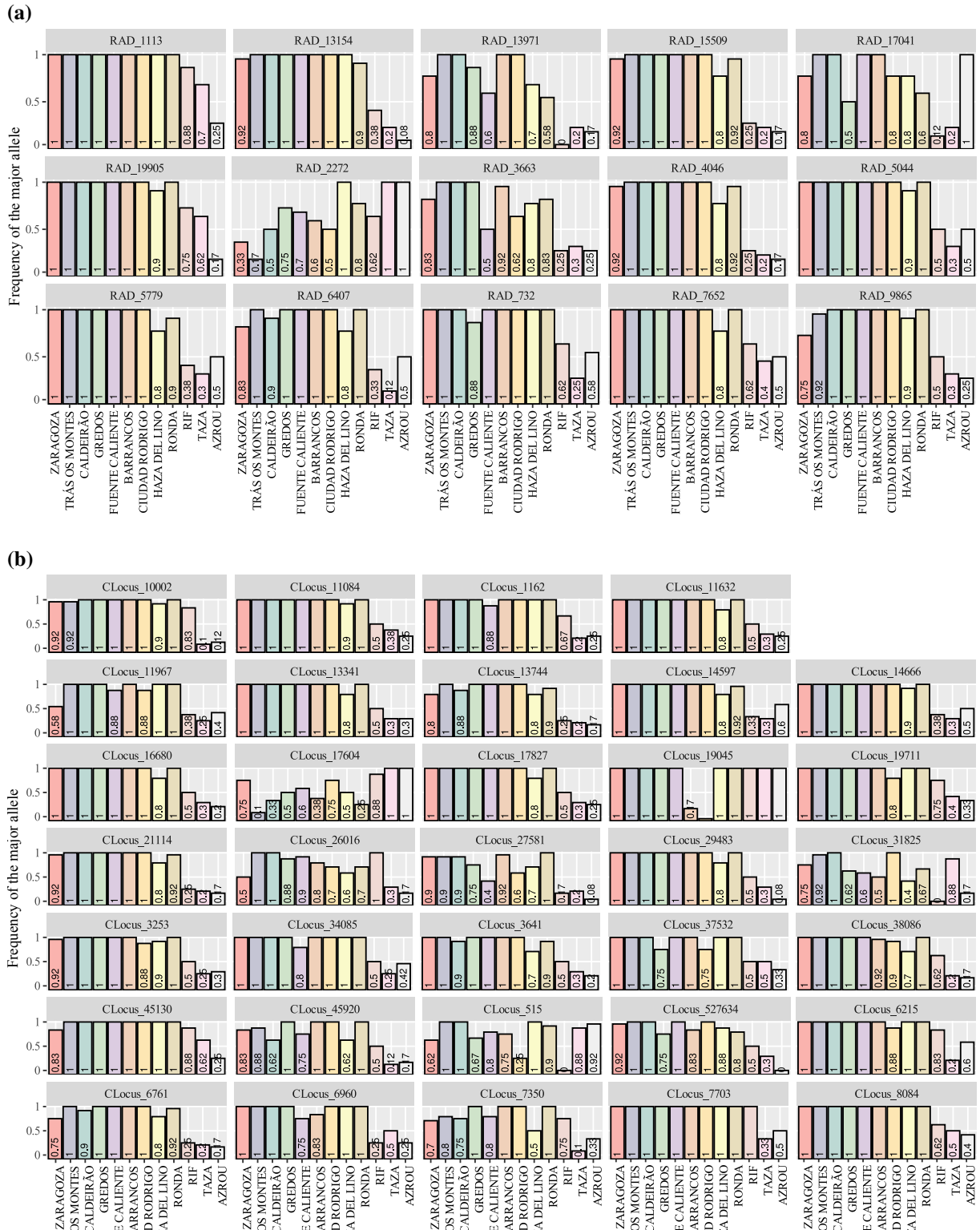


Figure 3.4: Frequency of the major allele for the outlier SNPs of datasets (a) 'ipr.centersnp_ib_mor' and (b) 'stk.centersnp_ib_mor', calculated for each *Q. rotundifolia* sampling location.

Table 3.4: Summary of results for (a) the outlier detection and (b) genetic-environmental association analyses, and subsequent annotation process, for datasets 'ipr.centersnp_ib', 'ipr.centersnp_ib_mor', 'stk.centersnp_ib' and 'stk.centersnp_ib_mor'. Columns show the number of outlier and associated SNPs/loci in each dataset, the total number of sequence matches against the reference genome of *Q. lobata*, the number of unique loci in the dataset with at least a match, and the number of unique genes matched in *Q. lobata*'s reference genome.

(a)

Dataset	Outlier SNPs/loci	Reference genome matches	Matched dataset loci	Matched reference genome genes
(iii) ipr.centersnp_ib	4	0	–	–
(vii) stk.centersnp_ib	6	0	–	–
(iv) ipr.centersnp_ib_mor	16	2	2	2
(viii) stk.centersnp_ib_mor	34	3	3	3

(b)

Dataset	Associated SNPs/loci	Reference genome matches	Matched dataset loci	Matched reference genome genes
(iii) ipr.centersnp_ib	853	66	59	64
(vii) stk.centersnp_ib	2337	333	306	282
(iv) ipr.centersnp_ib_mor	136	7	6	7
(viii) stk.centersnp_ib_mor	545	94	79	84

SNPs and 5 associated variables: bio4 (temperature seasonality), bio8 (mean daily mean air temperatures of the wettest quarter), bio9 (mean daily mean air temperatures of the driest quarter), bio15 (precipitation seasonality) and bio17 (mean monthly precipitation amount of the driest quarter). 'stk.centersnp_ib' returned a total of 3321 associations, with 2337 associated SNPs and 8 associated variables: bio1 (mean annual air temperature), bio2 (mean diurnal air temperature range), bio3 (isothermality), bio4 (temperature seasonality), bio9 (mean daily mean air temperatures of the driest quarter), bio12 (mean annual precipitation amount), bio15 (precipitation seasonality) and bio17 (mean monthly precipitation amount of the driest quarter).

Datasets for the Iberia and Morocco regions ('ipr.centersnp_ib_mor' and 'stk.centersnp_ib_mor') returned far fewer associations with $BF > 10$ (13% of 'ipr.centersnp_ib' and 17% of 'stk.centersnp_ib', respectively). 'ipr.centersnp_ib_mor' returned 139 associations, with 136 associated SNPs and 3 associated variables: bio12 (annual precipitation amount), bio15 (precipitation seasonality) and bio17 (mean monthly precipitation amount of the driest quarter). 'stk.centersnp_ib_mor' returned 579 associations, with 545 associated SNPs and 7 associated variables: bio1 (mean annual air temperature), bio3 (isothermality), bio4 (temperature seasonality), bio6 (mean daily minimum air temperature of the coldest month), bio10 (mean daily mean air temperatures of the warmest quarter), bio12 (annual precipitation amount) and bio17 (mean monthly precipitation amount of the driest quarter).

Every variable appeared associated with at least one dataset except for bio7 (which was only included in the analyses involving the datasets 'ipr.centersnp_ib_mor' and 'stk.centersnp_ib_mor'). Only one variable, bio17, was consistently found as associated in all datasets, showing decisive evidence of association in all of them. The proportion of loci associated with bio17 was much higher for the datasets containing Iberian and Moroccan samples (Supplemental Figure S8a). bio12 was found as associated in all datasets

Table 3.5: Genetic-environmental association results for datasets (a) 'ipr_centersnp_ib', (b) 'ipr_centersnp_ib_mor', (c) 'stk_centersnp_ib' and (d) 'stk_centersnp_ib_mor', analyzed under *BayPass* v.2.31's auxiliary covariate mode. Columns show the number of associations, the number of associated SNPs, and their percentage in relation to the total number of SNPs in the dataset (Table 3.1), and the associated variables for the different levels of association strength and in total. Association strength is defined by the BF value, based on Jeffreys' rule according to Gautier et al. (2015) [76].

(a) 'ipr_centersnp_ib'

Strength of association	Number of associations	Number of associated SNPs	Associated variables
Strong ($10 \leq BF < 15$)	664	596 (20%)	bio4, bio8, bio9, bio15, bio17
Very strong ($15 \leq BF < 20$)	254	244 (8%)	bio4, bio8, bio9, bio15, bio17
Decisive ($BF \geq 20$)	82	80 (3%)	bio4, bio8, bio9, bio15, bio17
TOTAL	1000	853	

(b) 'ipr_centersnp_ib_mor'

Strength of association	Number of associations	Number of associated SNPs	Associated variables
Strong ($10 \leq BF < 15$)	86	85 (3%)	bio12, bio15, bio17
Very strong ($15 \leq BF < 20$)	45	45 (2%)	bio12, bio15, bio17
Decisive ($BF \geq 20$)	8	10 (0.4%)	bio12, bio17
TOTAL	139	136	

(c) 'stk_centersnp_ib'

Strength of association	Number of associations	Number of associated SNPs	Associated variables
Strong ($10 \leq BF < 15$)	2095	1637 (34%)	bio1, bio2, bio3, bio4, bio9, bio12, bio15, bio17
Very strong ($15 \leq BF < 20$)	880	799 (16%)	bio1, bio2, bio3, bio4, bio9, bio12, bio15, bio17
Decisive ($BF \geq 20$)	346	325 (7%)	bio1, bio2, bio3, bio4, bio9, bio12, bio15, bio17
TOTAL	3321	2337	

(d) 'stk_centersnp_ib_mor'

Strength of association	Number of associations	Number of associated SNPs	Associated variables
Strong ($10 \leq BF < 15$)	376	362 (8%)	bio1, bio3, bio4, bio6, bio10, bio12, bio17
Very strong ($15 \leq BF < 20$)	143	140 (3%)	bio1, bio3, bio10, bio12, bio17
Decisive ($BF \geq 20$)	60	60 (1%)	bio1, bio12, bio17
TOTAL	579	545	

except 'ipr_centersnp_ib', making up a reasonable proportion of associated variables in those datasets (Supplemental Figure S8a). bio15 appeared as associated in all datasets except 'stk_centersnp_ib_mor'. bio6 appeared as associated to only two loci, both in the 'stk_centersnp_ib_mor' dataset.

3.5.2 Allele frequency analysis

Allele frequencies of associated SNPs were plotted along the respective associated bioclimatic variable as a way to visualize the pattern of allele turnover along the environmental gradients. This was done first to the 3 individual SNPs with the highest strength of association for each dataset, restricted to the variables common to datasets 'ipr_centersnp_ib' and 'stk_centersnp_ib' (bio4, bio9, bio15, and bio17) and to datasets 'ipr_centersnp_ib_mor' and 'stk_centersnp_ib_mor' (bio12 and bio17) (Figures 3.5 and 3.6, Supplemental Figures S11, S12, S13, S13, S14S10, S11, S12, and S13). These results showed that, for most plotted SNPs, the allele frequency turnover was not linear. Instead, there was a "turnover threshold", a value or range of values where the change in allele frequency tended to occur, and allele frequencies appeared fixed either below or above that threshold.

SNPs were then categorized into two groups depending on their pattern of allele turnover along the environmental variable they were associated with, based on if the allele frequencies tended to be fixed above or below the "turnover threshold". For the two groups of SNPs, the average allele frequency was calculated for each pair of dataset/bioclimatic variable and plotted along the environmental variable. These results reveal a similar pattern to that observed for the individual SNPs, where there is a tendency for alleles to be fixed either above or below a certain value or range of values, wherein most of the allele frequency turnover tends to occur. This pattern appears consistent when using every SNP, regardless of level of association strength (Figure 3.7 for bioclimatic variables bio4, bio9, bio15, bio12, and bio17; Supplemental Figure S9 for all associated variables), and when using only SNPs in the decisive level of association strength (S10).

Common variables in the Iberia region

The results for common associated variables in the Iberian region datasets, show the following:

Temperature seasonality (bio4): Seasonality seems to increase towards the center of the Iberian Peninsula, being higher in locations like Gredos and Fuente Caliente, and lower towards the southernmost locations, especially closer to the coast, as seen in Caldeirão, Haza del Lino, and Ronda (Figure 3.5). In terms of allele frequency turnover, the threshold seems to be around the 550 - 600 °C/100 mark (as observed in Figure 3.7 a)), with some loci being fixed above this temperature (RAD_16767, CLocus_7016, and CLocus_758) and others being fixed below (RAD_18847, RAD_5717, and CLocus_644896)(Figure 3.5).

Mean daily mean air temperatures of the driest quarter (bio9): Temperatures seem to be similar in most locations, ranging between 20-25 °C. Zaragoza seems to be an outlier, showing a much lower temperature value (Supplemental Figure S11). In terms of allele frequency turnover, the threshold seems to be located between 20-22.5 °C (Figure 3.7 a)). Some loci appear clearly fixed for higher temperatures (RAD_14198 and RAD_5462). The opposite pattern, loci fixed for lower temperatures, is not so clear: for CLocus_33447, CLocus_34093, CLocus_57127, and RAD_11361, the major allele is always fixed in Gredos, at lower temperatures, but tends to reduce its frequency in some sampling locations at higher temperatures (Supplemental Figure S11).

Precipitation seasonality (bio15): Seasonality seems to increase from northeast to southwest, being the lowest in Zaragoza and highest towards the southern coast of the peninsula (Supplemental Figure

S12). In terms of allele frequency turnover, the threshold seems to be around 55-60 °C (Figure 3.7 a)), with the major allele either fixed above (RAD_4895, RAD_2598, CLocus_48084, and CLocus_6370) or below this value of precipitation seasonality (RAD_14709 and CLocus_44371) (Supplemental Figure S12).

Mean monthly precipitation amount of the driest quarter (bio17): Precipitation seems to decrease from north to south, with Ciudad Rodrigo, Trás-os-Montes, and Zaragoza having the highest precipitation values for the driest quarter, and locations towards the southern coast having the lowest (Supplemental Figure S13). In terms of allele frequency turnover, the threshold appears roughly around 30-50 kg m⁻² (Figure 3.7 a)), with some loci fixed below this value (CLocus_24976, CLocus_18822, and CLocus_468842) and others fixed above (RAD_3752, RAD_18822, RAD_468842) (Supplemental Figure S13).

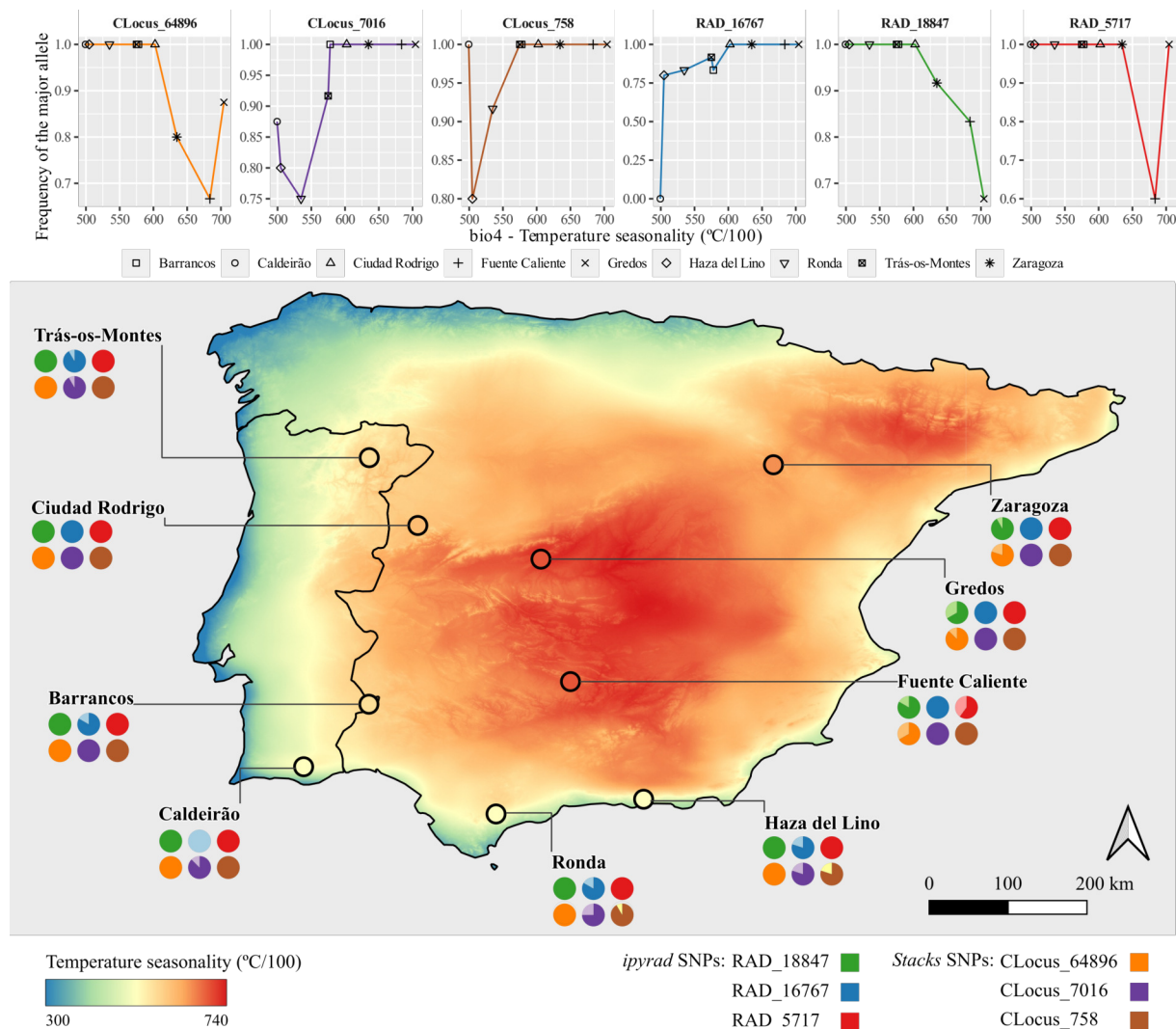


Figure 3.5: Representation of allele frequency turnover for loci associated with bioclimatic variable bio4 (temperature seasonality). The three strongest associated loci are shown for datasets ipr_centersnp_ib and stk_centersnp_ib. Allele frequencies for each sampling location are plotted as a function of temperature seasonality values (above). The scale of the yy axis may differ across plots. Allele frequencies for each sampling location are represented along the geographical area of the Iberian Peninsula (below). The range of temperature seasonality values is represented by a color scale. Values for each sampling location are represented above their geographical position.

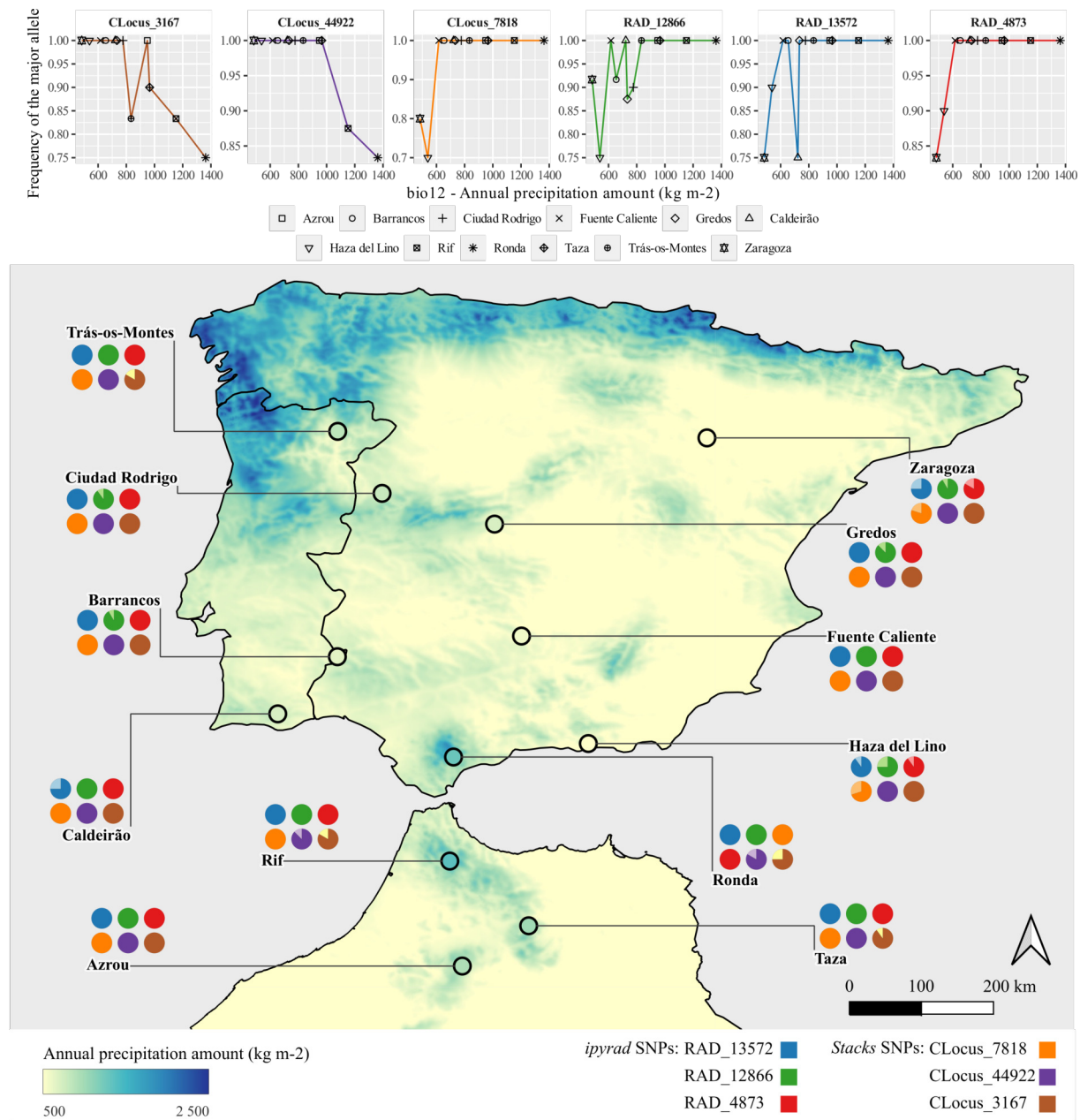
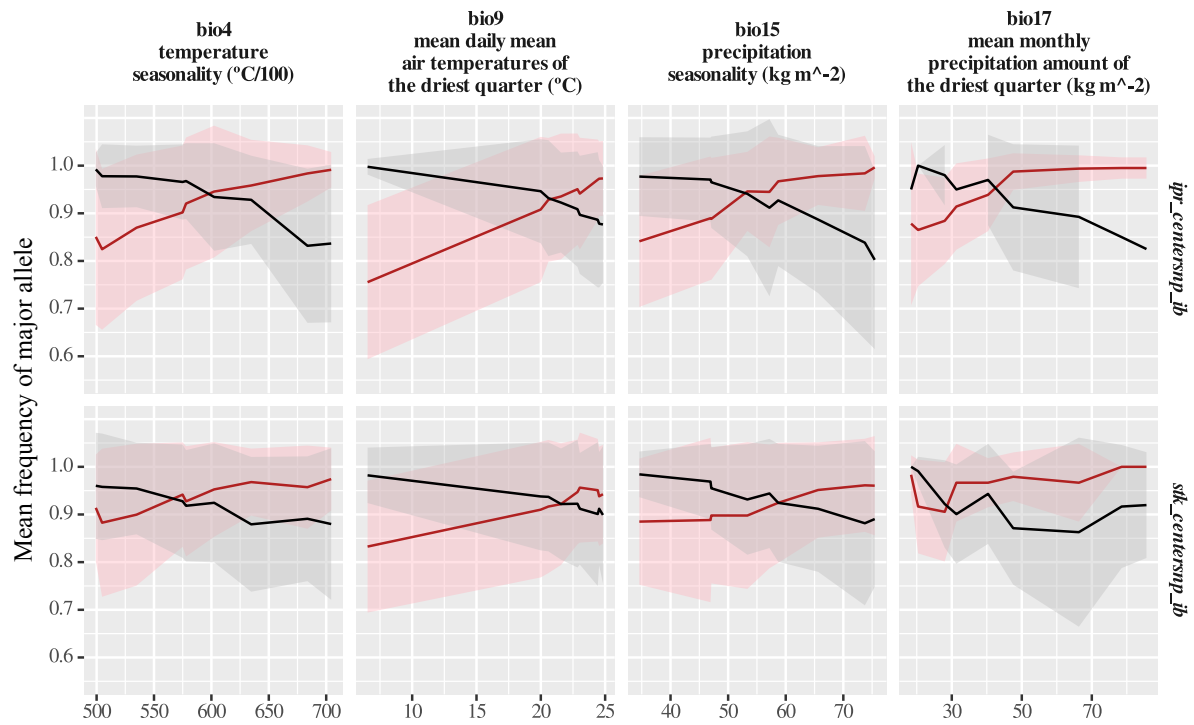


Figure 3.6: Representation of allele frequency turnover for loci associated with variable bio12 (annual precipitation amount). The three strongest associated loci are shown for datasets *ipr_centersnp_ib_mor* and *stk_centersnp_ib_mor*. Allele frequencies for each sampling location are plotted as a function of annual precipitation values (above). The scale of the yy axis may differ across plots. Allele frequencies for each sampling location are represented along the geographical area of the Iberian Peninsula and northern Morocco (below). The range of annual precipitation is represented by a color scale. Values for each sampling location are represented above their geographical position.

(a) Iberia



(b) Iberia and Morocco

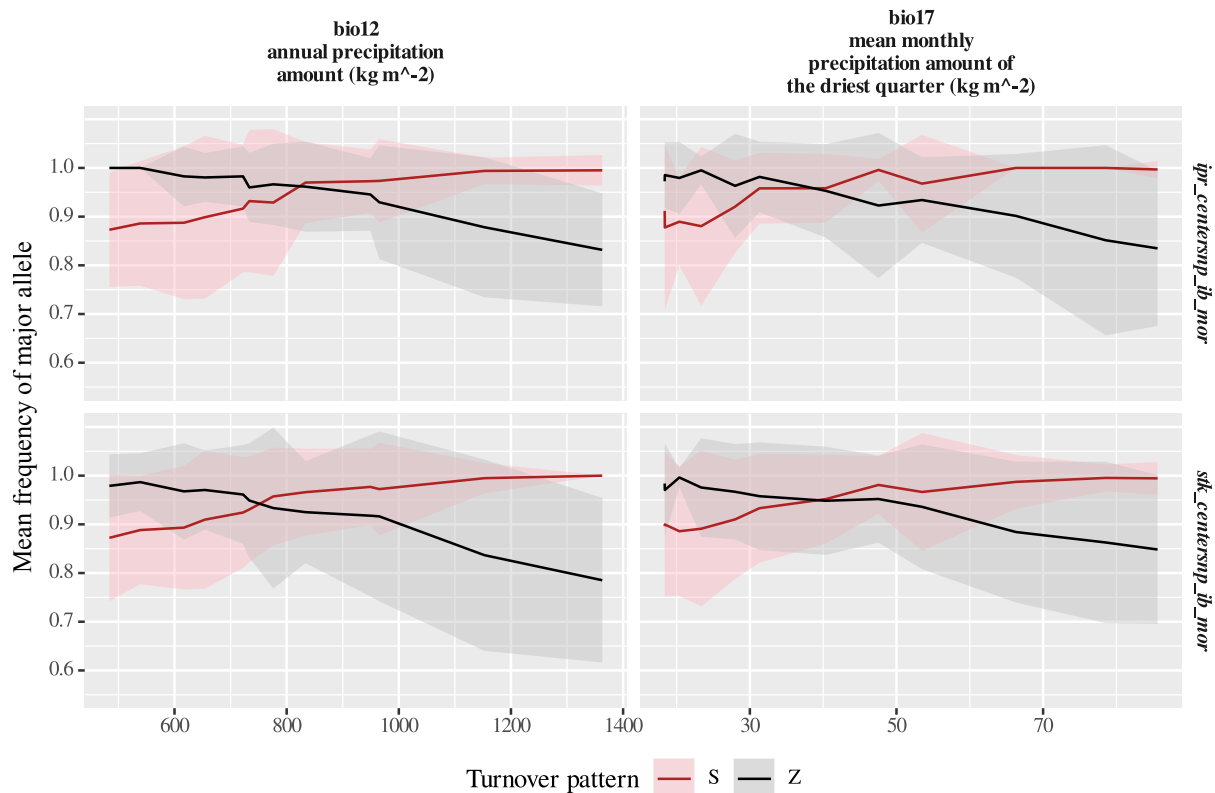


Figure 3.7: Average pattern of allele turnover along the associated bioclimatic variables common (a) to datasets *ipr_centersnp_ib* and *stk_centersnp_ib*, and (b) to datasets *ipr_centersnp_ib_mor* and *stk_centersnp_ib_mor*. Each SNP associated with an environmental variable, for all levels of association strength, was classified according to its pattern of allele turnover (upward pattern, with a tendency of allele frequencies to be fixed above turnover threshold; downward pattern, with a tendency of allele frequencies to be fixed below turnover threshold). The average allele frequency of the major alleles was calculated for SNPs in each turnover pattern, for each sampling location/environmental value. Colored lines represent the average allele frequencies along the bioclimatic variable. Standard deviation is represented by the colored shadow.

Common variables in the Iberia and Morocco regions

For the common associated variables in the Iberian and Moroccan region datasets, the following is observable:

Annual precipitation amount (bio12): Besides Ronda, Moroccan locations display the highest amounts of annual precipitation. Most Iberian locations display lower levels of annual precipitation, especially Zaragoza and Haza del Lino. In terms of allele frequency turnover, the threshold appears to be around 800-1000 kg m⁻², although, for a few loci, it appears to be lower. Some loci appear to be fixed below this threshold (CLocus_44922 and CLocus_3167), while others seem to be fixed above (RAD_13572, RAD_12866, RAD_4873, and CLocus_7818) (Figure 3.6).

Mean monthly precipitation amount of the driest quarter (bio17): As noted before, precipitation seems to decrease from north to south in the Iberian Peninsula. In Morocco, it seems to decrease from south to north, with Rif and Caldeirão having the lowest amount of precipitation during this period. In terms of allele frequency turnover, as observed for the Iberian sample datasets, the threshold appears roughly around the 50 kg m⁻² mark, although not every allele shows this clearly. Some loci appear fixed below this value (RAD_16369 and CLocus_9612) and others fixed above (RAD_21311, RAD_8811, CLocus_18664, and CLocus_2034) (Supplemental Figure S14).

3.6 Loci annotation and enrichment analyses

Reference genome sequence hits

For the outlier loci annotation, of the four datasets queried against the reference genome of *Q. lobata*, only 'ipr.centersnp_ib_mor' and 'stk.centersnp_ib_mor' returned sequence hits, with 2 and 3 similar *Q. lobata* genes found, respectively (Table 3.4 (a)). The outlier SNPs present in the matched loci were RAD_3663 and RAD_19905 for 'ipr.centersnp_ib_mor' and CLocus_8084, CLocus_11623, and CLocus_7350 for 'stk.centersnp_ib_mor' (Table 3.4). The name of the SNPs and the similar *Q. lobata* gene is expressed in Supplementary Table S6.

For the associated loci annotation, all datasets returned sequence hits against the reference genome, with 64, 282, 7, and 84 similar *Q. lobata* genes found for 'ipr.centersnp_ib', 'stk.centersnp_ib', 'ipr.centersnp_ib_mor' and 'stk.centersnp_ib_mor', respectively (Table 3.4 (b)). The number of matches was relatively low for every dataset, and the percentage of loci in the outlier/associated datasets that matched at least a gene in the reference genome ranged from 14.9% for 'stk.centersnp_ib_mor' associated loci to 4.4% for 'ipr.centersnp_ib_mor' associated loci.

Gene overlap between datasets

There was no overlap of genes obtained for *ipyrad* datasets (ipr.centersnp_ib and 'ipr.centersnp_ib_mor'), be it for outlier or associated loci (Supplemental Figure S16 (a)). For *Stacks* datasets, 'stk.centersnp_ib' associated loci showed 2 genes in common with the 'stk.centersnp_ib_mor' outlier loci and 50 genes in common with the 'stk.centersnp_ib_mor' associated loci (Supplemental Figure S16 (b) and (c)). A comparison of genes obtained for all datasets for the associated loci annotation showed: 10 common genes between ipr.centersnp_ib' and 'stk.centersnp_ib', one common gene between ipr.centersnp_ib_mor' and 'stk.centersnp_ib', one common gene between 'ipr.centersnp_ib' and 'stk.centersnp_ib_mor' (Supplemental Figure S16 (c)).

Exploration of potential role in heat and water stress response

The 50 *Q. lobata* genes found to be common to datasets 'stk_centersnp_ib' and 'stk_centersnp_ib_mor' were explored in terms of biological function. Eight genes showed a potentially interesting biological function in terms of response to abiotic stressors such as heat or water stress: 'F-box domain-containing protein', 'Zeaxanthin epoxidase, chloroplastic', 'TTF-type domain-containing protein', 'SBP-type domain-containing protein', 'SBP-type domain-containing protein', 'non-specific serine/threonine protein kinase', 'C2 domain-containing protein', 'Xanthine dehydrogenase 1-like', and 'BHLH domain-containing protein'. These *Q. lobata* genes and the respective associated SNPs / bioclimatic variables, as well as their putative role in heat and water stress response, are described in Supplementary Table S7

Enrichment analyses

Regarding the enrichment analyses with *g:Profiler*, only the following cases obtained results: for the genes resulting from the outlier detection, only 'stk_centersnp_ib_mor' returned results; for the genes resulting from the genetic-environmental association, only 'ipr_centersnp_ib', 'stk_centersnp_ib' and 'stk_centersnp_ib_mor'. Only three datasets wielded results for the enrichment analyses with *g:Profiler*: 'ipr_centersnp_ib' and 'stk_centersnp_ib' for the associated loci, and 'stk_centersnp_ib_mor' for both outlier and associated loci. The plotted results for the enrichment analyses can be found in Supplemental Figure S17

Chapter 4

Discussion

4.1 Observed differences between SNP calling programs

Filtering and SNP calling process

Two SNP calling software programs, *ipyrad* and *Stacks*, were used to ensure our findings were consistent and would not be skewed by the use of a single software. The first difference observed was the number of kept paired-end reads after the filtering steps (*ipyrad* steps 1 and 2; *Stacks*' *process_radtags* module). *Stacks* kept 198,500,534 paired-end reads while *ipyrad* kept 263,590,360. This 25% difference between the two software programs can mainly be attributed to the way they deal with the removal of the adapter sequences. While *ipyrad* is able to work with different-sized reads, trimming those that contain adapters, *Stacks* works only with same-sized reads. This means that when a read contains an adapter sequence, *Stacks* excludes that read completely, leading to the lower read count observed.

Another difference between both programs was the number of called SNPs in the final datasets, which was consistently higher for *Stacks*. This is especially evident for all-SNPs datasets, where *Stacks* returned 77,050 SNPs and *ipyrad* returned 19,652, a difference of 74%. The difference was smaller for the center-SNP datasets, with *ipyrad* returning around 45% of the number of SNPs in the *Stacks* dataset.

A decrease in the number of SNPs was observed when converting the all-SNPs datasets into center-SNP datasets (90% for *ipyrad*, 95% for *Stacks*). This difference might be explained by the existence of a high number of assembled loci with multiple variable sites at the end of the SNP calling processes. This would mean that, during the creation of center-SNP datasets, a considerable amount of loci had their information restricted to a single SNP, leading to a marked reduction of the final SNP count.

A difference in the final SNP count was observed between programs. This could be the result of (i) the different way *ipyrad*'s or *Stacks*' algorithms work, leading the programs to output a higher/lower number of SNPs, or (ii) the selected parameter values for each program, which may have, inadvertently, boosted the number of SNPs outputted by *Stacks* and/or reduced the number outputted by *ipyrad*. These two hypotheses are not mutually exclusive and might have both played a role. Regarding the latter, the selection of parameter values for *ipyrad* was based mainly on the assembly parameters used in Pina-Martins et al. (2019) [59], while for *Stacks*, the selection was not based on previous work, but mainly on parameter testing. Furthermore, parameter testing for both programs was not exhaustive, meaning there is certainly still room for improvement in terms of the final SNP count for both programs.

Impact on analyses results

The consistency of results between *ipyrad* and *Stacks* datasets varied slightly depending on the analysis. Population structure results were consistent in terms of the overall pattern of population structure, with results from *Stacks* datasets having slightly more resolution. For example, in *Stacks*' *fineRADstructure* results (dataset 'stk_allsnps_allsamp'), differentiation inside the *Q. ilex* and Moroccan *Q. rotundifolia* populations is more easily observed, and for the *ALStructure* results (dataset 'stk_centersnp_allsamp') the separation of Tunisian samples into a separate cluster is visible for lower values of K. Genetic diversity results differed very slightly between the different program's datasets, but the overall pattern of diversity was consistent.

The outlier detection and genetic-environmental association analyses used specific sub-datasets. For these, despite the difference in the number of SNPs between *ipyrad* and *Stacks* being slightly less pronounced, *ipyrad* datasets always returned less outlier/associated loci and, for the latter, fewer associated variables. This, in turn, led to a lower number of matched loci against the reference genome of *Q. lobata*, which made the comparison between datasets more difficult and compromised enrichment results for the

majority of *ipyrad* datasets.

4.2 Genetic diversity and population structure of Holm oak

Population structure analyses pointed to the division of our samples into three distinct genetic populations: (i) *Q. ilex*, including samples from Italy, France, Catalonia, and Tunisia, (ii) Iberian *Q. rotundifolia*, including samples from Portugal and Spain, and (iii) Moroccan *Q. rotundifolia*, including samples from Morocco. Our populations showed a global F_{st} of 0.122/0.123. *Q. ilex* and *Q. rotundifolia* samples showed considerable values of genetic differentiation (pairwise F_{st} = 0.129/0.128). Inside *Q. rotundifolia*, the Iberian and Moroccan groups showed a lower, yet considerable amount of differentiation (pairwise F_{st} = 0.077/0.080). This points to a relatively well-structured nature of this taxa.

All populations showed similar values of genetic diversity. Iberian *Q. rotundifolia* showed slightly lower values of heterozygosity (H_o = 0.093/0.084 and H_e = 0.110/0.100) when compared to the other two populations. Overall expected heterozygosity (H_e = 0.120/0.108) was slightly lower than what was observed for *Q. suber* [60], for example. Inbreeding coefficient values were positive and similar between all populations. Moroccan *Q. rotundifolia* showed the lowest values of inbreeding coefficient (F_{is} = 0.111/0.138), while Iberian *Q. rotundifolia* and *Q. ilex* showed slightly higher values (F_{is} = 0.149/0.146 and 0.135/0.167, respectively). These values probably reflect the structured nature of Holm oak, which is evident when compared to inbreeding coefficient values from relatively unstructured European oak species, such as *Q. robur* and *Q. patrea* [102] or *Q. suber* [59].

Patterns of genetic differentiation along the distribution of *Q. ilex* and *Q. rotundifolia* similar to the one observed here have been previously described: Michaud et al. (1995) [103] observed a difference in local allozyme alleles between samples from Greece, Italy, and southern France, and samples from the Iberian Peninsula, and Lumaret et al. (2002) [28] reported the existence of distinct chlorotypes in samples from the Balkans, Italy, and south-eastern France, from North Africa, and from the Iberian Peninsula. Both Michaud et al. (1995) [103] and Lumaret et al. (2002) [28] pointed to the role of ice age refugia as explaining the observed pattern of differentiation. During Quaternary glaciations, areas such as the Iberian, Italian, and Balkan peninsulas served as a refuge for many species [8]. According to the hypothesis laid down by Lumaret et al. (2002) [28], regions such as the southern Iberian, Italian, and Balkan Peninsulas, as well as northern Africa, might have served as a refuge for Holm oak populations. Their isolation in different refugia may have configured a barrier to gene flow, leading to differentiation, and post-glacial range expansion from these refugia would have given rise to the phylogeographical pattern observed along the area of distribution [28].

However, recent work by Hipp et al. (2020) [104], using more than 20,000 RAD-seq loci, has dated the divergence between *Q. ilex* and *Q. rotundifolia* to the early Pliocene, indicating that a considerable degree of differentiation between these two species was probably present earlier than the Quaternary. This puts into question the role of Quaternary glaciations as the main factor behind the divergence and present differentiation between the two taxa. Despite this, it is still safe to affirm that the current phylogeographical pattern observed for these taxa has probably been shaped by post-glacial range expansion and colonization from different glacial refugia, with the current *Q. rotundifolia* populations descending from populations that found refuge in the Iberian Peninsula and North Africa, while *Q. ilex* populations that found refuge in the Italian Peninsula (and possibly some Mediterranean islands) originated not only the present-day Italian population but also expanded their range northward toward southern France and the northeastern Iberian Peninsula, as suggested by Lumaret et al. (2002) [28]. Although Lumaret et al. (2002) [28] also hypothesize about post-glacial expansion from refugia in the Balkan peninsula, our lack

of sampling in that area does not allow us to draw conclusions regarding the easternmost portion of the distribution.

Nonetheless, the marked pattern of population structure and the considerable degree of genetic differentiation between *Q. rotundifolia* and *Q. ilex* shown in our results might also help inform the taxonomic organization of Holm oak. The observed pairwise F_{st} values between the two groups (pairwise $F_{st} = 0.129/0.128$) are, for example, considerably higher than those observed between other *Quercus* species such as *Q. scytophylla*, *Q. hypoleucoides* and *Q. sideroxyla* (pairwise F_{st} ranging from 0.022 to 0.071, based on nuclear microsatellite markers) [105], between *Q. magnoliifolia* and *Q. resinosa* (pairwise $F_{st} = 0.016$, pairwise F_{st} ranging from 0.022 to 0.071, based on nuclear microsatellite markers) [106] or between four species of European White oaks (mean interchromosomal 10-kb estimates of F_{st} below 0.08 for all pairs of species, based on 31,894,340 SNPs) [107]. The marked genetic differences evidenced in this work, added to the already mentioned genetic [28, 103], morphological [30], biochemical [30], and ecological [29] differences between *Q. rotundifolia* and *Q. ilex*, constitute, in our opinion, enough evidence to solidify the status of these taxa as two distinct species.

Regarding *Q. rotundifolia* in particular, the degree of differentiation observed between the Iberian and Moroccan populations was an interesting result and is congruent with several previously described differences between these populations: Lumaret et al. (2002) [28] found distinct but related chlorotypes in the Iberian Peninsula and North Africa, with a possible evolutionarily intermediate chlorotype found in southern Spain, leading them to hypothesize the colonization of Iberia through Morocco; Michaud et al. (1995) [103] observed a number of allozyme alleles exclusive to Moroccan and Tunisian populations; Lebreton et al. (2001) [30] showed these groups to be similar in terms of leaf morphology, but distinct in terms of the biochemical composition of their leaves. Our results show this differentiation to be especially evident when comparing Iberian individuals with some individuals from Taza or Azrou. This seems to support the hypothesis laid down by Lumaret et al. (2002) [28] that, during the glacial periods, there were efficient barriers to gene flow between Iberian and Moroccan populations, probably due to geographical discontinuities, that led to the observed differentiation. Additionally, our population structure results seem to confirm that Moroccan/African *Q. rotundifolia* did not contribute to the post-glacial recolonization of the Northern Iberian Peninsula, or the peninsula in general [28], as evidenced by the overall differentiation values between the Iberian and Moroccan populations, or by our population structure and PCA results.

Our results also seem to evidence a difference in the way diversity is organized within both populations: in the Iberian Peninsula, *Q. rotundifolia* seems to constitute a big, homogeneous, unstructured population; in Morocco, there seems to be a more marked pattern of population structure, evident when comparing individuals from Rif with those from Taza and Azrou. In Iberia, post-glacial expansion from the southern refugia, with individuals colonizing a relatively homogeneous environment, and with the maintenance of a high degree of gene flow, could have led to the relatively homogeneous population evidenced by our results. The occurrence of gene flow between the several refugia during glacial periods, or contact between refugia during inter-glacial range expansions may have also contributed to this. In Morocco, on the other hand, a lot of the population structure evidenced in our results seems to stem from the shared genetic diversity between Iberia and some Moroccan individuals, mainly from the Rif. Individuals from Taza and, mainly, Azrou seem to form a more homogenous core within Moroccan *Q. rotundifolia*. This seems to indicate that this area of the distribution might have its origin in the post-glacial range expansion of relatively isolated glacial refugia populations, probably associated with the High and Middle Atlas mountain ranges. Nonetheless, to better characterize the population structure of Moroccan *Q. rotundifolia* additional sampling is required, in specific, towards the southernmost point of

distribution.

Inside the Moroccan *Q. rotundifolia* population, results showed that samples from Rif, in particular, seem to share a higher degree of genetic diversity with Iberian *Q. rotundifolia*. Pair-wise F_{st} values corroborate this, showing lower values of differentiation between Rif and Iberian sampling locations, than those observed between Rif and the other Moroccan sampling locations. This pattern of shared genetic diversity might have been the result of gene flow followed by introgression, mainly from Iberia to Morocco, either through secondary, post-glacial contact followed by introgression, or during glacial periods or inter-glacial range expansions, given the geographical proximity between the putative refugia in the Rif and in southern Iberia. Alternatively, this pattern might also be the result of incomplete lineage sorting. Statistics such as the D statistic can be used to detect introgression between populations and help distinguish between the occurrence of incomplete lineage sorting and gene flow [108]. This was not done in this work, so the process shaping this pattern of differentiation remains to be studied more deeply.

Nonetheless, the overall degree of differentiation observed between Iberian and Moroccan *Q. rotundifolia* samples might warrant the classification of one of the groups as a sub-species, as proposed by Lebreton et al. (2001) [30].

An unexpected result of this work was the placement of Tunisian samples as part of *Q. ilex*, since (i) all consulted species distribution information classified Tunisian individuals as *Q. rotundifolia*, (ii) morphometric analyses of leaves grouped Tunisian individuals with those from Morocco and Algeria [30], and (iii) Tunisian individuals have been shown to be in the same chlorotype group as others from North Africa [28]. Given this information, a possible explanation for the pattern of population structure observed here could be the occurrence of post-glacial, long-distance pollen flow from individuals in the southernmost portions of the *Q. ilex* ranges (possibly from Sardinia or Sicily) to Tunisian *Q. rotundifolia* individuals. Something similar has been suggested by Petit et al. (1997) [109] to explain chloroplastidial DNA spatial patterns in French populations of *Q. robur* and *Q. patrea*. This pollen flow, followed by introgression of *Q. ilex* genetic material could have led to a replacement of *Q. rotundifolia* genetic material in this area, which would explain the similarity between Tunisian individuals and other *Q. ilex* populations revealed when studying nuclear DNA. Additionally, the fact that organelle DNA is maternally inherited in oaks would mean that this pollen flow would leave the chloroplastidial genome unaffected, which would explain the observed pattern of chlorotype distribution. Regardless, the similarity between all North African populations in terms of leaf morphology could be explained by an inconsistency between morphological traits and nuclear polymorphisms, as suggested by Michaud et al. (1995) [103]. It has been shown that Holm oak morphotypes correspond to distinct ecotypes that differ in their response to climate dryness [29]. An explanation for this could be that, as individuals in Tunisia occupy environments more similar to those occupied by *Q. rotundifolia* than by *Q. ilex* in terms of dryness, leaf morphology might tend to converge despite their genetic differences. Nonetheless, sampling in southern Italy/Sicily and the unsampled stretch of the species distribution between Tunisia and Morocco will be essential to help us better understand the pattern being observed in Tunisia.

4.3 Detection of selection in Holm oak

With the objective of exploring the existence of local adaptation in Iberian and Moroccan populations of *Q. rotundifolia*, *BayPass* was used to investigate the existence of variable sites throughout the genome under the putative effect of natural selection. This was done first through the detection of F_{st} outliers, and second, through genetic-environmental associations between variable sites and several bioclimatic

variables.

Detection of F_{st} outliers

Outlier detection analyses uncovered 4 and 6 outlier SNPs for the datasets containing only Iberia samples ('ipr_centersnp_ib' and 'stk_centersnp_ib', respectively) and 16 and 34 outlier SNPs for the datasets containing Iberian and Moroccan samples ('ipr_centersnp_ib_mor' and 'stk_centersnp_ib_mor', respectively). For more differentiated populations with higher average values of F_{st} , it tends to be more difficult to detect F_{st} outliers based on departures from neutrality [63]. The fact that we detected a higher number of outliers in datasets containing both Iberian and Moroccan samples than in outliers containing only Iberian samples was unexpected. If anything, the higher level of differentiation between Iberian and Moroccan individuals should have made it harder to detect F_{st} outliers. This might be explained by the fact that the majority of SNPs detected as outliers tend to have really distinct allele frequencies between Iberian and Moroccan sampling locations. This would lead to F_{st} values high enough above the threshold of what is considered neutral variation for these SNPs to be detected as outliers.

Detection of genetic-environmental associations and allele turnover patterns

Genetic-environmental association analyses uncovered a high number of SNPs showing signs of association with 11 of the 12 climatic variables used. The percentage of SNPs in the highest level of association strength ranged from 0.4% to 7% of the total number of SNPs in the datasets. Variables related to precipitation (*mean monthly precipitation amount of the driest quarter*, *precipitation seasonality*, and *annual precipitation amount*) appeared as associated in most datasets, with *mean monthly precipitation amount of the driest quarter* appearing as associated in all. Of the variables related to temperature, *temperature seasonality*, *mean daily mean air temperature of the driest quarter*, and *mean annual air temperature* appeared in most datasets. These variables are similar to those found by Pina-Martins et al. (2019) [59] and Vanhove et al. (2021) [60] for *Q. suber*.

Plotting the associated SNPs allele frequencies along the respective bioclimatic gradients showed that the allele frequency turnover is not linear. This was observed for both individual SNPs, and when plotting the average allele frequencies along each bioclimatic variable. Looking at the pattern of allele turnover, there seems to be a 'turnover threshold', a value or range of values where the change in allele frequency tends to occur. Allele frequencies either tend to appear fixed up until that value, and start to decrease as the variable rises beyond it (downward pattern), or are fixed beyond that value, and start to decrease as the variable decreases past it (upward pattern). This was especially evident when SNPs were classified according to their turnover pattern (upward vs. downward pattern) and then the average allele frequencies for each of these categories were plotted along the respective bioclimatic variable.

Comparing the plotted allele frequencies of individual SNPs associated with the same variable, we can observe that the value (or range of values) where the 'turnover threshold' seems to be located tends to be relatively consistent between them. Additionally, the location of the 'turnover threshold' seems to be relatively consistent across datasets for each bioclimatic variable, regardless of the samples in the dataset (Iberia vs Iberia and Morocco) or the SNP calling software used to generate the dataset (*ipyrad* vs. *Stacks*). The latter is expected, assuming a considerable number of SNPs is common to both types of datasets, despite the use of different software.

Differences between Iberia and Morocco in terms of associated variables

We observed some differences in terms of the number of associations/associated SNPs, as well as associated variables, between datasets containing only Iberian ('ipr_centersnp_ib' and 'stk_centersnp_ib') and Iberian and Moroccan samples ('ipr_centersnp_ib_mor' and 'stk_centersnp_ib_mor'). The number of associations/associated SNPs in the former was consistently higher than in the latter. A difference in the number of SNPs between datasets could explain this pattern but, although datasets 'ipr_centersnp_ib_mor' and 'stk_centersnp_ib_mor' had slightly fewer SNPs, the difference seems too small to serve as a satisfying explanation. A more likely explanation is that the observed difference lies in the inclusion or exclusion of Moroccan samples in the analyses. The addition of environmental data from Moroccan sampling locations might have brought forward associations not identified in the datasets containing only Iberian samples while masking associations that had been found in those datasets. For example, for SNPs/loci where the variation of allele frequencies between sampling locations is small, the addition of three new sampling locations along the environmental gradient might influence if associations are detected or not.

Comparing both types of datasets regarding the type of associated variables found, some differences are clear. Although environmental variables related to precipitation appeared associated in all datasets, they were most prevalent in datasets containing Iberian and Moroccan samples: for 'ipr_centersnp_ib_mor', all associated variables were related to precipitation; for 'stk_centersnp_ib_mor', these variables made up the majority of associations. For the three bioclimatic variables *mean monthly precipitation amount of the driest quarter*, *precipitation seasonality*, and *annual precipitation amount*, Moroccan sampling sites tend to be in the extremes of the distribution: Moroccan sampling locations have the highest values of precipitation seasonality (Supplemental Figure S4), displaying some of the highest amounts of annual precipitation (Figure 3.6, Supplemental Figure S4). Also, Rif and Taza, show some of the lowest amounts of mean monthly precipitation in the driest quarter (Supplemental Figure S14). This points to a big variation in precipitation between the driest and wettest periods of the year, with an intense lack of precipitation during the dry season. These particular environmental conditions of Moroccan sampling locations might have led to associations with precipitation-related variables to make up the majority of associations for these datasets.

On the other hand, bioclimatic variables associated with temperature, which make up the great majority of associations in datasets containing only Iberian samples, make up a minority of associations in the datasets containing Iberian and Moroccan samples (Table 3.5, Supplemental Figure S8a). When we look at the distribution of bioclimatic variable values for all sampling locations (Supplemental Figure S4), we can see that for most of the temperature-related variables used in the analyses, the placement of Moroccan sampling locations is not as extreme as for the precipitation related variables. This might have led to a lower strength of association with variables associated with temperature when Moroccan sample data is present.

Putative role of outlier and associated loci in heat and water stress response

Exploring the putative function of the genomic regions where the outlier and associated SNPs were found can help inform us of their role in adaptation. For this, the loci sequences were queried against the reference genome of *Quercus lobata*. The number of sequence matches was rather small for all datasets, and those with the lowest number of loci (outlier loci for 'ipr_centersnp_ib' and 'stk_centersnp_ib') returned no sequence matches. The number of queried dataset loci that matched at least one *Q. lobata* gene was, on average, 10% (Table 3.4). This can probably be attributed to the distance between *Q. rotundifolia*/*Q. ilex* and *Q. lobata*.

The matched *Q. lobata* genes for each dataset were compared. The number of common genes between datasets was relatively low (Supplemental Figure S16). The lack of common genes between *ipyrad* datasets (Supplemental Figure S16 (a)) can probably be attributed to their lower number of loci when compared to their *Stacks* counterparts. Associated loci from datasets 'stk_centersnp_ib' and 'stk_centersnp_ib_mor' showed the most genes in common, with 50, followed by associated loci from datasets 'ipr_centersnp_ib' and 'stk_centersnp_ib' with 10.

A few outlier loci matched regions of the *Q. lobata* genome that point to potentially interesting roles relating to heat or water stress. *ipyrad* locus RAD_19905 matched a region annotated as "ABC1 domain-containing protein". Kinases from the ABC1K family, characterized by a conserved ABC1 domain, have been shown to play a role in the response to water stress [110]. *Stacks* CLocus_7350 matched a region annotated as "Protein SUPPRESSOR OF GENE SILENCING 3-like". The gene Protein SUPPRESSOR OF GENE SILENCING 3 has been described as being involved in mechanisms for quick response to heat and heritable trans-generational memory of heat acclimation [111].

The potential role of the 50 *Q. lobata* genes found to be common to datasets 'stk_centersnp_ib' and 'stk_centersnp_ib_mor' was also explored. The results are expressed in Supplemental Table S7. We found genes of *Q. lobata* with putative roles relating to heat and water stress. Among them was a region potentially related to plant F-box proteins, which are known to be involved in response to abiotic stress. In wheat, water stress has been shown to induce the expression of these proteins. In transgenic tobacco plants, over-expression of F-box gene *TaFBA1* has been shown to lead to drought-resistant phenotypes. In *Arabidopsis* and apple, F-box genes are associated with signal transduction mechanisms involved in stress response under drought conditions [112]. Also found was the gene Zeoxanthin epoxidase, which plays a role in the biosynthesis of abscisic acid (ABA), a very important hormone in water stress signal transduction [113] and acquired thermotolerance [114]. A region annotated as "TTF-type domain-containing protein" was found. Trihelix transcription factors (TTFs) have been found to play important roles in terms of response to drought stress [115, 116]. Another region was annotated as an "SBP-type domain-containing protein". Squamosa promoter binding proteins (SBP) have been shown to be involved in drought and heat-stress responses [117]. The genomic region with the most matched loci was annotated as "non-specific serine/threonine protein kinase". Serine-Threonine Kinases have been shown to be positive regulators of ABA-mediated drought-stress tolerance [118]. A region annotated as "C2 domain-containing protein" was also found. C2 domain genes have been shown to be important in the response to drought stress [119]. The gene Xanthine dehydrogenase was also found, which may be important in drought acclimatization through its role in purine metabolism [120]. And finally, a region annotated as "BHLH domain-containing protein". The bHLH family of transcription factors is known to regulate drought-tolerant responses through modulation ABA sensitivity or regulation of stomata, leaf trichomes, and root hair development [121].

Additionally, and in a simple and exploratory way, enrichment analysis was run for the *Q. lobata* genes that matched the outlier and associated loci for the different datasets. Overall, the results were weak: some datasets did not return results, and for those that did, the number of enriched GO terms was reduced. This may be attributed to (i) the reduced amount of data (low number of loci in many of our datasets, and the even lower number of *Q. lobata* genes that matched our loci), and (ii) the degree of annotation for the genes/proteins in the *Q. lobata* genome, for which the existence of most proteins is either inferred from homology or predicted, and assignment of Gene Ontology terms is mainly inferred from electronic annotation, which is based on the general GO term signature of homologous proteins. Nonetheless, some of the obtained GO terms show potentially interesting roles in the response to heat and water stress. For dataset 'ipr_centersnp_ib', the terms "asparagine metabolic process", "asparagine

biosynthetic process” and ”asparagine synthase (glutamine-hydrolyzing) activity” were enriched. Genes involved in asparagine metabolism have been linked to the response to stress conditions, particularly in wheat roots and leaves under drought stress [122]. Dataset ’stk_centersnp_ib_mor’ showed enrichment of terms ”pyroglutamyl-peptidase activity” and ”omega peptidase activity”. Pyroglutamyl-peptidases are part of the cysteine protease family, where some members have been shown to be involved in the response to abiotic stresses such as heat [123]. There is also growing evidence of the involvement of proteases in plant response to abiotic stresses such as drought [124]

4.4 Conclusion and final remarks

In this work, we present what is, as far as we know, the first study into the population structure of Holm oak using genome-wide nuclear genetic markers. We uncovered a marked pattern of population structure that supports the existence of two genetically, morphologically, and ecologically distinct species, *Quercus rotundifolia* and *Quercus ilex*, contributing towards the clarification of the taxonomy of Holm oak, which up until this point has suffered from a lack of consistency. Furthermore, we found a considerable degree of differentiation between the Iberian and Moroccan populations of *Q. rotundifolia*, probably attributable to barriers to gene flow resulting from the isolation of populations in different Iberian and Moroccan Quaternary refugia. Inside the Moroccan distribution of the species, our results show that samples from areas geographically closer to Iberia tend to show a lower degree of differentiation from the Iberian population, which may indicate the presence of gene flow between northern Morocco and southern Iberia. We also found the Iberian population of *Q. rotundifolia* to be relatively unstructured, a pattern probably originated from post-glacial colonization of the peninsula from southern Quaternary refugia, allied to a balance between gene flow and local adaptation, leading to homogeneity in terms of genetic diversity. Finally, we uncover the presence of *Q. ilex* in Tunisia, an area that, up until this point, was thought to be inhabited by *Q. rotundifolia*. This result reveals an incongruity between chlorotype distribution and the pattern of population structure inferred from nuclear DNA. We hypothesize this to be the result of hybridization through long-distance pollen flow, probably originating from Sicily or Sardinia. In general, these results offer new support to a previously hypothesized east-west colonization of northern Africa by a lineage originating from the central-east Mediterranean basin, which then colonized Iberia from Morocco, and gave rise to the current distribution of *Q. rotundifolia*.

We also used a Landscape Genomics approach to identify putative SNPs under selection in *Q. rotundifolia* samples of Iberia and Morocco. This was done through the detection of F_{st} -outliers and genetic-environmental associations between SNPs and several bioclimatic variables. We were able to identify a considerable number of SNPs associated with precipitation and temperature-related bioclimatic variables, indicating the existence of variation clines resulting from the adaptation of individuals to local conditions. We also observed that datasets that included both Moroccan and Iberian samples, instead of only Iberian samples, tended to show a dominance of associations with precipitation-related bioclimatic variables. We found a curious pattern of allele frequency turnover along the bioclimatic variables, where allele frequencies show a tendency to be fixed either above or below a certain range of values. We also found several of the outlier/associated SNPs to be potentially associated with genomic regions with a putative role in the response to abiotic stressors such as heat and drought. These results represent a first step in understanding the capacity of Mediterranean Holm oak populations to respond to climate change and open the door to more complex types of analyses, such as genomic prediction of maladaptation, which will help identify areas of the distribution especially sensitive to climate change and aid future management efforts towards the conservation of Holm oak in the Mediterranean Basin.

Bibliography

1. Mazor, T. *et al.* Global mismatch of policy and research on drivers of biodiversity loss. *Nature Ecology Evolution* **2**, 1071–1074. ISSN: 2397-334X. <https://www.nature.com/articles/s41559-018-0563-x> (7 May 2018).
2. Foden, W. B. *et al.* Climate change vulnerability assessment of species. *WIREs Climate Change* **10**. ISSN: 1757-7780. <https://onlinelibrary.wiley.com/doi/10.1002/wcc.551> (1 Jan. 2019).
3. Langsdorf, S. *et al.* *IPCC, 2022: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge, UK and New York, NY, USA, 2022). www.environmentalgraphiti.org.
4. Aurelle, D. *et al.* Biodiversity, climate change, and adaptation in the Mediterranean. *Ecosphere* **13**, e3915. ISSN: 2150-8925. <https://onlinelibrary.wiley.com/doi/10.1002/ecs2.3915> (4 Apr. 2022).
5. Lionello, P. *et al.* *The Mediterranean climate: An overview of the main characteristics and issues* (eds Lionello, P, Malanotte-Rizzoli, P & Boscolo, R) 2006. <https://www.sciencedirect.com/science/article/pii/S1571919706800030>.
6. Gil-Pelegrín, E., Ángel Saz, M., Cuadrat, J. M., Peguero-Pina, J. J. & Sancho-Knapik, D. Oaks Under Mediterranean-Type Climates: Functional Response to Summer Aridity, 137–193. https://link.springer.com/chapter/10.1007/978-3-319-69099-5_5 (2017).
7. Ramdani, M., Elkhiaiti, N. & Flower, R. *Lakes of Africa: North of Sahara* (ed Likens, G. E.) 2009. <https://linkinghub.elsevier.com/retrieve/pii/B9780123706263000351>.
8. Médail, F. & Diadema, K. Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *Journal of Biogeography* **36**, 1333–1345. ISSN: 03050270. <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2699.2008.02051.x> (7 July 2009).
9. Blondel, J. & Aronson, J. *Biology and wildlife of the Mediterranean* ISBN: 019850036X, 9780198500360 (Oxford University Press, Aug. 1999).
10. Cos, J. *et al.* The Mediterranean climate change hotspot in the CMIP5 and CMIP6 projections. *Earth System Dynamics* **13**, 321–340. ISSN: 2190-4987. <https://esd.copernicus.org/articles/13/321/2022/> (1 Feb. 2022).
11. Dubrovský, M. *et al.* Multi-GCM projections of future drought and climate variability indicators for the Mediterranean region. *Regional Environmental Change* **14**, 1907–1919. ISSN: 1436-3798. <http://link.springer.com/10.1007/s10113-013-0562-z> (5 Oct. 2014).

12. Jacobeit, J., Hertig, E., Seubert, S. & Lutz, K. Statistical downscaling for climate change projections in the Mediterranean region: methods and results. *Regional Environmental Change* **14**, 1891–1906. ISSN: 1436-3798. <http://link.springer.com/10.1007/s10113-014-0605-0> (5 Oct. 2014).
13. Thuiller, W. *et al.* Predicting global change impacts on plant species' distributions: Future challenges. *Perspectives in Plant Ecology, Evolution and Systematics* **9**, 137–152. ISSN: 14338319. <https://linkinghub.elsevier.com/retrieve/pii/S1433831907000376> (3–4 Mar. 2008).
14. Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T. & Curtis-McLane, S. Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications* **1**, 95–111. ISSN: 17524571. <https://onlinelibrary.wiley.com/doi/10.1111/j.1752-4571.2007.00013.x> (1 Feb. 2008).
15. Ghalambor, C. K., Mckay, J. K., Carroll, S. P. & Reznick, D. N. Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. *Functional Ecology* **21**, 394–407. <https://besjournals.onlinelibrary.wiley.com/doi/10.1111/j.1365-2435.2007.01283.x> (2007).
16. Nicotra, A. *et al.* Plant phenotypic plasticity in a changing climate. *Trends in Plant Science* **15**, 684–692. ISSN: 13601385. <https://linkinghub.elsevier.com/retrieve/pii/S1360138510001986> (12 Dec. 2010).
17. Huey, R. B., Hertz, P. E. & Sinervo, B. Behavioral drive versus behavioral inertia in evolution: A null model approach. *American Naturalist* **161**, 357–366. ISSN: 00030147. https://www.researchgate.net/publication/10801795_Behavioral_Drive_versus_Behavioral_Inertia_in_Evolution_A_Null_Model_Approach (3 Mar. 2003).
18. Miryeganeh, M. & Saze, H. Epigenetic inheritance and plant evolution. *Population Ecology* **62**, 17–27. ISSN: 1438-3896. <https://onlinelibrary.wiley.com/doi/10.1002/1438-390X.12018> (1 Jan. 2020).
19. Barrett, R. & Schluter, D. Adaptation from standing genetic variation. *Trends in Ecology & Evolution* **23**, 38–44. ISSN: 01695347. <https://linkinghub.elsevier.com/retrieve/pii/S0169534707002868> (1 Jan. 2008).
20. Kremer, A. *et al.* Long-distance gene flow and adaptation of forest trees to rapid climate change. *Ecology Letters* **15**, 378–392. ISSN: 1461-023X. <https://onlinelibrary.wiley.com/doi/10.1111/j.1461-0248.2012.01746.x> (4 Apr. 2012).
21. Denk, T., Grimm, G. W., Manos, P. S., Deng, M. & Hipp, A. L. *An Updated Infrageneric Classification of the Oaks: Review of Previous Taxonomic Schemes and Synthesis of Evolutionary Patterns* 2017. http://link.springer.com/10.1007/978-3-319-69099-5_2.
22. Caudullo, G., Welk, E. & San-Miguel-Ayanz, J. Chorological maps for the main European woody species. *Data in Brief* **12**, 662–666. ISSN: 23523409. <https://linkinghub.elsevier.com/retrieve/pii/S2352340917301981> (June 2017).
23. Terradas, J. *Holm Oak and Holm Oak Forests: An Introduction* 1999. http://link.springer.com/10.1007/978-3-642-58618-7_1.

24. Amimi, N., Ghouil, H., Zitouna-Chebbi, R., Joët, T. & Ammari, Y. Intraspecific variation of *Quercus ilex* L. seed morphophysiological traits in Tunisia reveals a trade-off between seed germination and shoot emergence rates along a thermal gradient. *Annals of Forest Science* **80**, 12. ISSN: 1297-966X. <https://annforsci.biomedcentral.com/articles/10.1186/s13595-023-01179-7> (1 Feb. 2023).
25. Bensaci, O. A., Beghami, R. & Gouaref, K. First report of *Apiognomonina errabunda* on *Quercus ilex* in Algeria. *Folia Forestalia Polonica* **63**, 10–20. ISSN: 2199-5907. <https://www.sciendo.com/article/10.2478/ffp-2021-0002> (1 Mar. 2021).
26. Joffre, R., Rambal, S. & Ratte, J. P. The dehesa system of southern Spain and Portugal as a natural ecosystem mimic. *Agroforestry Systems* **45**, 57–79. ISSN: 01674366. <https://link.springer.com/article/10.1023/A:1006259402496> (1-3 1999).
27. Ferrer-Galego, P. P. & Sáez, L. Type designation of the oaks *Quercus ballota* and *Q. rotundifolia* (Fagaceae). *Nordic Journal of Botany* **37**, njb.02387. ISSN: 0107-055X. <https://onlinelibrary.wiley.com/doi/10.1111/njb.02387> (6 June 2019).
28. Lumaret, R., Mir, C., Michaud, H. & Raynal, V. Phylogeographical variation of chloroplast DNA in holm oak (*Quercus ilex* L.) *Molecular Ecology* **11**, 2327–2336. ISSN: 09621083. <http://doi.wiley.com/10.1046/j.1365-294X.2002.01611.x> (11 Oct. 2002).
29. Peguero-Pina, J. J. *et al.* Morphological and physiological divergences within *Quercus ilex* support the existence of different ecotypes depending on climatic dryness. *Annals of Botany* **114**, 301–313. ISSN: 1095-8290. <https://academic.oup.com/aob/article-lookup/doi/10.1093/aob/mcu108> (2 Aug. 2014).
30. Lebreton, P., Barbéro, M. & Quézel, P. Contribution morphométrique et biochimique à la structuration et à la systématique du complexe spécifique Chêne vert *Quercus ilex* L. *Acta Botanica Gallica* **148**, 289–317. ISSN: 1253-8078. <http://www.tandfonline.com/doi/abs/10.1080/12538078.2001.10515917> (4 Jan. 2001).
31. Jiang, X. *et al.* East Asian origins of European holly oaks (*Quercus* section *Ilex* Loudon) via the Tibet-Himalaya. *Journal of Biogeography* **46**, 2203–2214. ISSN: 0305-0270. <https://onlinelibrary.wiley.com/doi/10.1111/jbi.13654> (10 Oct. 2019).
32. Kovar-Eder, J. Vegetation dynamics in Europe during the Neogene. *Distribution and Migration of Tertiary Mammals in Eurasia. A Volume in Honour of Hans De Bruijn - DEINSEA* **10** (eds Reumer, J. & Wessels, W) 373–392. ISSN: 0923-9308 (Dec. 2003).
33. Barbero, M., Loisel, R. & Quézel, P. *Biogeography, ecology and history of Mediterranean Quercus ilex ecosystems* 1992. http://link.springer.com/10.1007/978-94-017-2836-2_2.
34. Muller, S. D. *et al.* Vegetation history of the western Rif mountains (NW Morocco): origin, late-Holocene dynamics and human impact. *Vegetation History and Archaeobotany* **24**, 487–501. ISSN: 09396314. https://www.researchgate.net/publication/268980912_Vegetation_history_of_the_western_Rif_mountains_NW_Morocco_origin_late-Holocene_dynamics_and_human_impact (4 Nov. 2015).
35. Martín-Sánchez, R. *et al.* Summer and winter can equally stress holm oak (*Quercus ilex* L.) in Mediterranean areas: A physiological view. *Flora* **290**, 152058. ISSN: 03672530. <https://linkinghub.elsevier.com/retrieve/pii/S036725302200055X> (May 2022).

36. Peguero-Pina, J. J. *et al.* Evidence of vulnerability segmentation in a deciduous Mediterranean oak (*Quercus subpyrenaica* E. H. del Villar). *Trees* **29**, 1917–1927. ISSN: 0931-1890. <http://link.springer.com/10.1007/s00468-015-1273-5> (6 Dec. 2015).
37. Pinto, C. A. *et al.* Drought-induced embolism in current-year shoots of two Mediterranean evergreen oaks. *Forest Ecology and Management* **285**, 1–10. ISSN: 03781127. <https://linkinghub.elsevier.com/retrieve/pii/S037811271200480X> (Dec. 2012).
38. Urli, M. *et al.* The high vulnerability of *Quercus robur* to drought at its southern margin paves the way for *Quercus ilex*. *Plant Ecology* **216**, 177–187. ISSN: 1385-0237. <http://link.springer.com/10.1007/s11258-014-0426-8> (2 Feb. 2015).
39. Corcuera, L., Camarero, J. J. & Gil-Pelegrín, E. Effects of a severe drought on *Quercus ilex* radial growth and xylem anatomy. *Trees - Structure and Function* **18**, 83–92. ISSN: 09311890. <https://link.springer.com/article/10.1007/s00468-003-0284-9> (1 Jan. 2004).
40. Mediavilla, S. & Escudero, A. Stomatal responses to drought at a Mediterranean site: a comparative study of co-occurring woody species differing in leaf longevity. *Tree Physiology* **23**, 987–996. ISSN: 0829-318X. <https://academic.oup.com/treephys/article-lookup/doi/10.1093/treephys/23.14.987> (14 Oct. 2003).
41. Mediavilla, S. & Escudero, A. Stomatal responses to drought of mature trees and seedlings of two co-occurring Mediterranean oaks. *Forest Ecology and Management* **187**, 281–294. ISSN: 03781127. <https://linkinghub.elsevier.com/retrieve/pii/S0378112703003827> (2-3 Jan. 2004).
42. Früchtenicht, E., Neumann, L., Klein, N., Bonal, D. & Brüggemann, W. Response of *Quercus robur* and two potential climate change winners—*Quercus pubescens* and *Quercus ilex*—To two years summer drought in a semi-controlled competition study: I—Tree water status. *Environmental and Experimental Botany* **152**, 107–117. ISSN: 00988472. <https://linkinghub.elsevier.com/retrieve/pii/S0098847218300091> (Aug. 2018).
43. Morales-Saldaña, S. *et al.* Even more oak species in Mexico? Genetic structure and morphological differentiation support the presence of at least two specific entities within *Quercus laeta*. *Journal of Systematics and Evolution* **60**, 1124–1139. ISSN: 1674-4918. <https://onlinelibrary.wiley.com/doi/10.1111/jse.12818> (5 Sept. 2022).
44. Fernández, V. *et al.* Wettability, Polarity, and Water Absorption of Holm Oak Leaves: Effect of Leaf Side and Age. *Plant Physiology* **166**, 168–180. ISSN: 0032-0889. <https://academic.oup.com/plphys/article/166/1/168/6113185> (1 Aug. 2014).
45. Abeyrama, D. K., Boyle, B. & Burg, T. M. Comparison of genotyping by sequencing procedures to determine population genetic structure. *Functional Integrative Genomics* **23**, 9. ISSN: 1438-793X. <https://link.springer.com/10.1007/s10142-022-00929-6> (1 Mar. 2023).
46. Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R. & Hohenlohe, P. A. Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology* **22**, 2841–2847. ISSN: 09621083. <https://onlinelibrary.wiley.com/doi/10.1111/mec.12350> (11 June 2013).
47. Elshire, R. J. *et al.* A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* **6** (ed Orban, L.) e19379. ISSN: 1932-6203. <https://dx.plos.org/10.1371/journal.pone.0019379> (5 May 2011).

48. Davey, J. W. *et al.* Special features of RAD Sequencing data: implications for genotyping. *Molecular ecology* **22**, 3151–3164. ISSN: 1365-294X. <https://pubmed.ncbi.nlm.nih.gov/23110438/> (11 June 2013).
49. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **12**, 443–451. ISSN: 1471-0056. <http://www.nature.com/articles/nrg2986> (6 June 2011).
50. Eaton, D. A. R. & Overcast, I. ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics (Oxford, England)* **36**, 2592–2594. ISSN: 1367-4811. <http://www.ncbi.nlm.nih.gov/pubmed/31904816> (8 Apr. 2020).
51. Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. & Cresko, W. A. Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124–3140. ISSN: 09621083. <https://onlinelibrary.wiley.com/doi/10.1111/mec.12354> (11 June 2013).
52. Catchen, J. *et al.* *Stacks Manual* <https://catchenlab.life.illinois.edu/stacks/manual/#intro>.
53. Liu, J., Shen, Q. & Bao, H. Comparison of seven SNP calling pipelines for the next-generation sequencing data of chickens. *PLOS ONE* **17** (ed Wu, S.-B.) e0262574. ISSN: 1932-6203. <https://dx.plos.org/10.1371/journal.pone.0262574> (1 Jan. 2022).
54. Liu, X., Han, S., Wang, Z., Gelernter, J. & Yang, B.-Z. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE* **8** (ed Khanin, R.) e75619. ISSN: 1932-6203. <https://dx.plos.org/10.1371/journal.pone.0075619> (9 Sept. 2013).
55. Hwang, S., Kim, E., Lee, I. & Marcotte, E. M. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports* **5**, 17875. ISSN: 2045-2322. <https://www.nature.com/articles/srep17875> (1 Dec. 2015).
56. Sandmann, S. *et al.* Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Scientific Reports* **7**, 43169. ISSN: 2045-2322. <https://www.nature.com/articles/srep43169> (1 Feb. 2017).
57. Hwang, K.-B. *et al.* Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. *Scientific Reports* **9**, 3219. ISSN: 2045-2322. <https://www.nature.com/articles/s41598-019-39108-2> (1 Mar. 2019).
58. Gézsi, A. *et al.* VariantMetaCaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering. *BMC Genomics* **16**, 875. ISSN: 1471-2164. <http://www.biomedcentral.com/1471-2164/16/875> (1 Dec. 2015).
59. Pina-Martins, F., Baptista, J., Pappas, G. & Paulo, O. S. New insights into adaptation and population structure of cork oak using genotyping by sequencing. *Global Change Biology* **25**, 337–350. ISSN: 1354-1013. <https://onlinelibrary.wiley.com/doi/10.1111/gcb.14497> (1 Jan. 2019).
60. Vanhove, M. *et al.* Using gradient Forest to predict climate response and adaptation in Cork oak. *Journal of Evolutionary Biology* **34**, 910–923. ISSN: 1010-061X. <https://onlinelibrary.wiley.com/doi/10.1111/jeb.13765> (6 June 2021).
61. Novembre, J. Pritchard, Stephens, and Donnelly on Population Structure. *Genetics* **204**, 391–393. ISSN: 1943-2631. <https://academic.oup.com/genetics/article/204/2/391/6072054> (2 Oct. 2016).

62. Sbordoni, V., Allegrucci, G. & Cesaroni, D. *Population Structure* Jan. 2012. <https://linkinghub.elsevier.com/retrieve/pii/B9780123838322000906>.
63. Hoban, S. *et al.* Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *The American Naturalist* **188**, 379–397. ISSN: 0003-0147. <https://www.journals.uchicago.edu/doi/10.1086/688018> (4 Oct. 2016).
64. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic Maps of Human Gene Frequencies in Europeans. *Science* **201**, 786–792. ISSN: 0036-8075. <https://www.science.org/doi/10.1126/science.356262> (4358 Sept. 1978).
65. Malinsky, M., Trucchi, E., Lawson, D. J. & Falush, D. RADpainter and fineRADstructure: Population Inference from RADseq Data. *Molecular Biology and Evolution* **35** (ed Takezaki, N.) 1284–1290. ISSN: 0737-4038. <https://academic.oup.com/mbe/article/35/5/1284/4883220> (5 May 2018).
66. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**, 945–959. ISSN: 00166731. <https://academic.oup.com/genetics/article/155/2/945/6048111> (2 June 2000).
67. Lawson, D. J., van Dorp, L. & Falush, D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications* **9**, 3258. ISSN: 2041-1723. <https://www.nature.com/articles/s41467-018-05257-7> (1 Aug. 2018).
68. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664. ISSN: 1088-9051. <http://genome.cshlp.org/lookup/doi/10.1101/gr.094052.109> (9 Sept. 2009).
69. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of Population Structure using Dense Haplotype Data. *PLOS Genetics* **8**, e1002453. ISSN: 1553-7404. <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002453> (1 Jan. 2012).
70. Cabrer0s, I. & Storey, J. D. A Likelihood-Free Estimator of Population Structure Bridging Admixture Models and Principal Components Analysis. *Genetics* **212**, 1009–1029. ISSN: 1943-2631. <https://academic.oup.com/genetics/article/212/4/1009/5931257> (4 Aug. 2019).
71. Mita, S. D. *et al.* Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology* **22**, 1383–1399. ISSN: 09621083. <https://onlinelibrary.wiley.com/doi/10.1111/mec.12182> (5 Mar. 2013).
72. Lotterhos, K. E. & Whitlock, M. C. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology* **23**, 2178–2192. ISSN: 1365-294X. <https://onlinelibrary.wiley.com/doi/full/10.1111/mec.12725>
<https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12725>
<https://onlinelibrary.wiley.com/doi/10.1111/mec.12725> (9 May 2014).
73. Szép, E., Sachdeva, H. & Barton, N. H. Polygenic local adaptation in metapopulations: A stochastic eco-evolutionary model. *Evolution; international journal of organic evolution* **75**, 1030–1045. ISSN: 1558-5646. <https://pubmed.ncbi.nlm.nih.gov/33742441/> (5 May 2021).
74. Lotterhos, K. E. & Whitlock, M. C. The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology* **24**, 1031–1046. ISSN: 09621083. <https://onlinelibrary.wiley.com/doi/10.1111/mec.13100> (5 Mar. 2015).

75. Sork, V. L. *et al.* Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genetics Genomes* **9**, 901–911. ISSN: 1614-2942. <http://link.springer.com/10.1007/s11295-013-0596-x> (4 Aug. 2013).
76. Gautier, M. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics* **201**, 1555–1579. ISSN: 1943-2631. <https://academic.oup.com/genetics/article/201/4/1555/5930067> (4 Dec. 2015).
77. Coop, G., Witonsky, D., Rienzo, A. D. & Pritchard, J. K. Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185**, 1411–1423. ISSN: 1943-2631. <https://pubmed.ncbi.nlm.nih.gov/20516501/> (4 Aug. 2010).
78. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS genetics* **8**. ISSN: 1553-7404. <https://pubmed.ncbi.nlm.nih.gov/23166502/> (11 Nov. 2012).
79. Murray, K. D. & Borevitz, J. O. Axe: rapid, competitive sequence read demultiplexing using a trie. *Bioinformatics* **34**, 3924–3925. ISSN: 1367-4803. <https://academic.oup.com/bioinformatics/article/34/22/3924/5026649> (22 Nov. 2018).
80. Bushnell, B. *BBMap* <https://sourceforge.net/projects/bbmap/>.
81. Choi, M. N. *et al.* The complete mitochondrial genome sequence of *Populus davidiana* Dode. *Mitochondrial DNA Part B* **2**, 113–114. ISSN: 2380-2359. <https://www.tandfonline.com/doi/full/10.1080/23802359.2017.1289346> (1 Jan. 2017).
82. Chung, H. Y., Lee, T.-H., Kim, Y.-K. & Kim, J. S. Complete chloroplast genome sequences of Wonwhang (*Pyrus pyrifolia*) and its phylogenetic analysis. *Mitochondrial DNA Part B* **2**, 325–326. ISSN: 2380-2359. <https://www.tandfonline.com/doi/full/10.1080/23802359.2017.1331328> (1 Jan. 2017).
83. Raymond, O. *et al.* The Rosa genome provides new insights into the domestication of modern roses. *Nature Genetics* **50**, 772–777. ISSN: 1061-4036. <http://www.nature.com/articles/s41588-018-0110-3> (6 June 2018).
84. Alexander, L. W. & Woeste, K. E. Pyrosequencing of the northern red oak (*Quercus rubra* L.) chloroplast genome reveals high quality polymorphisms for population management. *Tree Genetics Genomes* **10**, 803–812. ISSN: 1614-2942. <http://link.springer.com/10.1007/s11295-013-0681-1> (4 Aug. 2014).
85. Yang, Y. *et al.* Comparative Analysis of the Complete Chloroplast Genomes of Five *Quercus* Species. *Frontiers in Plant Science* **07**, 959. ISSN: 1664-462X. <http://journal.frontiersin.org/Article/10.3389/fpls.2016.00959/abstract> (June 2016).
86. Andrews, S. *FastQC A Quality Control tool for High Throughput Sequence Data* 2010. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
87. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158. ISSN: 1367-4811. <https://academic.oup.com/bioinformatics/article/27/15/2156/402296> (15 Aug. 2011).
88. R Core Team. *R: A Language and Environment for Statistical Computing* 2022. <https://www.R-project.org/>.

89. Pina-Martins, F., Silva, D. N., Fino, J. & Paulo, O. S. *Structure_threader*: An improved method for automation and parallelization of programs *Structure*, *fastStructure* and *MavericK* on multicore CPU systems. *Molecular Ecology Resources* **17**, e268–e274. ISSN: 1755098X. <https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.12702> (6 Nov. 2017).
90. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328. ISSN: 1367-4811. <https://academic.oup.com/bioinformatics/article/28/24/3326/245844> (24 Dec. 2012).
91. Jombart, T. *adeigenet*: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405. ISSN: 1367-4811. <https://academic.oup.com/bioinformatics/article/24/11/1403/191127> (11 June 2008).
92. Goudet, J. hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**, 184–186. ISSN: 1471-8278. <https://onlinelibrary.wiley.com/doi/10.1111/j.1471-8286.2004.00828.x> (1 Mar. 2005).
93. Morrison, M. L., Alcalá, N. & Rosenberg, N. A. FSTruct: An F_{ST} -based tool for measuring ancestry variation in inference of population structure. *Molecular Ecology Resources* **22**, 2614–2626. ISSN: 1755-098X. <https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13647> (7 Oct. 2022).
94. Karger, D. N. *et al.* Climatologies at high resolution for the earth's land surface areas. *Scientific Data* **4**, 170122. ISSN: 2052-4463. <https://www.nature.com/articles/sdata2017122> (1 Sept. 2017).
95. Brun, P., Zimmermann, N. E., Hari, C., Pellissier, L. & Karger, D. N. *CHELSEA-BIOCLIM+ A novel set of global climate-related predictors at kilometre-resolution* 2022. https://www.envodat.ch/dataset/bioclim_plus.
96. Brun, P., Zimmermann, N. E., Hari, C., Pellissier, L. & Karger, D. N. Global climate-related predictors at kilometer resolution for the past and future. *Earth System Science Data* **14**, 5573–5603. ISSN: 1866-3516. <https://essd.copernicus.org/articles/14/5573/2022/> (12 Dec. 2022).
97. QGIS Development Team. *QGIS Geographic Information System* 2022. <http://qgis.osgeo.org>.
98. Sork, V. L. *et al.* High-quality genome and methylomes illustrate features underlying evolutionary success of oaks. *Nature Communications* 2022 **13**, 1–15. ISSN: 2041-1723. <https://www.nature.com/articles/s41467-022-29584-y> (1 Apr. 2022).
99. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 1–9. ISSN: 14712105. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-421> (1 Dec. 2009).
100. Hahsler, M. & Nagar, A. rBLAST: R Interface for the Basic Local Alignment Search Tool. R package version 0.99.2. <https://github.com/mhahsler/rBLAST> (2019).
101. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* **47**, W191–W198. ISSN: 0305-1048. <https://academic.oup.com/nar/article/47/W1/W191/5486750> (W1 July 2019).

102. Guichoux, E. *et al.* Outlier loci highlight the direction of introgression in oaks. *Molecular Ecology* **22**, 450–462. ISSN: 1365-294X. <https://onlinelibrary.wiley.com/doi/full/10.1111/mec.12125> (2 Jan. 2013).
103. Michaud, H. *et al.* Effect of geographical discontinuity on genetic variation in *Quercus ilex* L. (holm oak). Evidence from enzyme polymorphism. *Heredity* **74**, 590–606. ISSN: 0018-067X. <http://www.nature.com/articles/hdy199583> (6 June 1995).
104. Hipp, A. L. *et al.* Genomic landscape of the global oak phylogeny. *New Phytologist* **226**, 1198–1212. ISSN: 0028-646X. <https://onlinelibrary.wiley.com/doi/10.1111/nph.16162> (4 May 2020).
105. Peñaloza-Ramírez, J. M. *et al.* Interspecific gene flow in a multispecies oak hybrid zone in the Sierra Tarahumara of Mexico. *Annals of Botany* **105**, 389–399. ISSN: 1095-8290. <https://academic.oup.com/aob/article-lookup/doi/10.1093/aob/mcp301> (3 Mar. 2010).
106. Albarrán-Lara, A. L. *et al.* Low genetic differentiation between two morphologically and ecologically distinct giant-leaved Mexican oaks. *Plant Systematics and Evolution* **305**, 89–101. ISSN: 0378-2697. <http://link.springer.com/10.1007/s00606-018-1554-8> (1 Jan. 2019).
107. Leroy, T. *et al.* Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. *New Phytologist* **226**, 1183–1197. ISSN: 1469-8137. <https://onlinelibrary.wiley.com/doi/full/10.1111/nph.16039> (4 May 2020).
108. Sousa, V. & Hey, J. Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics* **14**, 404–414. ISSN: 1471-0064. <https://www.nature.com/articles/nrg3446> (6 May 2013).
109. Petit, R. J. *et al.* Chloroplast DNA footprints of postglacial recolonization by oaks. *Proceedings of the National Academy of Sciences* **94**, 9996–10001. ISSN: 0027-8424. <https://pnas.org/doi/full/10.1073/pnas.94.18.9996> (18 Sept. 1997).
110. Manara, A., DalCorso, G. & Furini, A. The Role of the Atypical Kinases ABC1K7 and ABC1K8 in Absciscic Acid Responses. *Frontiers in Plant Science* **7**. ISSN: 1664-462X. <http://journal.frontiersin.org/Article/10.3389/fpls.2016.00366/abstract> (Mar. 2016).
111. Liu, J. *et al.* An H3K27me3 demethylase-HSFA2 regulatory loop orchestrates transgenerational thermomemory in *Arabidopsis*. *Cell Research* **29**, 379–390. ISSN: 1001-0602. <http://www.nature.com/articles/s41422-019-0145-8> (5 May 2019).
112. Zhang, X. *et al.* F-Box Proteins in Plants. *Annual Plant Reviews online* **2**, 307–328. ISSN: 26393832. <https://onlinelibrary.wiley.com/doi/10.1002/9781119312994.apr0701> (1 Feb. 2019).
113. Jakab, G. *et al.* Enhancing *Arabidopsis* Salt and Drought Stress Tolerance by Chemical Priming for Its Absciscic Acid Responses. *Plant Physiology* **139**, 267–274. ISSN: 1532-2548. <https://academic.oup.com/plphys/article/139/1/267/6113377> (1 Sept. 2005).
114. Larkindale, J., Hall, J. D., Knight, M. R. & Vierling, E. Heat Stress Phenotypes of *Arabidopsis* Mutants Implicate Multiple Signaling Pathways in the Acquisition of Thermotolerance. *Plant Physiology* **138**, 882–897. ISSN: 1532-2548. <https://academic.oup.com/plphys/article/138/2/882/6112929> (2 June 2005).

115. Cheng, X. *et al.* The trihelix family of transcription factors: Functional and evolutionary analysis in Moso bamboo (*Phyllostachys edulis*). *BMC Plant Biology* **19**, 1–20. ISSN: 14712229. <https://bmcpplantbiol.biomedcentral.com/articles/10.1186/s12870-019-1744-8> (1 Apr. 2019).
116. Magwanga, R. O. *et al.* Genome wide identification of the trihelix transcription factors and over-expression of *Gh_A05G2067* (*GT-2*), a novel gene contributing to increased drought and salt stresses tolerance in cotton. *Physiologia Plantarum* **167**, 447–464. ISSN: 0031-9317. <https://onlinelibrary.wiley.com/doi/10.1111/pp1.12920> (3 Nov. 2019).
117. Abdullah, M. *et al.* Genome-Wide Analysis Characterization and Evolution of SBP Genes in *Fragaria vesca*, *Pyrus bretschneideri*, *Prunus persica* and *Prunus mume*. *Frontiers in Genetics* **9**. ISSN: 1664-8021. <http://journal.frontiersin.org/article/10.3389/fgene.2018.00064/full> (Mar. 2018).
118. Lim, J., Lim, C. W. & Lee, S. C. Pepper Novel Serine-Threonine Kinase CaDIK1 Regulates Drought Tolerance via Modulating ABA Sensitivity. *Frontiers in Plant Science* **11**, 1133. ISSN: 1664-462X. <https://www.frontiersin.org/article/10.3389/fpls.2020.01133/full> (July 2020).
119. Sun, Y. *et al.* Genome-Wide Analysis of the C2 Domain Family in Soybean and Identification of a Putative Abiotic Stress Response Gene GmC2-148. *Frontiers in Plant Science* **12**, 19. ISSN: 1664-462X. <https://www.frontiersin.org/articles/10.3389/fpls.2021.620544/full> (Feb. 2021).
120. Watanabe, S., Nakagawa, A., Izumi, S., Shimada, H. & Sakamoto, A. RNA interference-mediated suppression of xanthine dehydrogenase reveals the role of purine metabolism in drought tolerance in *Arabidopsis*. *FEBS letters* **584**, 1181–1186. ISSN: 1873-3468. <https://pubmed.ncbi.nlm.nih.gov/20153325/> (6 Mar. 2010).
121. Qian, Y. *et al.* Regulatory Mechanisms of bHLH Transcription Factors in Plant Adaptive Responses to Various Abiotic Stresses. *Frontiers in Plant Science* **12**, 1143. ISSN: 1664-462X. <https://www.frontiersin.org/articles/10.3389/fpls.2021.677611/full> (June 2021).
122. Curtis, T. Y., Bo, V., Tucker, A. & Halford, N. G. Construction of a network describing asparagine metabolism in plants and its application to the identification of genes affecting asparagine metabolism in wheat under drought and nutritional stress. *Food and Energy Security* **7**, e00126. ISSN: 20483694. <https://onlinelibrary.wiley.com/doi/10.1002/fes3.126> (1 Feb. 2018).
123. Niño, M. C., Kim, M.-S., Kang, K. K. & Cho, Y.-G. Genome-wide identification and molecular characterization of cysteine protease genes in rice. *Plant Biotechnology Reports* **14**, 69–87. ISSN: 1863-5466. <http://link.springer.com/10.1007/s11816-019-00583-8> (1 Feb. 2020).
124. Kidrič, M., Kos, J. & Sabotič, J. Proteases and their endogenous inhibitors in the plant response to abiotic stress. *Botanica SERBICA* **38**, 139–158. www.chem.qmul.ac.uk/iubmb (1 Jan. 2014).

Supplementary Material

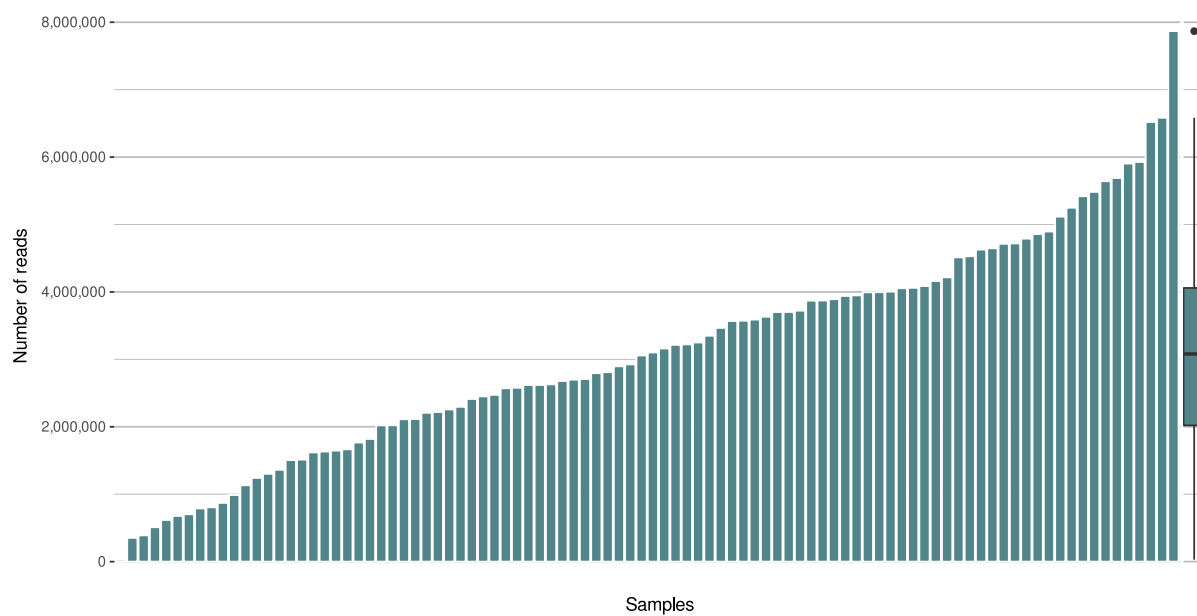


Figure S1: Distribution of the number of raw reads per sample after the GBS sequencing protocol. The total number of reads was 293,145,449. The mean number of reads per sample was 3,118,569.

Table S1: Identification, country of origin, sampling location name, and geographical coordinates for the 94 samples used in this study. Bold Sample IDs indicate samples removed due to missing data.

Sample ID	Country	Sampling location	Latitude (Decimal degrees)	Longitude (Decimal degrees)
AZO_153	Morocco	Azrou	33.5417362	-5.3153128
AZO_154	Morocco	Azrou	33.4295518	-5.2032826
AZO_155	Morocco	Azrou	33.4348333	-5.1843198
AZO_158	Morocco	Azrou	33.420788	-5.1671908
AZO_160	Morocco	Azrou	33.4826218	-5.1521205
AZO_161	Morocco	Azrou	33.493777	-5.1481254
BAR_004	Portugal	Barrancos	38.0976991	-7.2083361
BAR_006	Portugal	Barrancos	38.088715	-7.1918516
BAR_013	Portugal	Barrancos	38.113327	-7.0441189
BAR_015	Portugal	Barrancos	38.1232201	-7.0190476
BAR_017	Portugal	Barrancos	38.1235343	-6.9988389
BAR_019	Portugal	Barrancos	38.1321876	-6.9829109
CAL_021	Portugal	Caldeirão	37.3010447	-8.0228781
CAL_025	Portugal	Caldeirão	37.3142209	-8.0131762
CAL_027	Portugal	Caldeirão	37.301664	-7.9504139
CAL_034	Portugal	Caldeirão	37.1819522	-7.9398292
CAL_040	Portugal	Caldeirão	37.1743712	-7.9274833
CAT_122	Spain	Catalonia	42.3946285	2.8389134
CAT_123	Spain	Catalonia	42.4003249	2.8304845
CAT_124	Spain	Catalonia	42.4060748	2.8142974
CAT_126	Spain	Catalonia	42.4081897	2.7990403
CAT_130	Spain	Catalonia	42.4224049	2.8012565
COR_683	France	Corsica	41.7753869	8.901265
COR_685	France	Corsica	41.7745493	8.8805302
COR_686	France	Corsica	41.7544299	8.8568312
COR_688	France	Corsica	41.7742222	8.9012563
COR_690	France	Corsica	41.7644434	8.9028197
CIR_062	Spain	Cuidad Rodrigo	40.6069012	-6.4431163
CIR_067	Spain	Cuidad Rodrigo	40.5954904	-6.4049656
CIR_070	Spain	Cuidad Rodrigo	40.5784072	-6.4048942
CIR_073	Spain	Cuidad Rodrigo	40.5452028	-6.3897072
CIR_076	Spain	Cuidad Rodrigo	40.5057838	-6.366327
FUE_101	Spain	Fuente Caliente	38.4415738	-4.3161179
FUE_102	Spain	Fuente Caliente	38.4262685	-4.310854
FUE_103	Spain	Fuente Caliente	38.4243455	-4.3016393
FUE_106	Spain	Fuente Caliente	38.4257821	-4.2926764
FUE_110	Spain	Fuente Caliente	38.3990189	-4.2997477
GRE_111	Spain	Gredos	40.1157697	-4.689008
GRE_112	Spain	Gredos	40.1189338	-4.7125329
GRE_116	Spain	Gredos	40.1090127	-4.7212484
GRE_118	Spain	Gredos	40.0952886	-4.7254632
GRE_120	Spain	Gredos	40.0952071	-4.6796494
HZL_093	Spain	Haza del Lino	36.80042	-3.2985612
HZL_095	Spain	Haza del Lino	36.8035274	-3.2922325
HZL_097	Spain	Haza del Lino	36.8076982	-3.3038163
HZL_098	Spain	Haza del Lino	36.8094546	-3.3062008
HZL_100	Spain	Haza del Lino	36.823739	-3.2947736

Table S1: (cont.)

Sample ID	Country	Sampling location	Latitude (Decimal degrees)	Longitude (Decimal degrees)
RIF_162	Morocco	Rif	35.072928	-5.322744
RIF_163	Morocco	Rif	35.0366419	-5.361306
RIF_164	Morocco	Rif	35.0256431	-5.3870468
RIF_165	Morocco	Rif	35.0396275	-5.4622251
RON_081	Spain	Ronda	36.6973731	-5.2618628
RON_083	Spain	Ronda	36.659107	-5.2849854
RON_085	Spain	Ronda	36.6304484	-5.3320357
RON_087	Spain	Ronda	36.5993561	-5.3617616
RON_088	Spain	Ronda	36.5247917	-5.3645072
RON_090	Spain	Ronda	36.5293488	-5.3578694
SAR_611	Italy	Sardinia	39.30514	9.39325
SAR_612	Italy	Sardinia	39.30844	9.39797
SAR_613	Italy	Sardinia	39.09328	8.78947
SAR_614	Italy	Sardinia	39.08669	8.79547
SAR_617	Italy	Sardinia	40.88175	9.03483
TAZ_141	Morocco	Taza	34.1157537	-4.2950763
TAZ_144	Morocco	Taza	34.0621733	-4.2639979
TAZ_146	Morocco	Taza	34.0539858	-4.1662072
TAZ_149	Morocco	Taza	34.0625935	-4.1339634
TAZ_150	Morocco	Taza	34.0674635	-4.1236426
TOM_041	Portugal	Trás-os-Montes	41.5179201	-7.0271724
TOM_047	Portugal	Trás-os-Montes	41.5301854	-7.0479675
TOM_050	Portugal	Trás-os-Montes	41.5121427	-7.0898351
TOM_051	Portugal	Trás-os-Montes	41.4901549	-7.1022524
TOM_055	Portugal	Trás-os-Montes	41.4724078	-7.1011956
TOM_060	Portugal	Trás-os-Montes	41.4598431	-7.0470636
TUN_B1	Tunisia	Tunisia	36.361533	10.102767
TUN_B3	Tunisia	Tunisia	36.359883	10.1029
TUN_B4	Tunisia	Tunisia	36.359117	10.103367
TUN_B5	Tunisia	Tunisia	36.359	10.10235
TUN_B8	Tunisia	Tunisia	36.352633	10.117233
TUS_001	Italy	Tuscany	42.397779	11.906297
TUS_003	Italy	Tuscany	42.4151674	11.8932853
TUS_004	Italy	Tuscany	42.4581099	11.7385
TUS_005	Italy	Tuscany	42.4578433	11.7333082
TUS_008	Italy	Tuscany	42.4631939	11.7490059
VAR_009	France	Var	43.2521056	6.4025379
VAR_010	France	Var	43.2484542	6.3791808
VAR_012	France	Var	43.2445191	6.3714762
VAR_013	France	Var	43.2462416	6.3679384
VAR_015	France	Var	43.2368768	6.3342056
ZAR_131	Spain	Zaragoza	41.390979	-1.5512734
ZAR_133	Spain	Zaragoza	41.3996836	-1.5353276
ZAR_134	Spain	Zaragoza	41.4065805	-1.5234337
ZAR_136	Spain	Zaragoza	41.4129826	-1.5196421
ZAR_137	Spain	Zaragoza	41.3981384	-1.4859897
ZAR_140	Spain	Zaragoza	41.3893705	-1.4741743

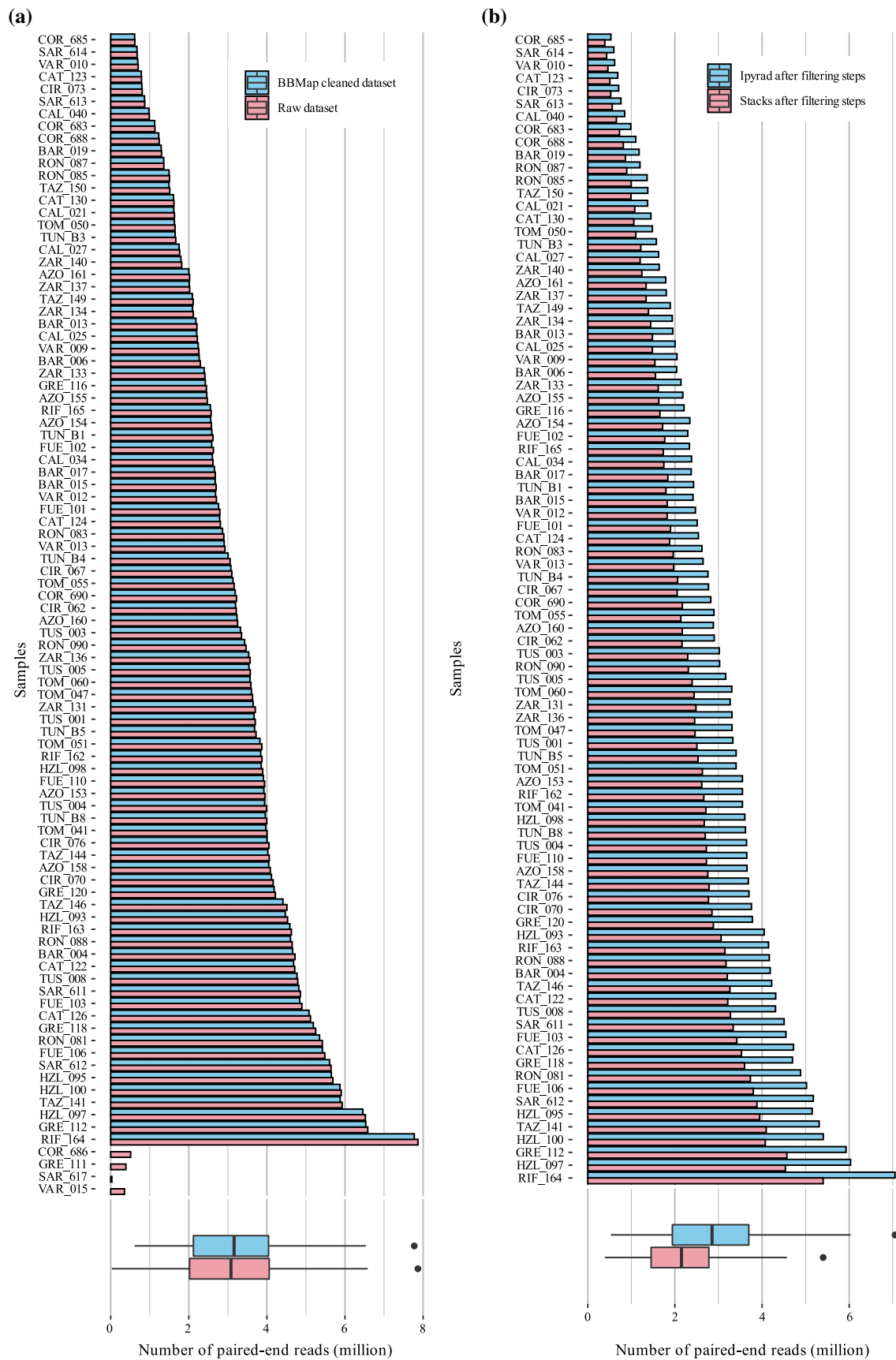


Figure S2: Distribution of the number of paired-end reads per sample for (a) the raw dataset after demultiplexation and after the removal of mitochondrial and chloroplastidial reads with *BMap*, and (b) after *ipyrad* and *Stacks* filtering steps.

Table S2: Results for *ipyrad* assemblies run with different values for the *clustering_threshold* parameter. Represented for all-SNPs and center-SNPs datasets are the number of variable sites directly out of *ipyrad* (and *vcf_parser.py* for center-SNP datasets) and after *VCFTools* filtering, and the number of samples remaining in the datasets after removing those with >65% of missing data.

Parameter (<i>Clustering_threshold</i>)	Nr variable sites out of <i>ipyrad</i>		Nr variable sites after <i>VCFTools</i>		Remaining samples	
	All-SNPs	Center-SNP	All-SNPs	Center-SNP	All-SNPs	Center-SNP
0.85	235.964	22.686	19.652	1.921	90	90
0.90	352.915	31.942	23.518	2.098	89	90
0.95	504.593	52.380	26.560	2.305	89	90

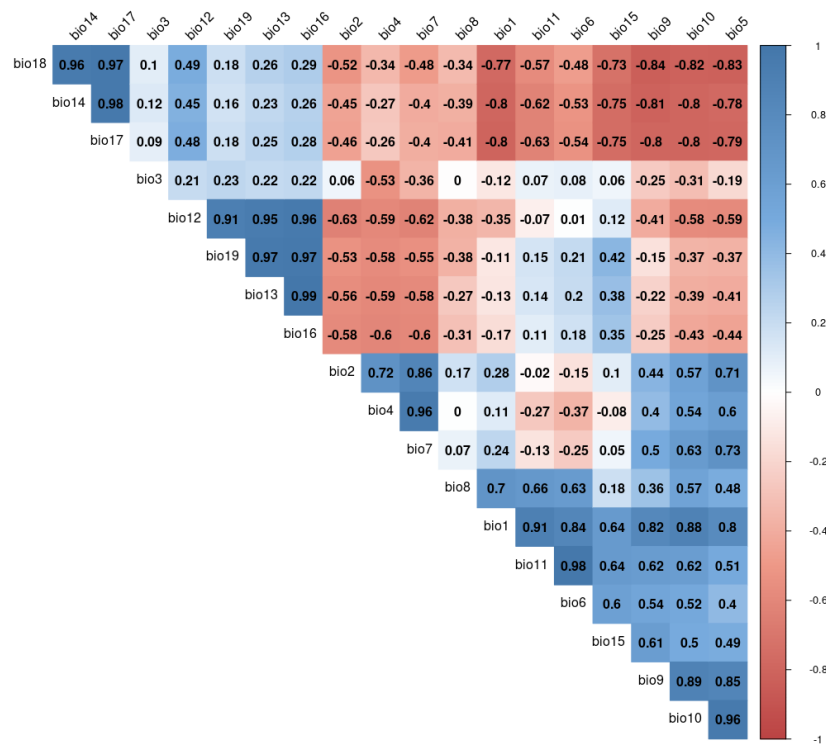
Table S3: Results for *Stacks* assemblies run with different values for parameters *-m*, *-M*, *-n*, and *-R*. Represented for all-SNPs and center-SNPs datasets are the number of variable sites directly out of *ipyrad* (and *vcf_parser.py* for center-SNP datasets) and after *VCFTools* filtering, and the number of samples remaining in the datasets after removing those with >65% of missing data.

Parameters				Nr variable sites before filter		Nr variable sites after filter		Remaining samples	
<i>-m</i>	<i>-M</i>	<i>-n</i>	<i>-R</i>	All-SNPs	Random-SNP	All-SNPs	Random-SNP	All-SNPs	Random-SNP
3	2	1	0.09	1.012.123	105.879	58.887	2.993	90	90
3	10	10	0.09	1.215.862	71.350	77.050	3.202	90	90
5	10	10	0.09	769.788	47.478	42.237	2.494	86	86
8	2	1	0.09	487.561	49.628	10.584	522	83	83
8	5	5	0.09	562.885	39.581	19.242	455	83	83
8	10	10	0.09	466.769	31.010	8.693	393	82	82

Table S4: List of bioclimatic variables kept for each geographical area after the correlation analyses. Represented for each variable are the short name, full name, unit, and a brief explanation.

Area	Shortname	Longname	Unit	Explanation
Iberian Peninsula	bio1	mean annual air temperature	°C	Mean annual daily mean air temperatures averaged over 1 year
	bio2	mean diurnal air temperature range	°C	Mean diurnal range of temperatures averaged over 1 year
	bio3	isothermality	°C	Ratio of diurnal variation to annual variation in temperatures
	bio4	temperature seasonality	°C/100	Standard deviation of the monthly mean temperatures
	bio6	mean daily minimum air temperature of the coldest month	°C	The lowest temperature of any monthly daily mean maximum temperature
	bio8	mean daily mean air temperatures of the wettest quarter	°C	The wettest quarter of the year is determined (to the nearest month)
	bio9	mean daily mean air temperatures of the driest quarter	°C	The driest quarter of the year is determined (to the nearest month)
	bio10	mean daily mean air temperatures of the warmest quarter	°C	The warmest quarter of the year is determined (to the nearest month)
	bio12	annual precipitation amount	kg m ⁻²	Accumulated precipitation amount over 1 year
	bio15	precipitation seasonality	kg m ⁻²	The Coefficient of Variation is the standard deviation of the monthly precipitation estimates expressed as a percentage of the mean of those estimates (i.e. the annual mean)
Iberian Peninsula and Morocco	bio17	mean monthly precipitation amount of the driest quarter	kg m ⁻²	The driest quarter of the year is determined (to the nearest month)
	bio1	Mean annual air temperature	°C	Mean annual daily mean air temperatures averaged over 1 year
	bio2	mean diurnal air temperature range	°C	Mean diurnal range of temperatures averaged over 1 year
	bio3	isothermality	°C	Ratio of diurnal variation to annual variation in temperatures
	bio4	temperature seasonality	°C/100	Standard deviation of the monthly mean temperatures
	bio6	mean daily minimum air temperature of the coldest month	°C	The lowest temperature of any monthly daily mean maximum temperature
	bio7	annual range of air temperature	°C	The difference between the Maximum Temperature of Warmest month and the Minimum Temperature of Coldest month
	bio8	mean daily mean air temperatures of the wettest quarter	°C	The wettest quarter of the year is determined (to the nearest month)
	bio10	mean daily mean air temperatures of the warmest quarter	°C	The warmest quarter of the year is determined (to the nearest month)
	bio12	annual precipitation amount	kg m ⁻²	Accumulated precipitation amount over 1 year
	bio15	precipitation seasonality	kg m ⁻²	The Coefficient of Variation is the standard deviation of the monthly precipitation estimates expressed as a percentage of the mean of those estimates (i.e. the annual mean)
	bio17	mean monthly precipitation amount of the driest quarter	kg m ⁻²	The driest quarter of the year is determined (to the nearest month)

(a) Iberia



(b) Iberia and Morocco

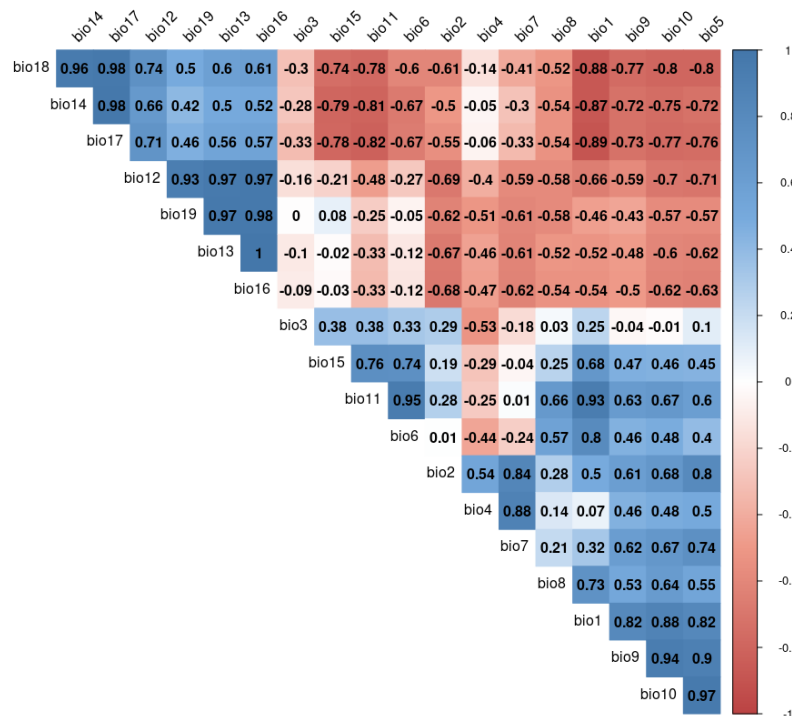


Figure S3: Spearman's correlation matrices for the 19 bioclimatic variables, calculated for the geographic areas of (a) Iberia and (b) Iberia and Morocco. Color corresponds to the strength of the correlation.

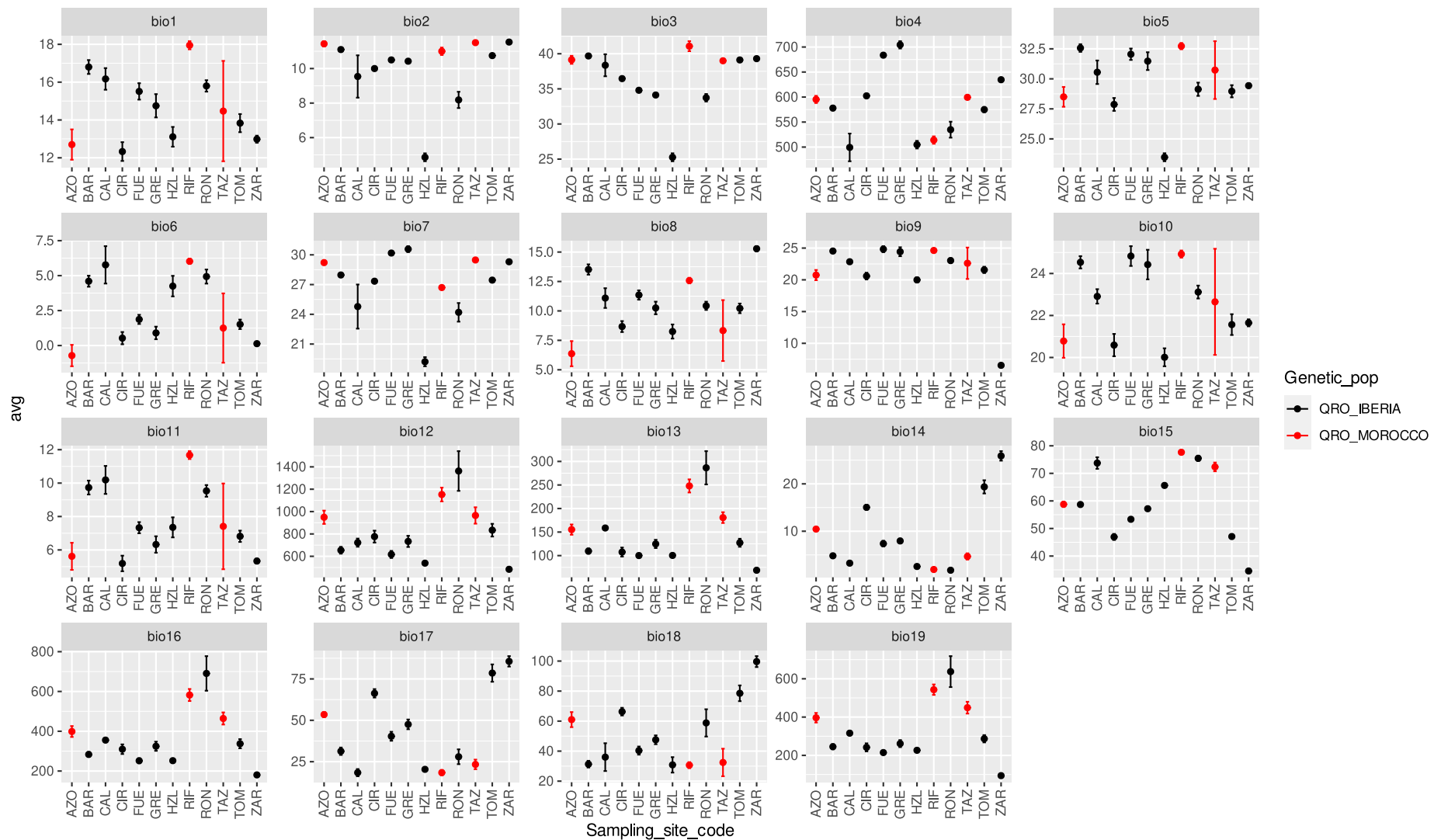


Figure S4: Environmental values for all 19 bioclimatic variables. The average environmental value and standard deviation are presented for each *Q. rotundifolia* sampling location. Sampling locations are colored according to their genetic population (QRO_IBERIA - Iberian *Q. rotundifolia*; QRO_MOROCCO - Moroccan *Q. rotundifolia*). The bioclimatic variables adopted for the genetic-environmental analyses were bio1, bio2, bio3, bio4, bio6, bio7, bio8, bio9, bio10, bio12, bio15, and bio17. Bioclimatic variable names can be consulted in Supplementary Table S4.

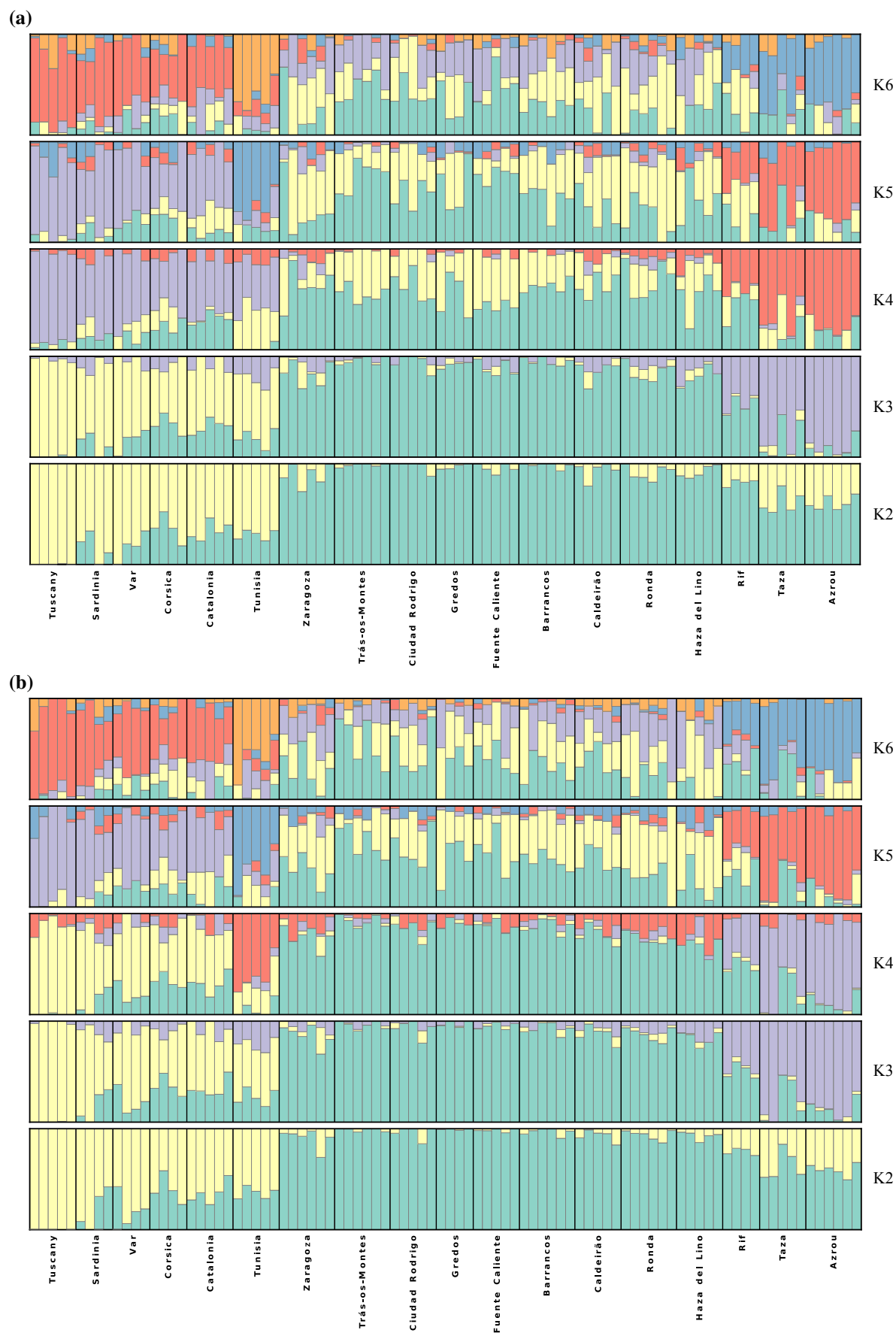


Figure S5: Full *ALStructure* results for K2 - K6 for datasets (a) *ipr_centersnp_ib* and (b) *stk_centersnp_ib*. Samples are organized by sampling location.

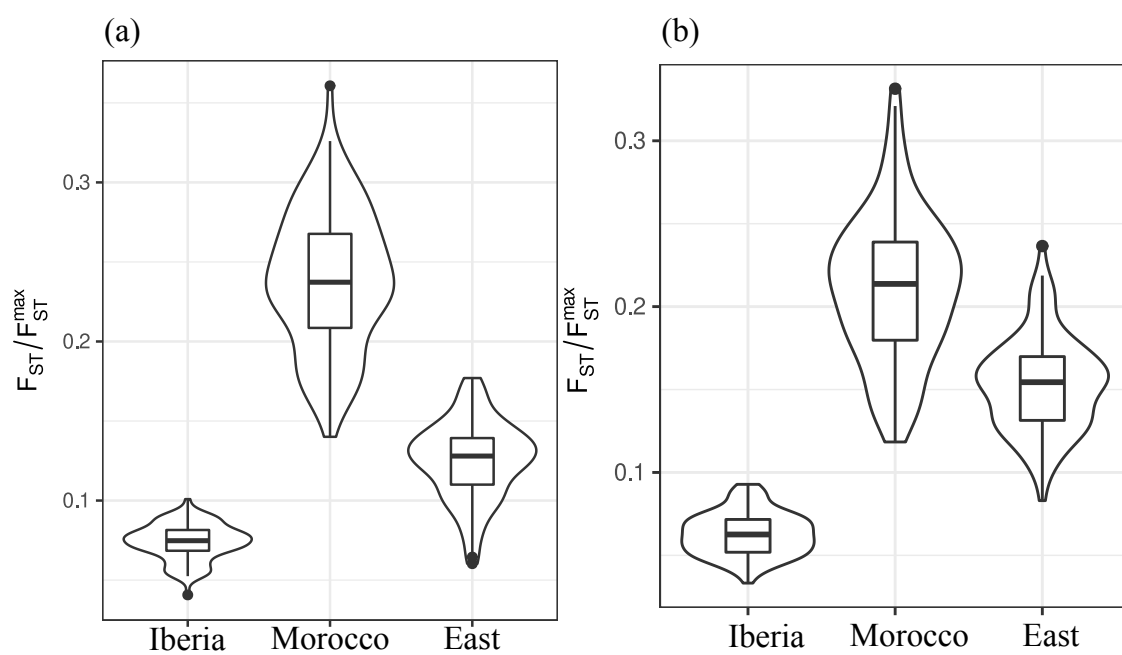


Figure S6: Bootstrap distributions of F_{ST}/F_{ST}^{max} values for each population, for datasets (a) *ipr.centersnp_ib* and (b) *stk.centersnp_ib*.

Table S5: Pairwise Fst values calculated for datasets 'ipr centersnp allsamps' (below diagonal) and 'stk centersnp allsamps' (above diagonal). Fst values were calculated pairwise for all sampling locations. Sampling location codes as in Table 2.1.

	AZO	BAR	CAL	CAT	CIR	COR	FUE	GRE	HZL	RIF	RON	SAR	TAZ	TOM	TUN	TUS	VAR	ZAR
AZO		0.115	0.130	0.112	0.131	0.134	0.119	0.125	0.105	0.050	0.113	0.193	0.004	0.133	0.098	0.212	0.139	0.110
BAR	0.134		-0.009	0.114	0.001	0.137	-0.006	-0.010	0.010	0.037	-0.004	0.237	0.093	-0.003	0.116	0.252	0.165	-0.006
CAL	0.111	-0.008		0.124	-0.001	0.142	0.004	0.001	0.018	0.045	0.003	0.254	0.100	0.005	0.113	0.267	0.173	0.007
CAT	0.133	0.136	0.111		0.123	0.025	0.115	0.117	0.108	0.089	0.108	0.038	0.103	0.130	0.040	0.065	0.009	0.089
CIR	0.14	0.005	-0.005	0.143		0.149	0.005	0.000	0.016	0.046	0.011	0.247	0.101	0.005	0.120	0.261	0.169	0.003
COR	0.141	0.15	0.113	0.031	0.15		0.148	0.142	0.146	0.110	0.139	0.052	0.119	0.162	0.048	0.059	-0.017	0.121
FUE	0.113	0.004	0.002	0.119	0.005	0.12		-0.005	0.011	0.045	0.009	0.242	0.094	0.007	0.116	0.254	0.171	0.007
GRE	0.119	0.002	-0.005	0.121	0.01	0.119	-0.004		0.008	0.041	0.001	0.245	0.095	0.006	0.114	0.261	0.173	-0.001
HZL	0.097	0.007	0.004	0.11	0.008	0.129	0.001	0.003		0.042	0.008	0.229	0.079	0.024	0.103	0.245	0.163	0.005
RIF	0.042	0.039	0.022	0.109	0.041	0.122	0.027	0.029	0.017		0.042	0.192	0.023	0.054	0.076	0.207	0.127	0.038
RON	0.11	0.012	-0.008	0.109	0.004	0.12	0.013	0.008	-0.003	0.036		0.229	0.085	0.008	0.103	0.246	0.167	0.000
SAR	0.185	0.237	0.216	0.037	0.233	0.062	0.215	0.222	0.201	0.188	0.202		0.182	0.253	0.079	0.020	0.010	0.210
TAZ	0.009	0.109	0.09	0.124	0.117	0.134	0.091	0.096	0.07	0.034	0.09	0.181		0.105	0.080	0.207	0.129	0.080
TOM	0.129	-0.001	0.003	0.136	0.002	0.153	0.007	0.008	0.02	0.049	0.014	0.237	0.116		0.125	0.267	0.181	0.009
TUN	0.111	0.138	0.123	0.047	0.141	0.053	0.121	0.116	0.106	0.098	0.114	0.066	0.1	0.14		0.087	0.049	0.093
TUS	0.209	0.252	0.244	0.054	0.259	0.073	0.234	0.238	0.22	0.201	0.222	0.028	0.205	0.252	0.073		0.023	0.227
VAR	0.161	0.192	0.17	0.027	0.19	0.023	0.168	0.166	0.154	0.151	0.16	0.032	0.153	0.18	0.056	0.037		0.144
ZAR	0.101	0.003	-0.002	0.09	0.003	0.113	0.005	0.006	0.007	0.019	0.003	0.186	0.082	0.013	0.095	0.201	0.14	

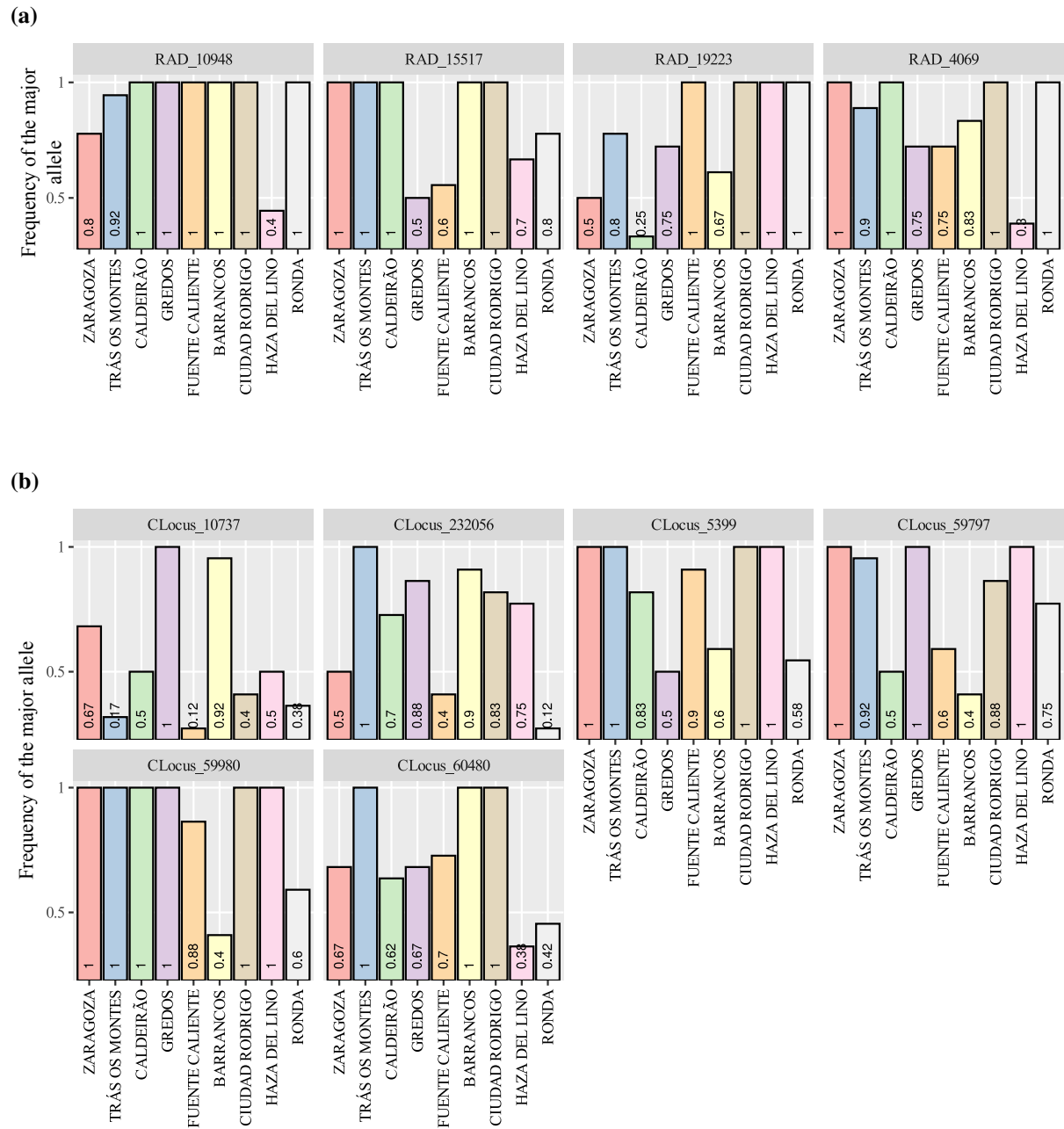


Figure S7: Frequency of the major allele for the outlier SNPs detected by *BayPass* core model mode, for datasets (a) *ipr_centersnp_ib* and (b) *stk_centersnp_ib*.

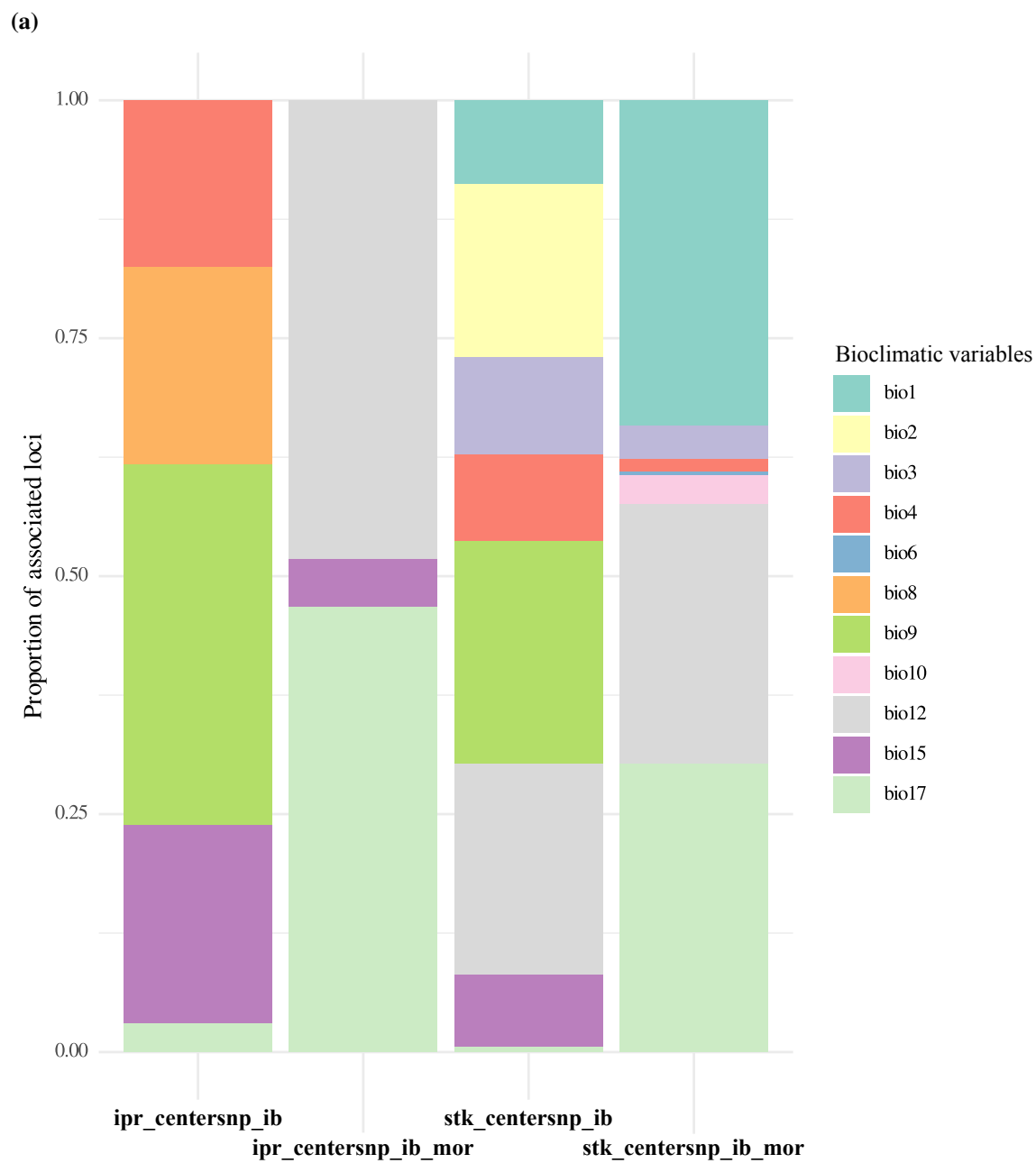


Figure S8: Proportion of associated SNPs (a) per bioclimatic variable, per dataset, and (b) per bioclimatic variable, per dataset, per strength of association, for datasets ipr_centersnp_ib, ipr_centersnp_ib_mor, stk_centersnp_ib, and stk_centersnp_ib_mor

(b)

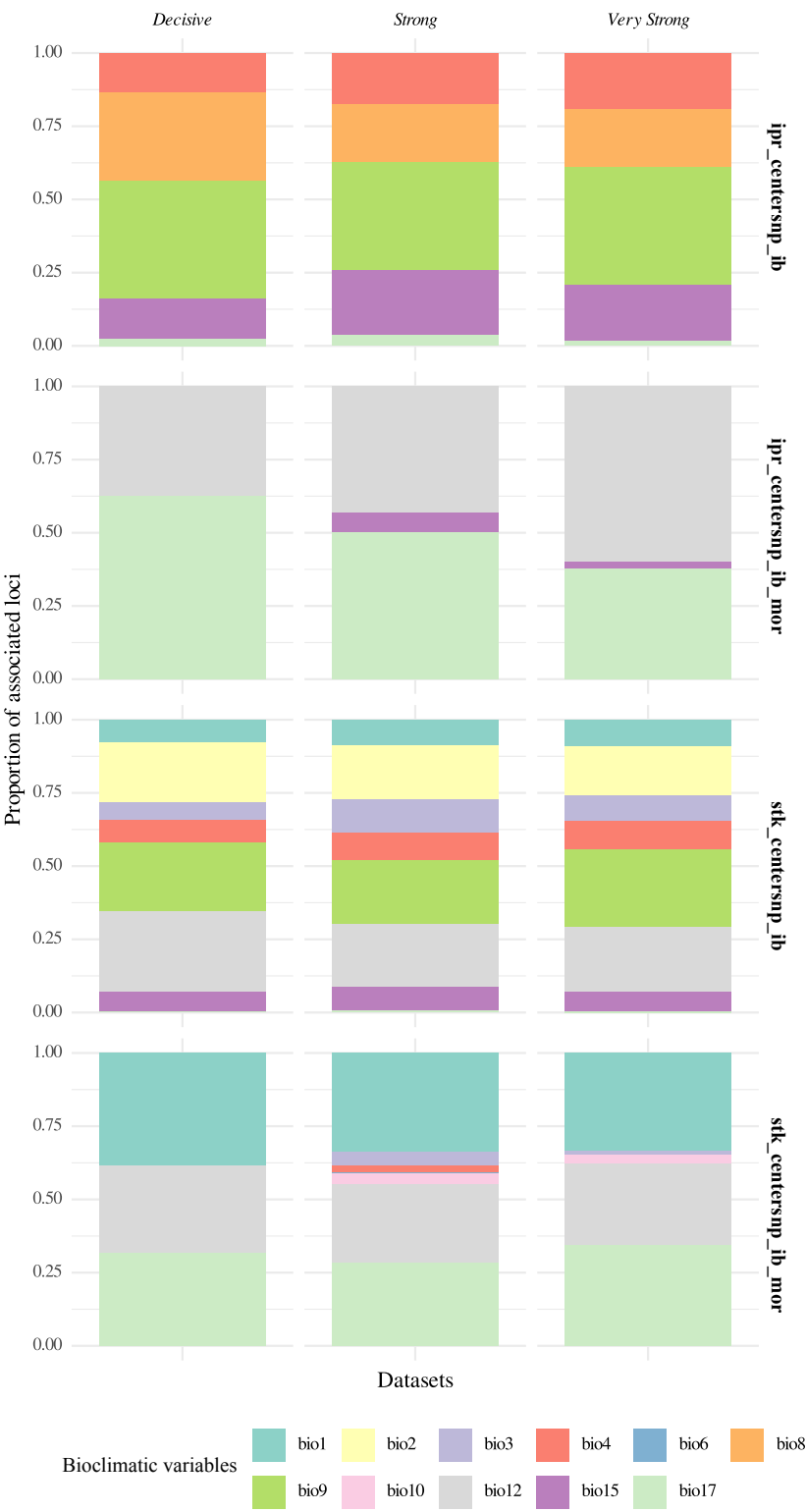


Figure S8: *cont.*

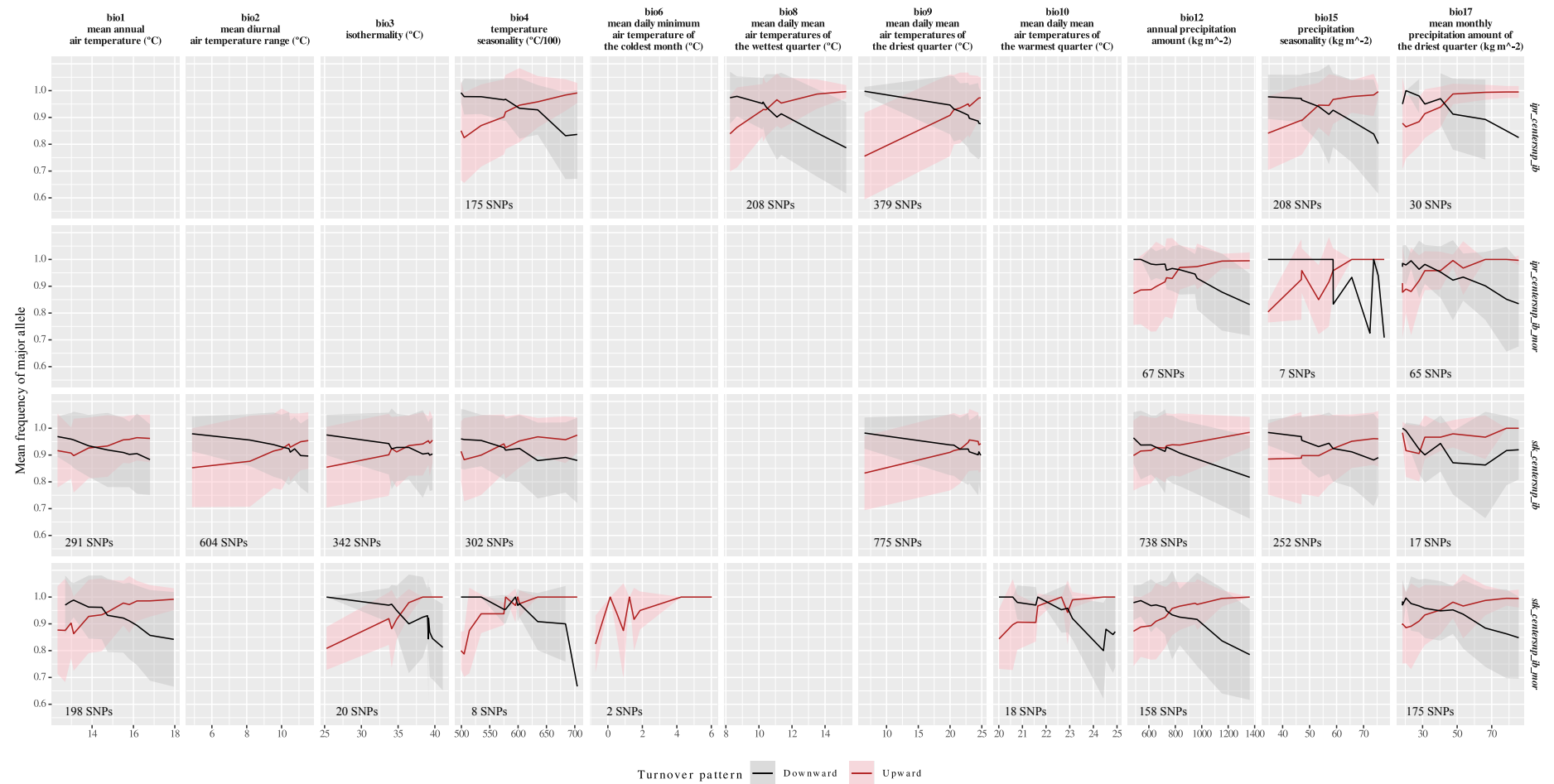


Figure S9: Average pattern of allele turnover along all associated bioclimatic variables, for the datasets *ipr_centersnp_ib*, *stk_centersnp_ib*, *ipr_centersnp_ib_mor*, and *stk_centersnp_ib_mor*. Each SNP associated with an environmental variable, for all levels of association strength, was classified according to its pattern of allele turnover (upward pattern, a tendency of allele frequencies to be fixed above turnover threshold; downward pattern, a tendency of allele frequencies to be fixed below turnover threshold). The average allele frequency of the major alleles was calculated for SNPs in each turnover pattern, for each sampling location/environmental value. The number of associated SNPs is represented for each dataset/bioclimatic variable. Colored lines represent the average allele frequencies along the bioclimatic variable. Standard deviation is represented by the colored shadow.

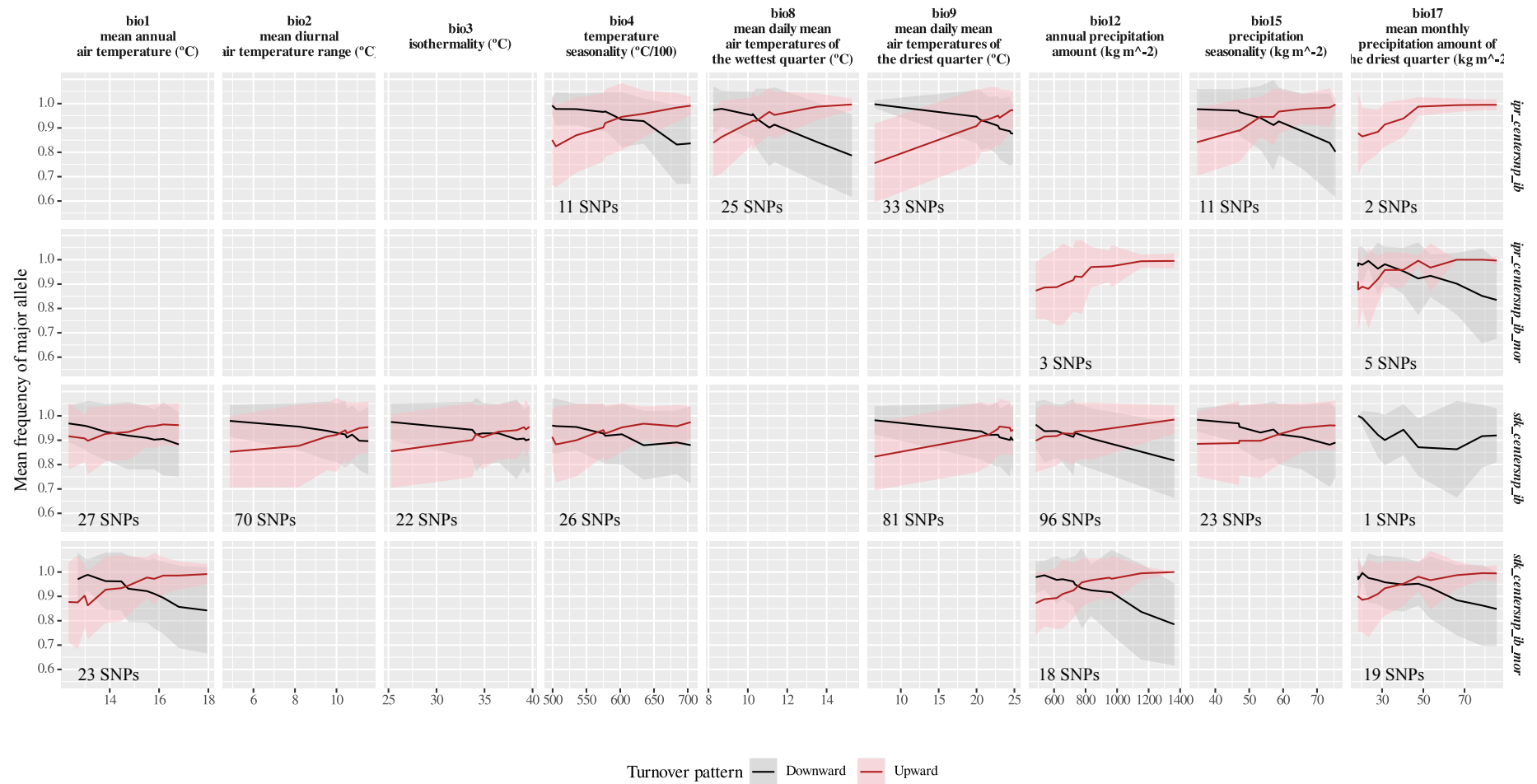


Figure S10: Average pattern of allele turnover along associated bioclimatic variables, for the datasets *ipr_centersnp_ib*, *stk_centersnp_ib*, *ipr_centersnp_ib_mor*, and *stk_centersnp_ib_mor*. Each SNP associated with an environmental variable, for the decisive level of association strength, was classified according to its pattern of allele turnover (upward pattern, a tendency of allele frequencies to be fixed above turnover threshold; downward pattern, a tendency of allele frequencies to be fixed below turnover threshold). The average allele frequency of the major alleles was calculated for SNPs in each turnover pattern, for each sampling location/environmental value. The number of associated SNPs is represented for each dataset/bioclimatic variable. Colored lines represent the average allele frequencies along the bioclimatic variable. Standard deviation is represented by the colored shadow.

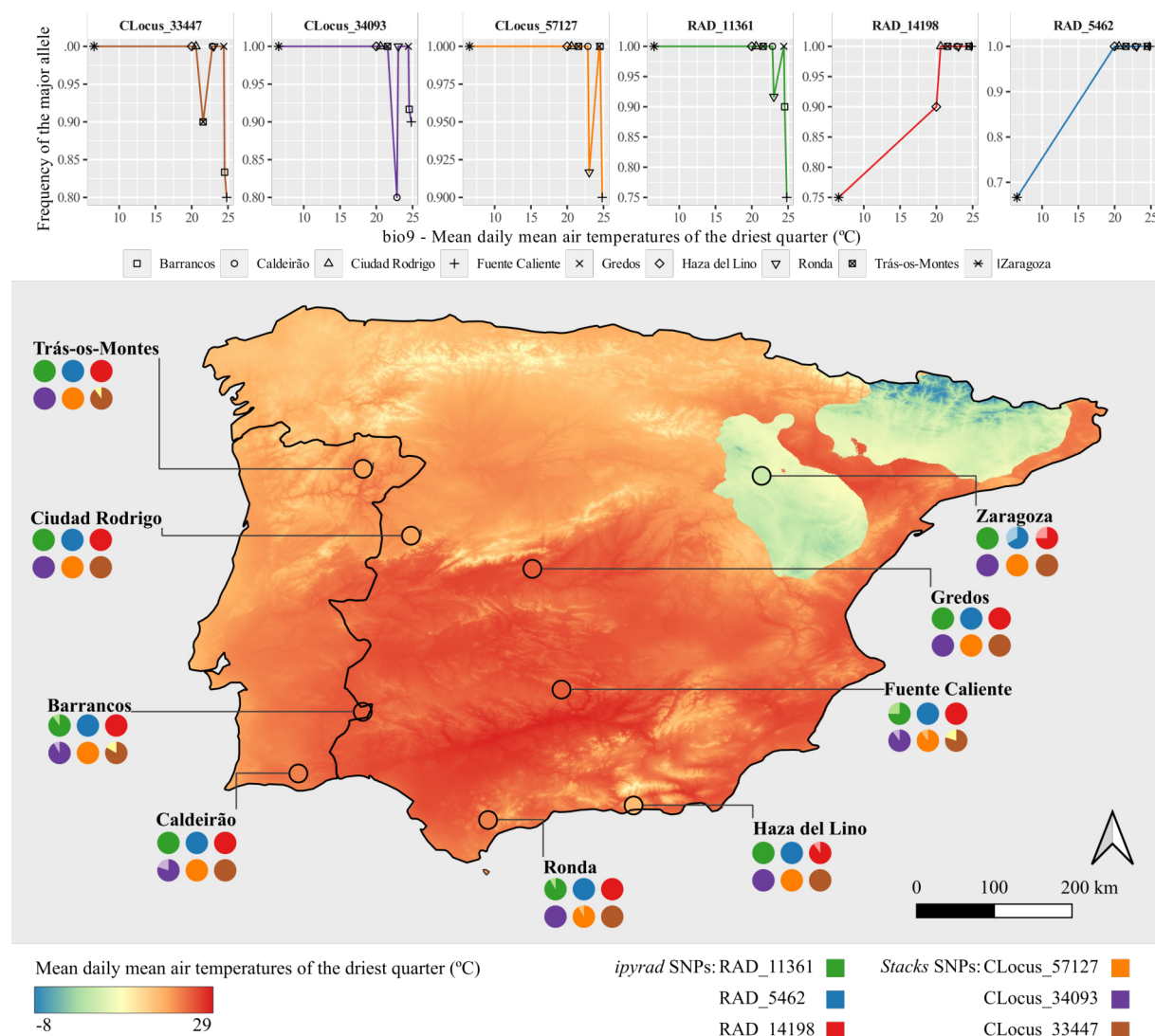


Figure S11: Representation of allele frequency turnover for loci associated with bioclimatic variable bio9 (mean daily mean air temperatures of the driest quarter). The three strongest associated loci are shown for datasets `ipr.centersnp_ib` and `stk.centersnp_ib`. Allele frequencies for each sampling location are plotted as a function of mean daily mean air temperature values (above). The scale of the yy axis may differ across plots. Allele frequencies for each sampling location are represented along the geographical area of the Iberian Peninsula (below). The range of mean daily mean air temperature is represented by a color scale. Values for each sampling location are represented above their geographical position.

Table S6: Summary of outlier SNP/loci hits against the reference genome of *Q. lobata* for datasets `'ipr.centersnp_ib_mor'` and `'stk.centersnp_ib_mor'`.

Dataset	SNP	<i>Q. lobata</i> gene
ipr.centersnp_ib_mor	RAD.3663	Nucleotid.trans domain-containing protein
ipr.centersnp_ib_mor	RAD.19905	ABC1 domain-containing protein
stk.centersnp_ib_mor	CLocus.7350	Protein SUPPRESSOR OF GENE SILENCING 3-like
stk.centersnp_ib_mor	CLocus.8084	DNA oxidative demethylase
stk.centersnp_ib_mor	CLocus.11623	Pentatricopeptide repeat-containing protein

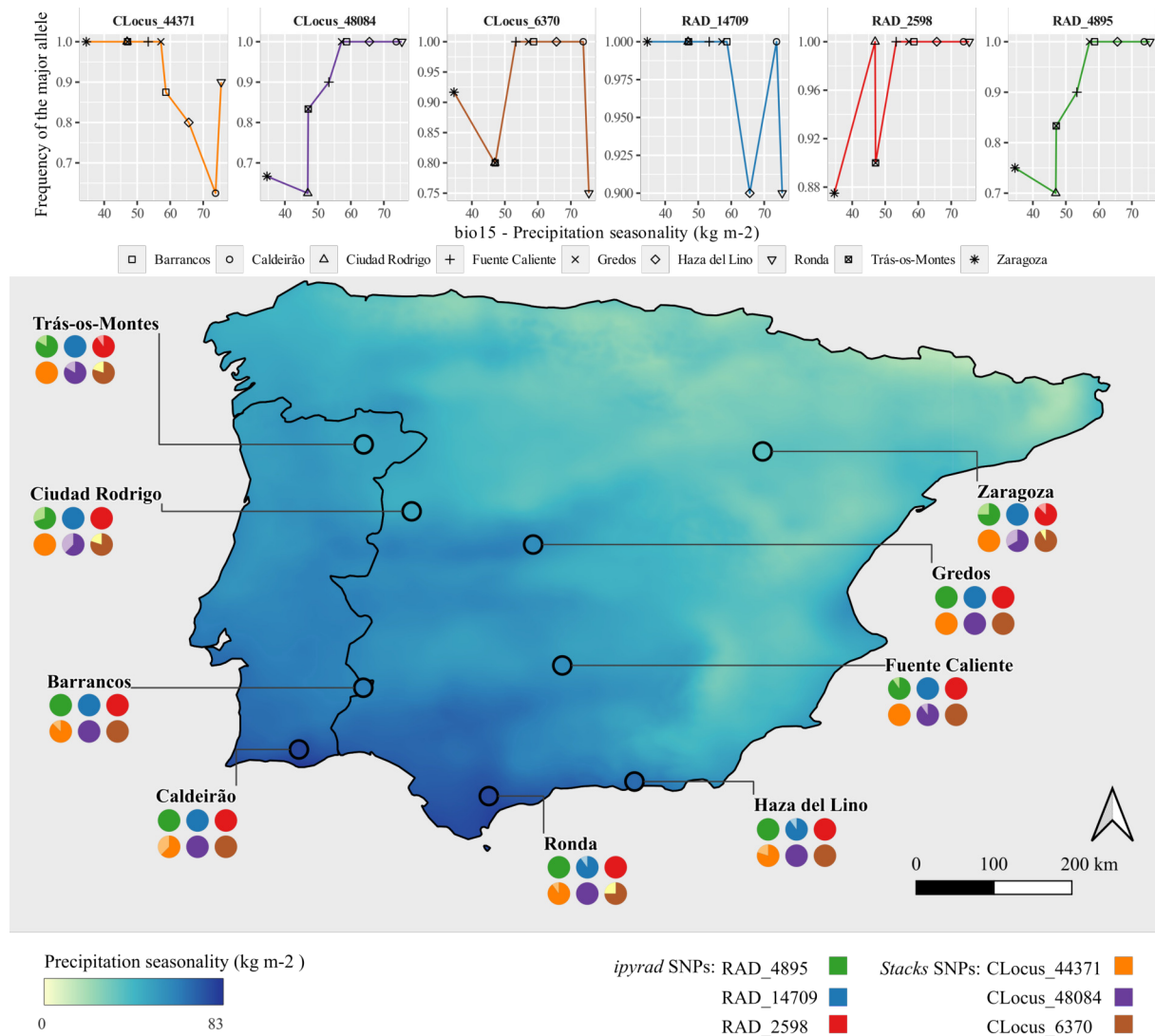


Figure S12: Representation of allele frequency turnover for loci associated with bioclimatic variable bio15 (precipitation seasonality). The three strongest associated loci are shown for datasets *ipr_centersnp_ib* and *stk_centersnp_ib*. Allele frequencies for each sampling location are plotted as a function of precipitation seasonality values (above). The scale of the yy axis may differ across plots. Allele frequencies for each sampling location are represented along the geographical area of the Iberian Peninsula (below). The range of precipitation seasonality is represented by a color scale. Values for each sampling location are represented above their geographical position.

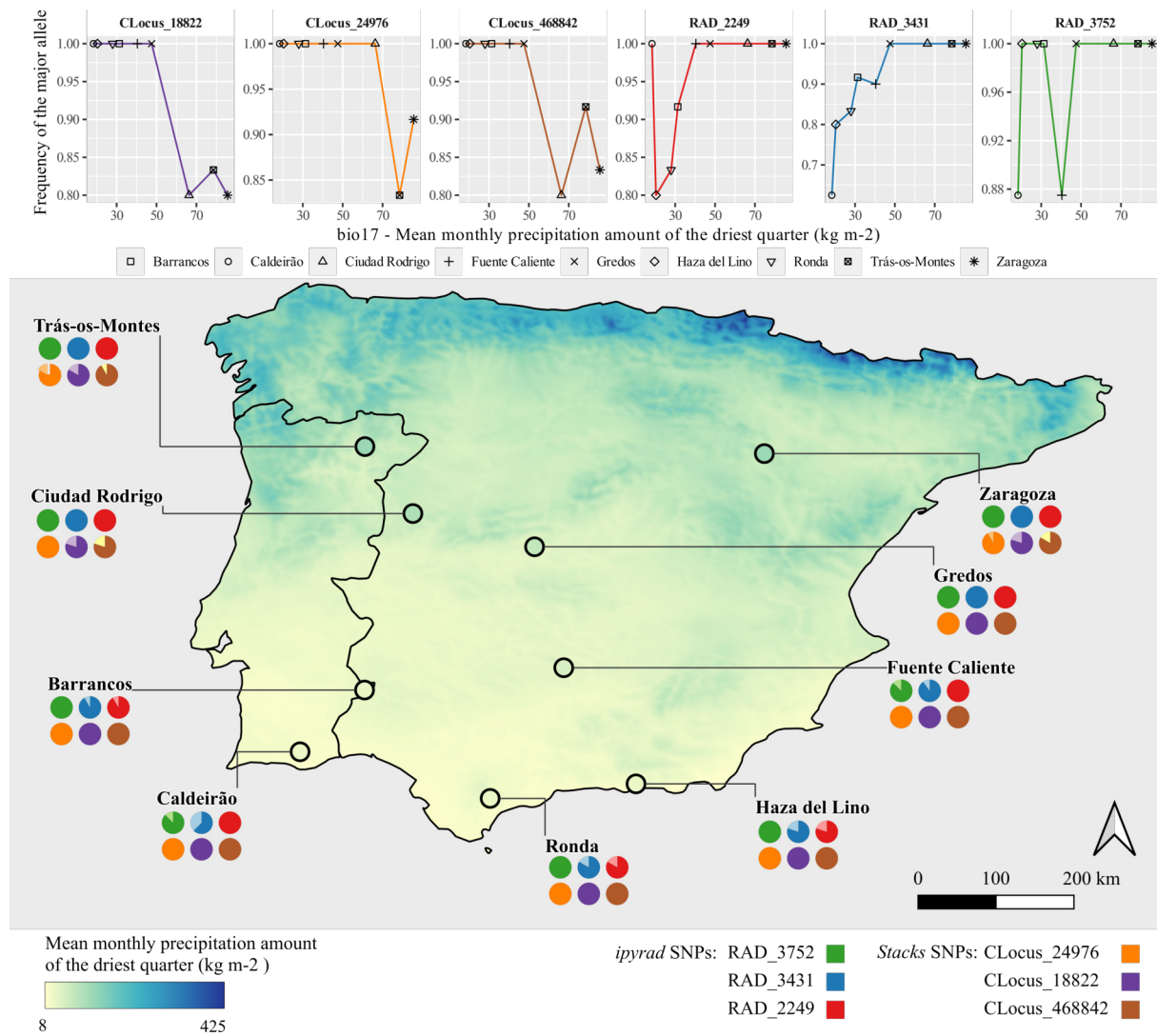


Figure S13: Representation of allele frequency turnover for loci associated with bioclimatic variable bio17 (mean monthly precipitation amount of the driest quarter). The three strongest associated loci are shown for datasets *ipr_centersnp_ib* and *stk_centersnp_ib*. Allele frequencies for each sampling location are plotted as a function of mean monthly precipitation values (above). The scale of the yy axis may differ across plots. Allele frequencies for each sampling location are represented along the geographical area of the Iberian Peninsula (below). The range of mean monthly precipitation is represented by a color scale. Values for each sampling location are represented above their geographical position.

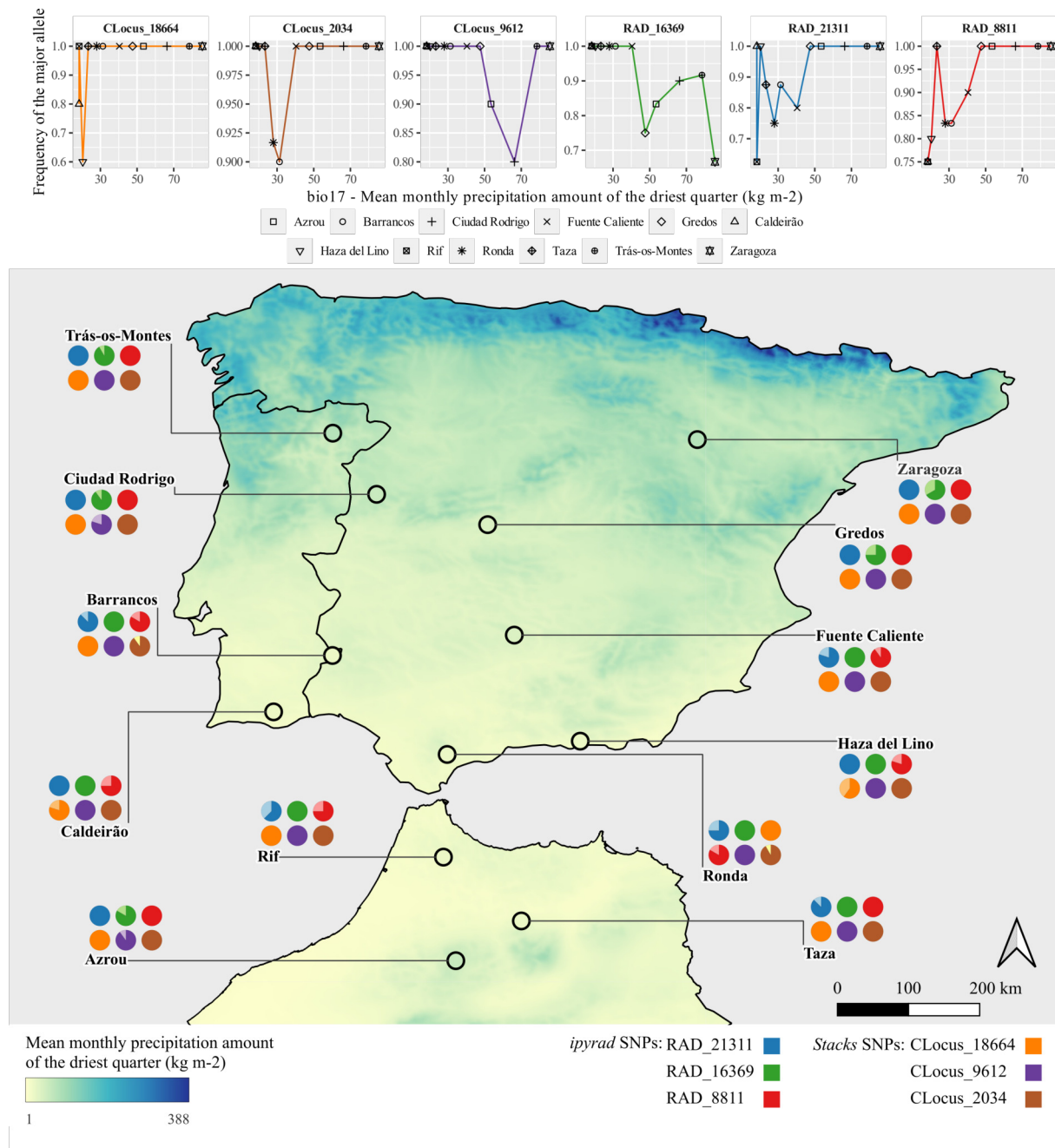


Figure S14: Representation of allele frequency turnover for SNPs/loci associated with bioclimatic variable bio17 (mean monthly precipitation amount of the driest quarter). The three strongest associated SNPs/loci are shown for datasets *ipr_centersnp_ib_mor* and *stk_centersnp_ib_mor*. Allele frequencies for each sampling location are plotted as a function of mean monthly precipitation values (above). The scale of the yy axis may differ across plots. Allele frequencies for each sampling location are represented along the geographical area of the Iberian Peninsula and northern Morocco (below). The range of mean monthly precipitation is represented by a color scale. Values for each sampling location are represented above their geographical position.

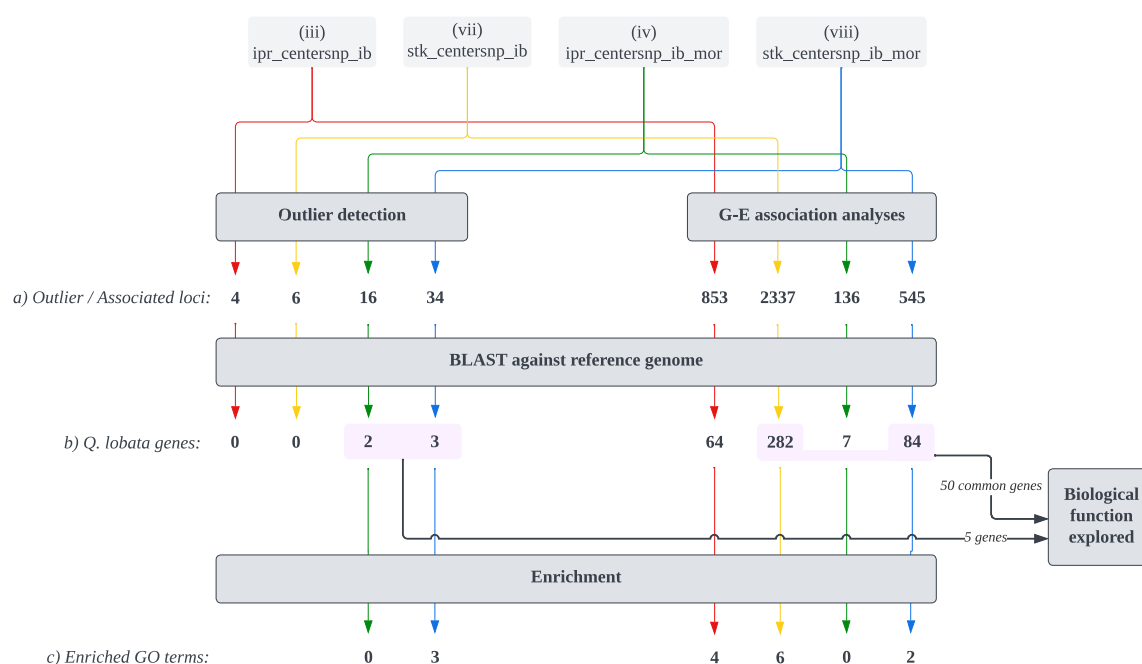


Figure S15: Simplified workflow of the outlier detection and genetic-environmental association analyses using *BayPass*, BLAST against the reference genome of *Q. lobata* and enrichment analysis using *g:Profiler*. This process is shown for datasets '*ipr_centersnp_ib*', '*stk_centersnp_ib*', '*ipr_centersnp_ib_mor*', and '*stk_centersnp_ib_mor*'. Represented are the number of outlier or associated SNPs/loci for each dataset after the respective analysis using *BayPass*, the number of unique *Q. lobata* genes found to be similar to the datasets' loci, and the number of enriched GO terms for the datasets that returned results. Also represented are the sets of genes that were explored in terms of biological function.

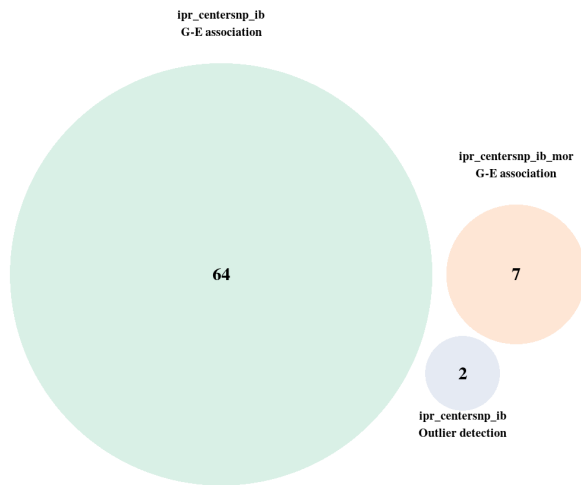
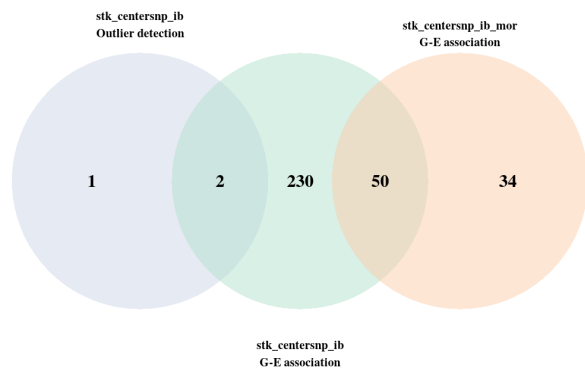
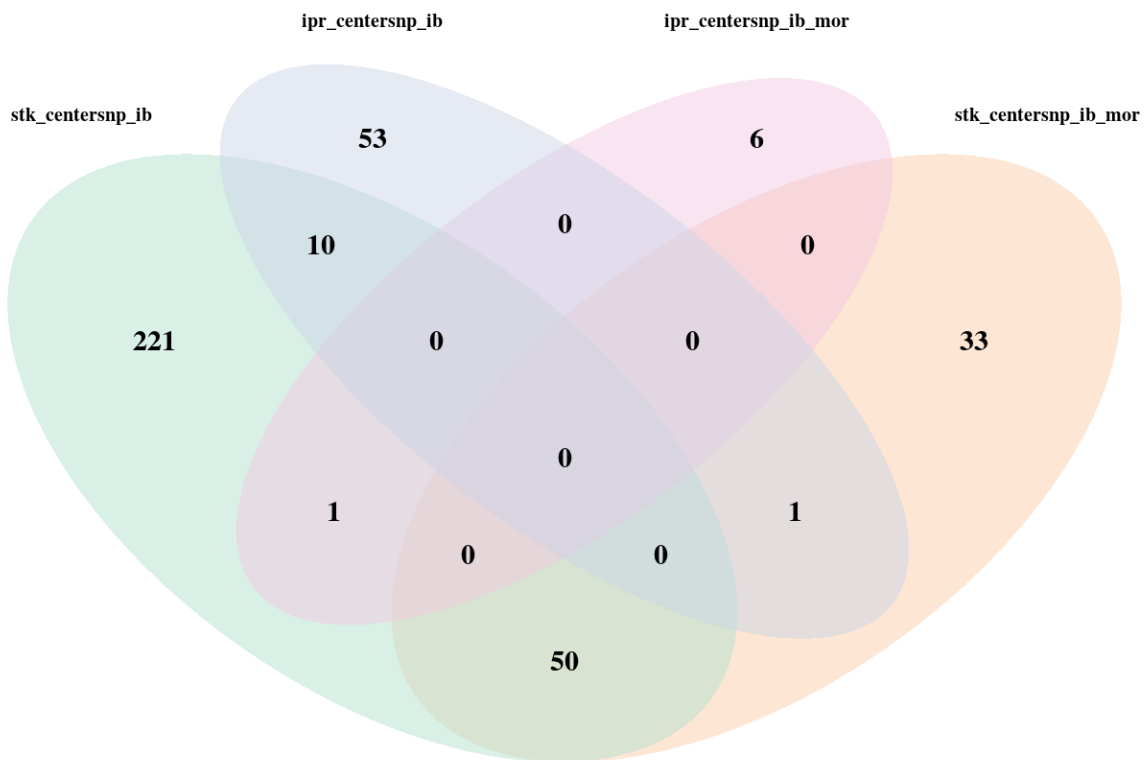
(a) *ipyrad* datasets(b) *Stacks* datasets(c) *ipyrad* and *Stacks* datasets, genetic-environmental association

Figure S16: Venn diagrams representing the overlap of genes found for the different datasets (ipr_centersnp_ib, ipr_centersnp_ib_mor, stk_centersnp_ib, and stk_centersnp_ib_mor) after annotation of outlier/ associated loci against the reference genome *Q. lobata*. Represented is the overlap for (a) *ipyrad* datasets after outlier detection and genetic-environmental association, (b) *Stacks* datasets after outlier detection and genetic-environmental association, and (c) the four datasets after genetic-environmental association analyses.

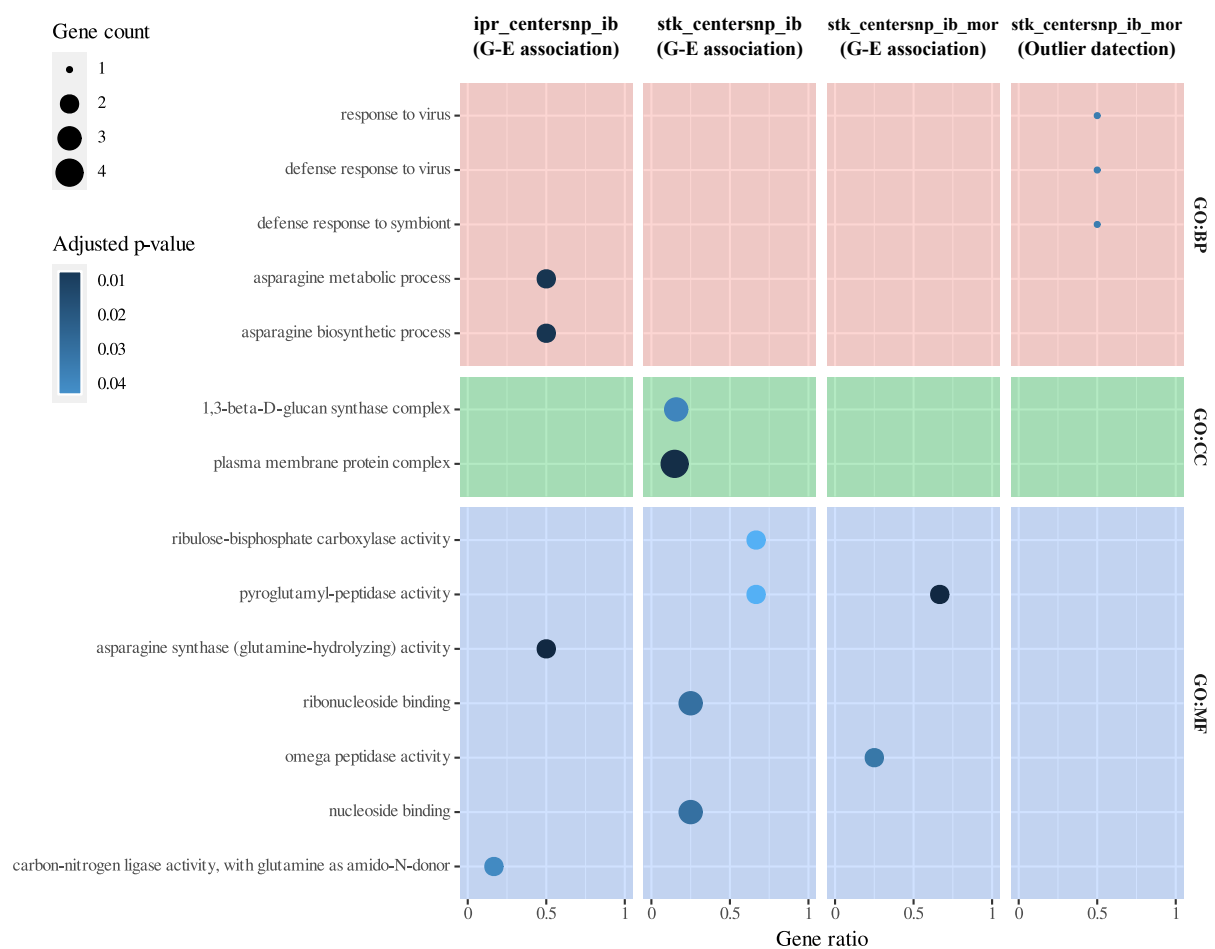


Figure S17: Enrichment analyses results for datasets ipr.centersnp_ib, ipr.centersnp_ib_mor, stk.centersnp_ib, and stk.centersnp_ib_mor). Analyses were run with *g:Profiler* for the list of *Q. lobata* genes similar to the loci obtained from the outlier detection and genetic-environmental analyses. Gene ratio represents the number of observed genes in our datasets divided by the number of genes in the reference genome annotated for each GO term). The gene count is the number of genes related to each GO term and is represented by the size of points. The adjusted p-values are a measure of the randomness of the intersection between the query and the ontology term and are represented by a color gradient. Results are grouped by the three GO term categories: *GO:BP* = Biological process; *GO:CC* = Celular component; *GO:MF* = Molecular function.

Table S7: Summary of *Q. lobata* genes with a potentially interesting biological function relating to heat and water stress response, among the 50 common genes between datasets (a) 'stk_centersnp_ib' and (b) 'stk_centersnp_ib_mor'. For each of the genes, the respective associated SNPs and bioclimatic variables are shown, as well as their putative role in the response to abiotic stressors such as temperature or drought.

Dataset	SNP	Associated bioclimatic variable	<i>Q. lobata</i> gene	Putative role
(a)	CLocus_2369	bio12	F-box domain-containing protein	Response to water stress, drought resistance [112]
(a)	CLocus_4420	bio17		
(b)	CLocus_14941	bio4		
(b)	CLocus_28758	bio17; bio9	Zeaxanthin epoxidase, chloroplastic	Water stress signal transduction [113]; acquired thermo-tolerance [114]
(a)	CLocus_5062	bio12		
(a)	CLocus_14421	bio17; bio9		
(b)	CLocus_31799	bio12	TTF-type domain-containing protein	Drought stress response [115, 116]
(a)	CLocus_5062	bio12		
(a)	CLocus_14421	bio17; bio9		
(b)	CLocus_31799	bio12	SBP-type domain-containing protein	Drought and heat stress response [117]
(a)	CLocus_17458	bio1		
(b)	CLocus_11532	bio9		
(a)	CLocus_20730	bio12	non-specific serine/threonine protein kinase	Drought stress tolerance [118]
(a)	CLocus_30911	bio17		
(b)	CLocus_19622	bio9		
(b)	CLocus_21139	bio2; bio9		
(b)	CLocus_12109	bio1; bio9		
(b)	CLocus_17301	bio9	C2 domain-containing protein	Drought stress response [119]
(a)	CLocus_18999	bio17		
(b)	CLocus_34652	bio2; bio12	Xanthine dehydrogenase 1-like	Drought acclimatization [120]
(a)	CLocus_28130	bio3; bio17		
(b)	CLocus_47571	bio2; bio3	BHLH domain-containing protein	Regulation of drought tolerance responses [121]
(a)	CLocus_39607	bio1		
(b)	CLocus_65627	bio4		