Extensive variation of leaf specialized metabolite production in sessile oak (*Quercus petraea*) populations is to a large extent genetically determined but not locally adaptive

Domitille Coq--Etchegaray[1*], Stéphane Bernillon[2, 3], Grégoire Le-Provost[1], Antoine Kremer[1], Alexis Ducousso[1], Céline Lalanne[1], Fabrice Bonne[4], Annick Moing[3], Christophe Plomion[1], Benjamin Brachi[1*]

[1] Univ. Bordeaux, INRAE, BIOGECO, UMR 1202, F-33610 Cestas, France

[2] INRAE, MycSA, F-33140 Villenave d'Ornon, France

[3] Univ. Bordeaux, INRAE, BFP, UMR 1332, Bordeaux Metabolome, MetaboHUB, F-33140 Villenave d'Ornon, France

[4] Univ. Lorraine, AgroParisTech, INRAE, UMR SILVA, F-54280 Champenoux, France

Corresponding authors:

Domitille Coq- -Etchegaray, INRAE, 69 route d'Arcachon, 33610 Cestas, France

Email: domitille.coq-etchegaray@inrae.fr

Benjamin Brachi, INRAE, 69 route d'Arcachon, Cestas, France

Email: benjamin.brachi@inrae.fr

Number of words: 6,485 (Introduction: 932, Material & Methods: 2,072, Results: 1,535, Discussion: 1,946)

Summary

- Specialized or secondary metabolites (SMs) play a key role in plant resistance against abiotic stresses and defences against bioaggressors. For example, in sessile oaks *Quercus petraea*, phenolics contribute to reduce herbivore damage and improve drought resistance. Here, we explored the natural variation of SMs in nine European provenances of sessile oaks and aimed to detect its underlying genetic bases.
- We sampled mature leaves from high and low branches on 225 sessile oak trees located in a common garden and used untargeted metabolomics to characterise the variation of 219 specialized metabolites. In addition, we used whole genome low-depth sequencing to genotype individuals for 1.4M genetic markers. We then performed genome-wide association analyses to identify the genetic bases underlying the variation of leaf SMs.
- We found that leaf SMs displayed extensive within-provenance variation, but very little differentiation between provenances. For ~10% of the metabolites we detected, most of this variation could be explained by a single genetic marker.
- Our results suggest that genetic variation for most leaf SMs is unlikely to be locally adaptive, and that selective pressures may act locally to maintain diversity at loci associated with leaf SM variation within oak populations.

Key words: genome-wide association study (GWAS), leaf specialized metabolites, European white oaks, biotic interactions, provenance trial, *Quercus petraea*, local adaptation, balanced selection

Introduction

One of the crucial mechanisms developed by plants to interact with their environment is the production of thousands of molecules, known as specialized or secondary metabolites (SMs). Each plant species produces thousands of different molecules, which range from extremely specific to shared across the plant kingdom (De Luca & St Pierre, 2000). In leaves, certain compounds, including tannins, flavonoids and glucosinolates were shown to have defensive effects against herbivores, including insects but also vertebrates (see Dearing *et al.*, 2005 for a review). Some compounds have become cues for specialist herbivores to find host plants (Wink, 2018) and others attract the predators of herbivores, hence reducing damage to the plant (McCormick *et al.*, 2012). Beyond interactions with herbivores, some SMs have antimicrobial properties and influence the leaf microbial community (Bailey *et al.*, 2005) like certain flavonoids in carnation (Galeotti *et al.*, 2008) and tomato (Vargas *et al.*, 2013). In addition, SMs production is also tightly linked to the plant immune system and phytohormones (Bednarek, 2012), making them key players in plant pathogen interactions and good candidates for quantitative resistance and even exapted resistance (Newcombe, 1998; Bartholomé *et al.*, 2020). Beyond the leaf, in root exudates, SMs including flavonoids help attract beneficial microbes and promote mycorrhizal symbiosis (reviewed in Walker *et al.*, 2003; Sebastiana *et al.*, 2021), hence enhancing growth and water-use efficiency.

The effects of SMs reach beyond the plant boundaries and influence the local biotic environment. Differences were detected in insect communities in the canopies of tree species producing different tannins (Forkner *et al.*, 2004), molecules which were also found to impact the litter microbial community (Schweitzer *et al.*, 2008). Finally, plant SMs play important roles for the plant protection against abiotic stresses (Sardans *et al.*, 2011). Many specialized compounds have antioxidant properties and may help scavenge reactive oxygen species (ROS) (Nakabayashi *et al.*, 2014). Thus, certain SMs, in particular anthocyanins, protect plants from the negative effects of ionising UV light (Rozema *et al.*, 2002), but can also enhance drought, heat and cold tolerance (Obata *et al.*, 2015).

Individuals within a plant species do not produce a homogeneous set of compounds and different chemotypes can emerge. Multiple studies have shown that the quantities and structure of SMs vary within species, either because they are plastic and respond to the

environment, or because of genetic differences between individuals. For example, in Arabidopsis and multiple members of the *Brassicaceae* family, glucosinolates and their metabolic pathways have been broadly investigated. Extensive genetic variation for specialized metabolites is present within species, with some compounds displaying broad geographical clines, suggesting adaptation to local herbivore communities (Zust *et al.*, 2012; Brachi *et al.*, 2015). In addition, length of the glucosinolate side-chain has been linked with interactions with insect herbivores (Burow *et al.*, 2010).

In pedunculate oaks (*Quercus robur L.*), the effect of specific compounds on defoliation *Tortrix viridana* was evidenced by a combination of metabolomics and transcriptomics and showed that within population variation in the production of defensive compounds largely influenced defoliation (Kersten *et al.*, 2013). Despite overall differences in the metabolic profile of trees among natural oak stands, all stands included resistant or susceptible oak trees (Bertić *et al.*, 2021). Leaves from resistant oak trees were enriched in defence-related polyphenolic compounds, while leaves from the susceptible oaks were enriched in growth-related substances such as carbohydrates and amino-acid derivatives.
These results are therefore consistent with both the adaptation of populations to local environmental conditions, but also the maintenance, within stands, of variation in the investment of trees in defence and growth.
European white oaks cover a large portion of European temperate forests and the decline of many populations, likely accelerated by climate change, could have dramatic consequences for the ecosystem services they provide including biodiversity. Indeed, a recent study showed that oaks support a rich community of organisms with 2,300 species associated with oak forests in the UK, of which 326 are directly associated with oaks (Mitchell *et al.*, 2019). Many of these interactions, and the central role of oak trees in forest ecosystems, could be partly mediated by SMs. With oak populations suffering from decline related to both abiotic and biotic stresses imposed by global change, studying the natural variation and genetic bases of leaf SMs is essential to better understand the evolutionary history of SMs variation in oak populations. In particular, quantifying how much of this variation is locally adaptive and how much is maintained within population, is an important addition to the study of traits adaptive to climate, such as phenology or drought tolerance (Sáenz-Romero *et al.*, 2017; Torres-Ruiz *et al.*, 2019), to estimate the adaptive potential of oak populations.

In this study, we investigated the natural variation and the genetic bases of non-volatile leaf SMs in nine European sessile oak (*Quercus petraea* (Matt.) Liebl.) provenances growing in a common garden, using untargeted metabolomics. We started by characterising the genetic structure of oak populations and examined the differentiation between provenances for leaf SMs. We then performed a genome-wide association study to identify the genetic bases and architectures of leaf SMs. We investigated candidate genes and molecular structures for metabolites presenting strong associations and interesting patterns of variation. Our analyses revealed very little differentiation among provenances for 83% of the metabolites quantified, with the vast majority of the phenotypic variation being present within provenances. Our genome-wide association study revealed that phenotypic variation for 63% of the metabolites was largely genetically determined, and displayed mono- to oligogenic architectures. Together our results indicated that the variation of the majority of the metabolites we quantified did not contribute to local adaptation among provenances, at a European geographical scale.

Materials and Methods

Sampling sessile oak leaves

We sampled sessile oak trees grown in a common garden experiment located in Eastern France (Sillegny, 48°59′ 24″ N 06°07′ 56″ E). This common garden is one of the four common gardens of a large-scale (106 oak populations) and long-term provenance trial on sessile oak *Quercus petraea* Matt. Liebl. (Ducousso *et al.*, 2022)

Within the 106 populations, we selected nine populations that were previously gathered along a latitudinal gradient spanning from the South-West of France to the North of Germany (Fig. **1a**, Table **S1**) and leaves in the common garden on September 7th and 8th 2016 on 225 trees (22 to 28 trees per population). Trees were between 29-35 years old. For each tree, we sampled four to six fully developed leaves from branches at two heights: low branches, mostly protected from direct sunlight by the canopy, and high branches exposed to sunlight. Leaves were collected either using a pole pruner when possible, or with a shotgun whenever the branches were too high to reach.

Four to six leaves per branch were stored in 20 mL plastic vials and frozen on dry-ice upon harvest (see Supplementary Material and Methods S1).

Sequencing and genotyping of sessile oak individuals

Briefly, we extracted DNA from freeze dried leaf material and sequenced individuals to a depth of ~10X using the Illumina NovaSeq system. Details are provided in the Supplementary Material and Methods S2.

We used a home-made bioinformatic pipeline, developed with Snakemake v5.8.1 workflow manager (Köster *et al.*, 2021) and Singularity containers (Kurtzer *et al.*, 2017). We removed Illumina TrueSeq adapters and trimmed reads using cutadapt v1.18 (Martin, 2011) and sickle v1.33 (Joshi & Fass, 2011). We aligned paired-end reads with bwa-mem2 v2.2.1 (Li, 2013) on *Quercus robur* reference genome (Plomion *et al.*, 2018). We marked duplicated reads using GATK v4.2.4.0 and clipped overlapping read pairs using BamUtils v1.0.15 (Jun *et al.*, 2015), recommended in GATK best practices (Van der Auwera *et al.*, 2013). We created mpileup files of the 224 individuals using samtools v1.9 (Danecek *et al.*, 2021) and converted them into pro files using sam2pro v0.8. We detected bi-allelic SNPs using the Bayesian genotype caller (BGC) described in (Maruki & Lynch, 2017). We created a file summarizing nucleotide read quartets for the 224 individuals from the pro files using a genotype-frequency estimator (GFE) (Maruki & Lynch, 2015) and called genotypes with the BGC software. We discarded SNPs with a minor allele frequency (MAF) lower than 10%, with more than 5% of missing calls and located in regions annotated as transposable elements in the Q. *robur* reference genome. These filters were achieved with a combination of awk command lines and PLINK v1.9 (Chang *et al.*, 2015). We discarded individuals with genome-wide genetic similarity above 95% (i.e., 95% of SNPs with identical genotypes between the two individuals) as they may have corresponded to the same individual sequenced twice.

Population genomics analysis of sessile oak populations

We estimated linkage disequilibrium (LD) decay along the oak genome using the r^2 function available in PLINK v1.9 with window size of 50kb.

To study the genetic structure of *Q. petraea* populations we pruned the SNPs according to LD previously estimated and removed highly correlated markers ($r^2 > 0.2$) using the "--indep-

pairwise" function of PLINK v1.9 (Chang *et al.*, 2015) with a window size of 2 kb and a step-size of 1. We computed an average Hudson F_{ST} (Bhatia *et al.*, 2013) index between pairs of populations, estimated across all SNPs in the pruned dataset using the "--fst" PLINK v2.0 (Chang *et al.*, 2015) with default parameters. We computed F_{ST} (Weir & Cockerham, 1984) across populations using the "--fst" function in PLINK v2.0 for each SNP. To investigate global patterns of population structure and investigate whether our SNP set allowed source identification, we performed a principal component analysis using the "--pca" function in PLINK v1.9 (Chang *et al.*, 2015).

Metabolomics: extraction, data acquisition, processing and filtering

Extraction and LC-MS analyses: Briefly, specialized metabolites were extracted using 70% methanol. For all methanolic extracts, the separation of extracts was achieved using reverse phase liquid chromatography with a C18 column over a 10mn gradient (see Supplementary Material and Methods S3).

We performed mass-spectrometry analyses with different instruments and settings to generate two datasets. The first dataset will be dedicated to genome-wide association analyses and the second one for annotations of metabolites.

The first dataset, hereafter the "GWAS set", included all samples. For this analysis the chromatography flow was directed to an electrospray ionisation probe set to positive mode (-500 V endplate offset, +3.5 kV capillary voltage, 2.4 bar nebulizer, dry gas flow of 8.0 L/min at 190˚ C) into a hybrid quadrupole time-of-flight (QTOF) mass spectrometer (Bruker, Bremen, Germany) with no collision energy. The mass-to-charge ratio ion scan was from m/z 50 to m/z 1500 with an acquisition frequency of 2 Hz. The mass spectrometer was m/z-calibrated with a 10 mM lithium formiate solution injected at the begin and the end of each chromatogram.

The second dataset, hereafter the "annotation dataset", was generated from 56 randomly selected samples in the nine populations. We used the data dependent mode of the LTQ-Orbitrap Elite (ThermoScientific, Bremen, Germany), to generate MS^2 spectra for all sufficiently intense fragments. We ran this analysis on the plates of methanolic extracts twice, with the HESI electrospray ionisation probe operating in the positive and negative modes. Parameter details are provided in Supplementary Material and Methods S4.

Data treatment and filtering: For the "GWAS dataset", we converted raw proprietary files from the instrument to the mzML format using ProteoWizzard program (Chambers *et al.*, 2012). We removed files with no calibration peaks and filtered remaining files to keep only data between 80 and 570 s using msconvert (Chambers *et al.*, 2012) to remove calibration peaks and column wash offs at the beginning and the end of the runs. Past these initial processing steps, all analyses were carried out in a home-made bioinformatic pipeline using Snakemake workflow manager, Singularity containers and R v4.1.1 (R Core team 2021) see (gitlab link). We used the IPO R package v1.18 (Libiseller *et al.*, 2015) to optimise parameters for retention time (RT) correction and peak picking implemented in the R package XCMS v3.14 (Smith *et al.*, 2006) (Table **S2**). We ran the "findChromPeaks" function to detect chromatographic peaks. Then, we ran the "adjustRtime" function to correct RT deviation and calculated adjusted retention times using a subset-based alignment based on the QC samples regularly injected. Finally, we ran the "PeakDensityParam" and "groupChromPeaks" function to match all the detected peaks between samples. We produced a table of peak intensities and identified pseudo-molecules, based on RT and intensity correlations among peak groups using the "groupFWHM" and "groupCorr" functions of the CAMERA R package v1.48.0 (Carsten Kuhl *et al.*, 2021).

After that step we filtered the dataset to remove sample outliers and peak outliers. Samples from the eighth plate were removed because the mass spectrometer failed to m/z-calibrate. First, we used the median peak intensity across all peak groups to remove samples that had values consistent with blanks and removed blanks that had median peak heights similar to regular samples. These rare cases probably arose from errors during the randomisation of samples into the 96 well plates. Second, we filtered the peak groups defined by XCMS using QCs and blanks, with the rationale that peaks had to be significantly higher in the QCs than in the blanks to be retained. We used a one-way variance analysis (ANOVA, p-*value*<0.05) on log-transformed intensity data for each peak to determine the average difference in intensity between blanks and QCs and assess significance. Because in this situation relaxing the threshold for significance resulted in a more conservative filtering, we did not apply correction for multiple testing (false discovery rate, FDR). In addition, we also removed peak groups with coefficients of variation above 20% in the QCs. The final step was to verify the quality of injection by checking that the internal standard included in the extraction buffer (quercetin) was still clearly present in the data set. QCs were not used to correct for intensity

drift and peak intensity was not corrected for the mass of material used in the extraction. (See Dataset S1)

Metabolite annotations: We used the "annotation dataset" to annotate pseudo-molecules using an integrated metabolomic workflow (Fraisier-Vannier *et al.*, 2020) (see Supplementary Material and Methods S5). We obtained annotations for the "annotation set" and created a correspondence table between the "annotation dataset" measured on the LTQ-ORBITRAP- and the "GWAS-dataset" measured using the QTOF-MS with a *m/z* tolerance of 0.03 and an RT tolerance of 0.6 min (Dataset S2). Then, we selected putative annotations for 21 metabolites associated with genetic variation in oaks described below and manually validated annotations using MS2 spectra.

Statistical analysis of leaf specialized metabolites variation

Specialized metabolite variation analysis: We studied global patterns of leaf metabolic profiles using a principal component analysis with unit variance scaling on raw signal intensity values of representative peaks.

To investigate the differentiation of phenotypic variation between populations at each branch height, we fitted models with a single random effect of the population on log transformed peak intensity (log(x+1)), and computed the proportion of variance explained by the population of trees at each branch height. We found that models using a student-t distribution yielded better fit to the data, thanks to greater tolerance for outliers. Specifically, the model is described below, along with the priors for each parameter.

$$y_{ip} \sim StudentT(\nu, \mu_{ip}, \sigma^2)$$

$$\mu_{ig} = \alpha + \lambda_p$$

$$\alpha \sim \mathcal{N}(m, 10^2)$$

$$\lambda_p \sim \mathcal{N}(0, \tau^2)$$

$$\tau \sim Exponential(0.1)$$

$$\sigma \sim Exponential(0.1)$$

$$\nu \sim Exponential(1)$$

Where Y_{ip} was the log(x+1) transformed phenotype (one pseudo-molecule, at one branch height) of the ith tree from the pth population, α is the intercept of the model, m is the mean of the log transformed phenotype, and λ_p is the population effect added to the model intercept for the population of the ith tree from the pth population. The proportion of variance explained (PVE) by the provenance effect was calculated as:

$$PVE = \frac{\tau^2}{\tau^2 + \sigma^2}$$

Models were fitted using the R package brms v2.18.0 and R v4.2.2 (Bürkner, 2018), running 5 chains of 4000 iterations, including 2000 of warmup. We reported the mean of the posterior distribution for PVE and the associated 85% credible interval for each molecule.

Specialized metabolite signal intensities correlation: We studied correlations among log-transformed signal intensities of SMs. We used all samples, excluding technical replicates and QCs, in both heights, calculated Spearman correlation coefficients between pairs of pseudo-molecules using the "corr.test" function of the psych R Package v2.2.9 (Revelle, 2022). We considered correlations above 0.5 and FDR corrected p-value <0.05 as significant (Perez De Souza *et al.*, 2020). We represented correlation networks using Cytoscape v3.9.1 (Shannon *et al.*, 2003).

Genome-wide association analysis

Genome-wide association studies: The genome-wide association analyses were performed on the log transformed signal intensities of individual pseudo-molecules using a univariate linear mixed model implemented in GEMMA v0.98.3, which accounts for the genetic relatedness among individuals, estimated as a centred genome-wide kinship (K) using the function provided in GEMMA.

We analysed both heights separately and controlled for multiple testing by applying a Benjamini-Hochberg correction to the *p*-values from the association tests. For this correction we considered the total number of tests performed (number of SNPs * number of phenotypes, equal to 607M) and a significance threshold of *p*-value<0.01. This resulted in considering associations as significant when $-\log_{10}$(*p*-value) > 6.79.

mQTLs regions: To facilitate the interpretation of significant association, we clustered SNPs by regions, hereafter mQTLs. We considered all significant associations ($-\log_{10}(p\text{-value}) > 6.79$), across all metabolites, and, walking along the genome, we grouped consecutive significant SNPs in the same cluster if they were located less than 1 Mb apart. If the distance between consecutive significant SNPs was larger than 1 Mb a new cluster was created. For each mQTL, we counted how many SNPs, and metabolites segregated within the mQTL.

Gene annotation of highly associated SNPs: We searched genes within regions near SNPs highly associated with metabolite variation. We used the estimated linkage disequilibrium decay distance (~3,000 bases) to create regions of 6,000 bases around SNPs with associations. We parsed the gene annotation file using the start and end position of regions to match with genes start and end position annotations and extract sequences using samtools faidx v1.6. Then we used BLAST+ v2.12.0 (blastx) (Camacho *et al.*, 2009) with default parameters to find similarity matches between translated nucleotide regions of 6 kb or gene sequences against the *Arabidopsis* proteome (Berardini *et al.*, 2015).

Results

Population genomics of sessile oak populations

In this study we sampled 225 sessile oak trees originating from nine populations (Fig. **1a**) and sequenced 224 of these trees to a depth of 10X (on average 36.5M paired-end reads per tree). After filtering, we were able to genotype trees for 1,408,029 SNPs with a minor allele frequency above 10%, a missing genotype rate of 10% and after excluding regions annotated as transposable elements on the *Q. robur* reference genome, the closest genome available. We obtained an average SNP density of 170 SNPs per window of 100 kb, ranging from 0 to 923 SNPs, and 90% of 100 kb windows included over 9 SNPs (Fig. **S1**). Three pairs of individuals displayed genetic similarity above 95% and were removed from the dataset, bringing the total number of individuals genotyped to 218.

Using these 1.4 million markers we estimated that linkage disequilibrium (LD, r^2) decreased to values below 0.2 over 2.9 kb in our collection of 9 populations spanning a large fraction of the species distribution range (Fig. **S2**).

We then pruned the dataset down to 356,413 un-correlated SNPs ($r^2<0.2$) in the 218 individuals to investigate population differentiation and genetic structure. First, we investigated the genetic differentiation between pairs of provenances using the Hudson fixation index F_{ST}. We found that F_{ST} between pairs of populations was on average 0.008, ranging from 0.002 to 0.014 (Fig. **1b**). We observed the highest differentiation (F_{ST}=0.0146) between two populations from France: Vachères and Saint-Sauvant (Fig. **1b**). Second, we investigated patterns of population structure using a PCA. The first principal component (PC) explained 7.2% of total variance and separated the two populations from the South-East of France (Vachères and Grésigne) from the other provenances (Fig. **1c**). The second PC explained 6.5% of total variance and mostly captured the difference between a group of individuals from Bezange from the other individuals included in the study (Fig. **1c**). Inspection of kinship values shows that these individuals were likely half-sibs. Overall, the PCA showed clustering of individuals according to their population of origin (Fig. **1c,d**).

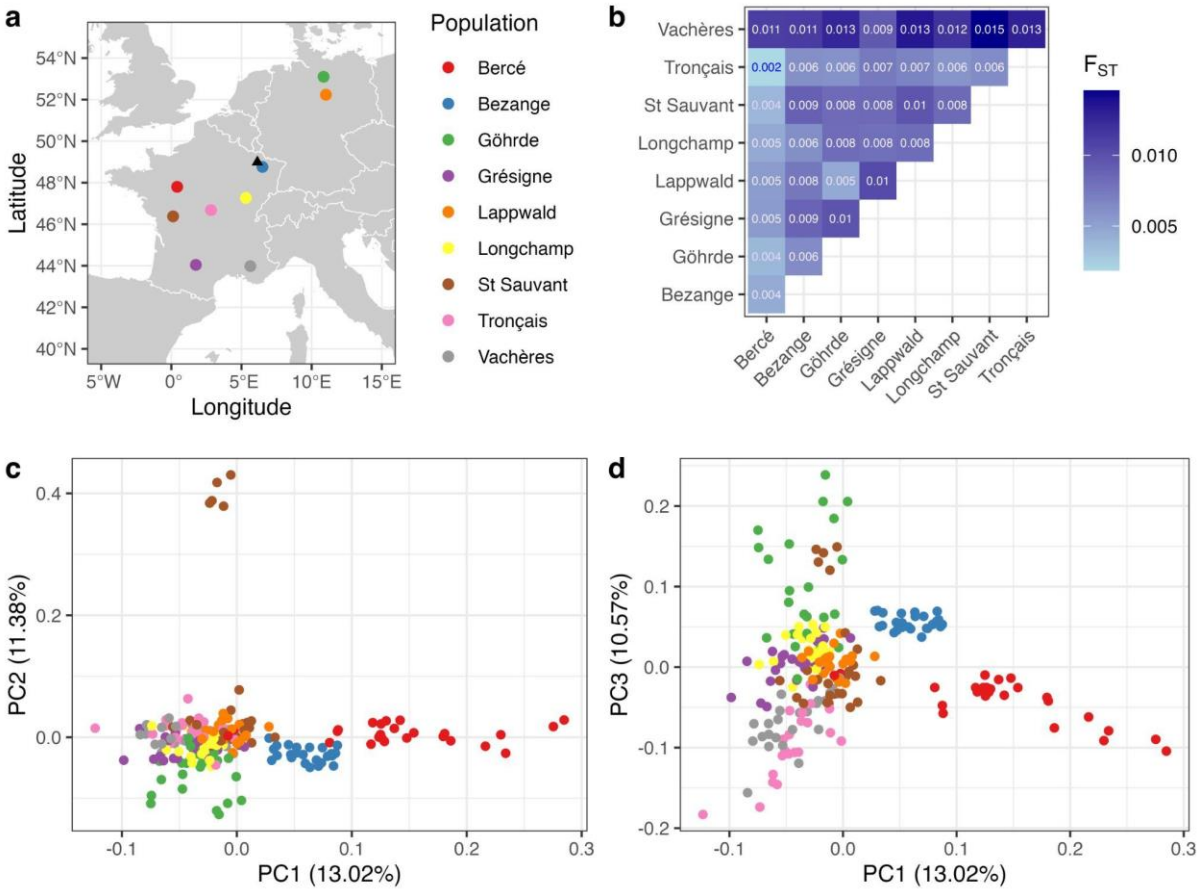

Fig.1 Structure and differentiation of sessile oak populations. **(a)** Map of sessile oak populations included in the study. The coloured dots on the map correspond to the geographical origin of the 9 populations we sampled. The location of the common garden in which the trees are installed is

marked by the black triangle. **(b)** Hudson pairwise F_{ST} between populations. Populations are indicated along the x and y axes. F_{ST} between pairs of populations are indicated in the half-matrix. The cells in the matrix are coloured in darker shades of blue for low F_{ST} values. **(c-d)** Genetic variation of oak populations visualised using a PCA. Each point corresponds to trees growing in a common garden (N=218), projected in the plane defined by the first three components **(c)** PC1, PC2 and **(d)** PC2, PC3 of a principal component analysis (scores plot) computed using 356,413 un-correlated ($r^2 < 0.2$) SNPs markers. Points are coloured according to populations as in **(a)**.

Leaf specialized metabolites (LSMs) characterisation

We measured leaf SMs within oak leaves for 225 trees at two heights using a high-throughput untargeted LC-MS approach. Our raw dataset contained 750,540 signal intensities. After filtering, we obtained intensity signals for 219 pseudo-molecules in leaves from low and high branches of 209 and 215 trees, respectively. Given our sampling and LC-MS protocols, the 219 pseudo-molecules analysed in this study corresponded to non-volatile, moderately polar specialized metabolites.

Based on MS2 data, we performed automatic annotation of the pseudo-molecules present in our leaf samples. We obtained putative annotations for 89 pseudo-molecules (out of 219) of the "GWAS set" (all samples, see material and methods) in negative and positive mode. These putative annotations were grouped in ontology classes, and included mostly flavonoids, terpenes, quinic acids and derivatives, hydrolyzable tannins, in addition to other rarer classes (Dataset S2).

Variation of leaf specialized metabolites

We visualised leaf SMs variation of individuals using a PCA on the signal intensity values of 219 pseudo-molecules for 424 leaf samples. PC1 explained 18.3% of total variance and clearly separated leaves sampled from high branches from leaves sampled from low branches (Fig. **2a**). While there was a large effect of branch height on pseudo-molecule intensities overall, variation of individual molecules was generally positively correlated between the two branch heights (Fig. **2c**). Across all 219 pseudo-molecules, Spearman rank correlations coefficients computed between phenotypic variation in high and low branches ranged from -0.03 to 0.91, with a median of 0.47. After FDR correction, we found significant positive correlations for 199 out of the 216 pseudo-molecules.

In contrast to patterns observed for the genetic variation, we found no differentiation of populations along the first two PCs in this analysis (Fig. **2b**). This was generally true for individual metabolites. We estimated the proportion of variance explained (PVE) by a random population effect for each molecule at each branch height. We found that only 31 pseudo-molecules displayed PVE by population effects with a low credible interval above 1% at both branch heights, and only 37 at, at least, one branch height. Point estimates (posterior means) for the PVE by population effects were below 10% at both branch heights for 165 pseudo-molecules out of 219 (~75 %) (Fig. **2d**). Metabolites displaying the largest provenance effects are represented in (Fig. S3).

We studied pairwise correlations among pseudo-molecule signal intensities for leaves sampled on low and high branches separately. We used pairwise correlations above 0.5 and qFDR < 0.05 to build correlation networks. This analysis revealed that nearly all significant pairwise correlations were positive with only one and two negative correlations for leaves collected on high and low branches, respectively. In addition, pseudo-molecules with similar retention time were often more strongly correlated and clustered together in the networks (Fig. **S4**).

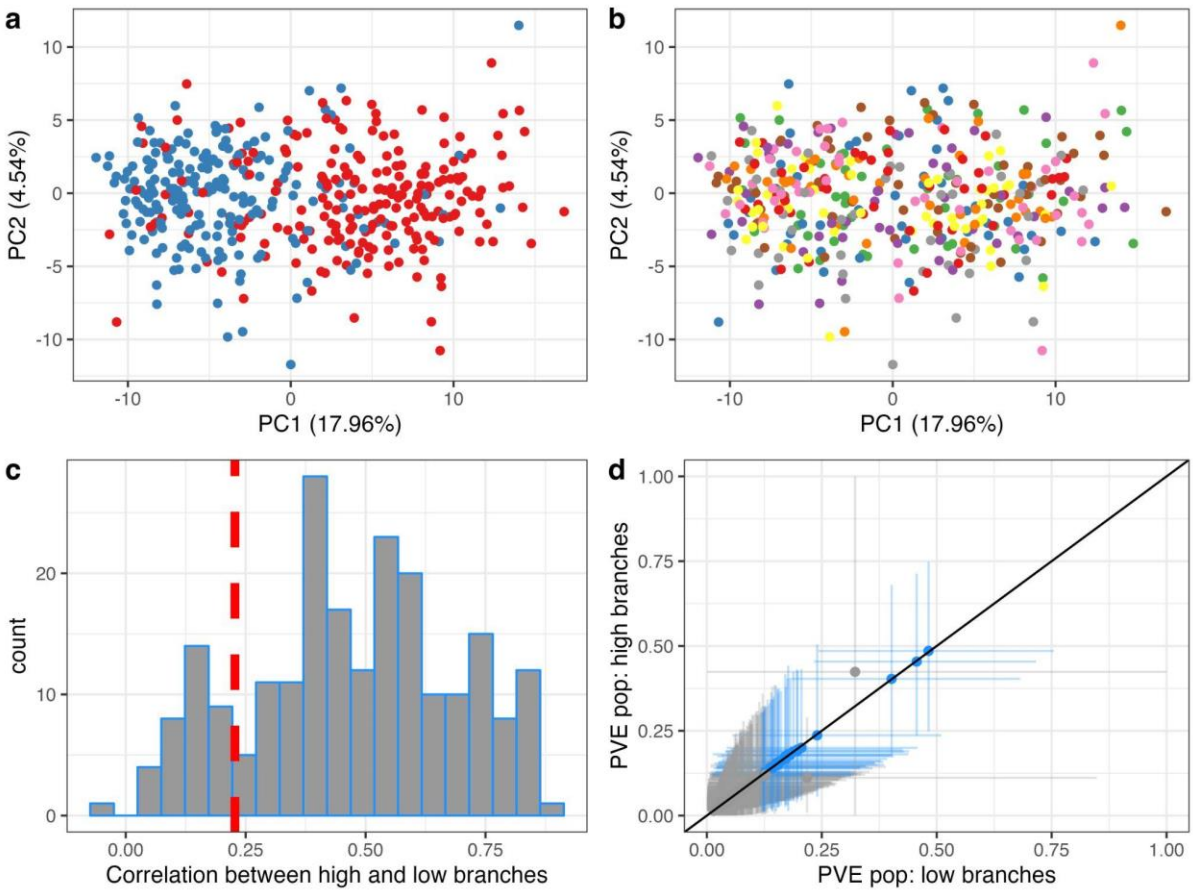

Fig. 2: Variation of leaf SMs in 225 oak trees from nine European populations. **a.** Phenotypic variation of leaf SMs visualised using a PCA. Each point corresponds to a sample in our study (424 samples, 219 variables) projected in the PC1 x PC2 plane (scores plot) of a principal component analysis of the scaled peak intensities of 219 pseudo-molecules. The points are coloured according to the height (high and low) at which leaves were collected in the trees. **b.** Same as **a.** but points are coloured by populations. **c.** Distribution of phenotypic correlations between high and low branches. The x-axis represents Spearman correlation coefficients between peak intensities in high and low branches, for each pseudo-molecule (N=219). The y-axis represents counts of pseudo-molecules.The red dashed vertical line represents the lowest significant correlation coefficient observed between high and low branches (p-value ≤ 0.05 after FDR correction). **d.** Proportion of variance explained by a random population intercept for each pseudo-molecule in leaves from low (x-axis) and high (y-axes) branches. The vertical and horizontal segments are 85% credible intervals. Points coloured in blue correspond to pseudo-molecules with the lower bound of the credible interval above 1% (N=31) at both branch heights. The black line is the line of intercept 0 and slope 1. The variation of the 5 pseudo-molecules with the largest provenance effects are represented in Figure S3.

Genetic basis of leaf specialized metabolites

To investigate the genetic basis underlying the variation of leaf LSMs in sessile oaks, we used a univariate linear mixed model implemented in GEMMA accounting for population structure. We performed association analysis separately for each of the 219 pseudo-molecules and for the two branch heights (219 molecules * 2 branch heights = 438 scans). All trees were genotyped for 1,386,405 SNPs (MAF>10% and SNPs missingness < 5%) and we performed association analyses in panels of 204 and 208 trees, for low and high branches, respectively (Fig. **3**). Considering the results from both branch heights, we obtained a total of 5,305 significant associations (after the Benjamini-Hochberg correction filter, *p*-value <0.01), corresponding to 2,272 individual SNPs associated with at least one of 138 pseudo-molecules (63% of pseudo-molecules) measured for at least one branch height. These significant associations clustered in 155 loci, or mQTLs hereafter, separated by at least 1 Mb (see Material and Methods). Note that we did not account for associated SNPs mapped to scaffolds not assigned to chromosomes.

For downstream analyses, we applied a more stringent filter and focused on 35 mQTLs including highly significant associations ($-\log_{10}$(*p*-value) > 10) for at least one pseudo-molecule. These 35 loci were distributed on all chromosomes of the oak genome except on chromosome 12, and had sizes ranging from 1 bp to 1.47 Mb (Fig. **3**). For each of the 35 mQTLs, we investigated candidate genes and selected the pseudo-molecule with the strongest association for structural annotation (N=21 pseudo-molecules Table 1, Dataset S3).

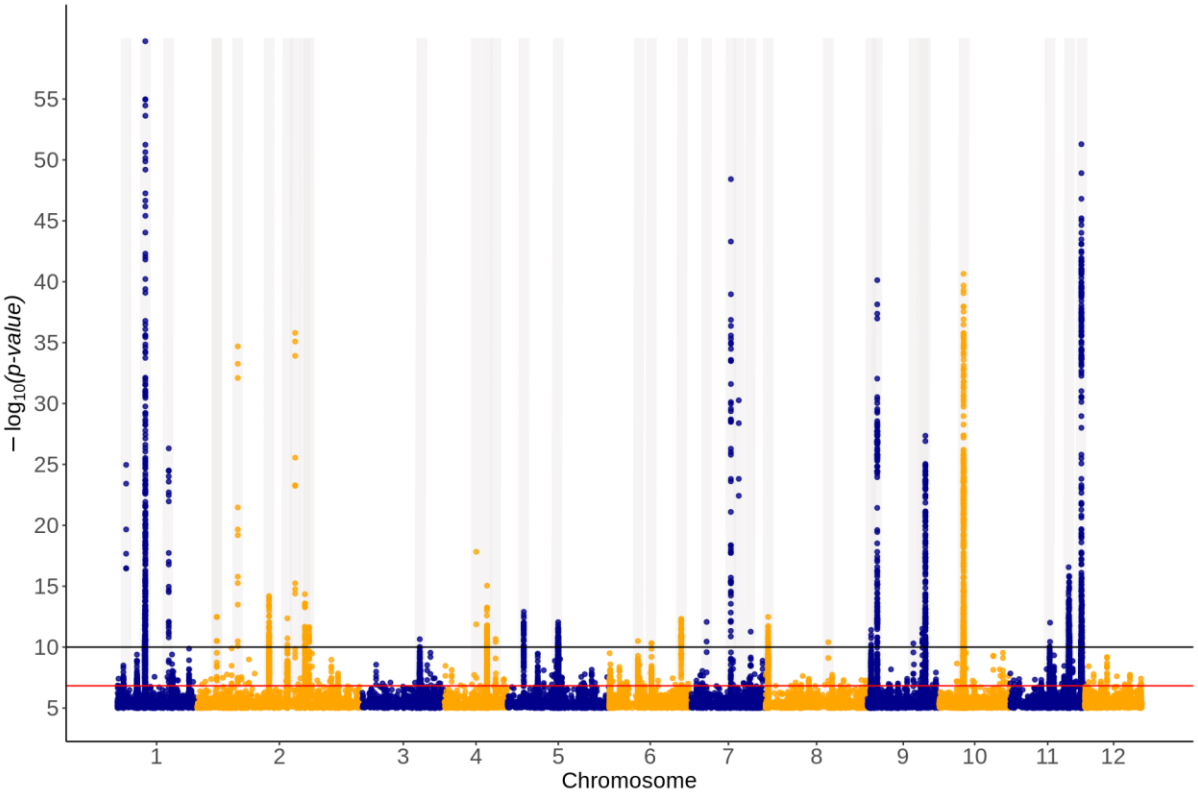

Fig. 3 Genome-wide association mapping of oak leaf specialized metabolites. Manhattan plot where each dot represents a genetic marker (N=1,386,405). The positions of genetic markers are given, along the 12 pseudo-chromosomes of the pedunculate oak reference genome, along the x-axis. The y-axis indicates the best association score (-$\log_{10}$(*p*-value)) identified for each marker over the 219 pseudo-molecules and the two branch heights investigated in our study. The red horizontal line represents the association score above which associations were considered significant (*p*-value < 0.01 after Benjamini-Hochberg correction). The black horizontal line represents the threshold above which associations were considered highly significant (-$\log_{10}$(*p*-value)>10). The vertical grey lines represent the position of the 35 mQTLs. For clarity, only markers with association scores above 5 are represented.

Among these 21 focal pseudo-molecules, 15 displayed associations (-$\log_{10}$(*p*-value) > 10) within one mQTL only. The rest displayed associated SNPs in two to six mQTLs. As expected from the correlation networks (Fig. **S4**), molecules with retention times within about a minute often displayed associations within the same mQTLs (Fig S4, Dataset S3). For 19 of the focal 21 pseudo-molecules, associations were detected within the same mQTL for both branch heights (Dataset S3).

The SNPs identified as associated with these 21 pseudo-molecules typically explained a large fraction of phenotypic variance (Fig. **4a**, Fig. **S5**). We investigated the genotype distribution within the nine populations of those SNPs. For the 21 pseudo-molecules, we observed that the heterozygous and homozygous genotypes were distributed within the nine provenances without differentiation among populations (Fig. **4a**, Fig. **S5**). Comparing the differentiation among populations for the genetic markers associated with the 21 pseudo-molecules to genome-wide patterns of differentiation revealed that only four out of these 21 markers displayed values above the 80th percentile ($F_{ST} > 0.02$) of the genome-wide F_{ST} distribution (Fig. **S6**). For five of the 21 pseudo-molecules, the heterozygous genotype was associated with higher signal peak intensities than the homozygous genotype (Fig. **4b**, Fig. **S5**). In addition, the homozygous genotype was not present in our dataset for 16 SNPs associated with seven pseudo-molecules (Fig. S5).

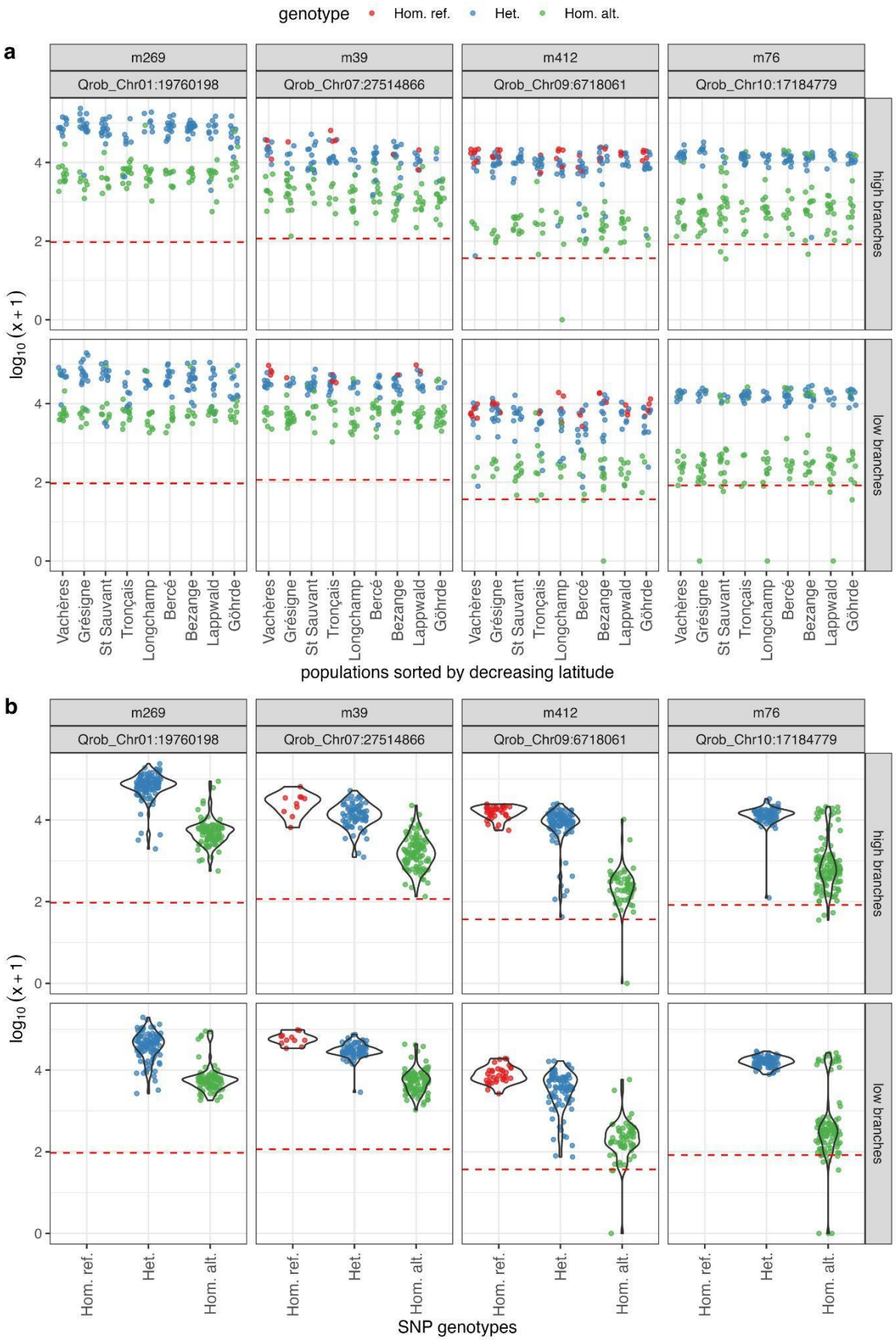
genotype
Hom. ref.
Het.
Hom. alt.
a
m269
Qrob_Chr01:19760198
m39
Qrob_Chr07:27514866
m412
Qrob_Chr09:6718061
m76
Qrob_Chr10:17184779
high branches
low branches
log10 (x + 1)
Vachères
Grésigne
St Sauvant
Tronçais
Longchamp
Bercé
Bezange
Lappwald
Göhrde
populations sorted by decreasing latitude
b
Hom. ref.
Het.
Hom. alt.
SNP genotypes

Fig. 4 Variation of many leaf SMs can be explained by genetic variation at a single genetic marker. In all panels, the plots represent the variation of the peak intensity (y-axis, $\log_{10}(x+1)$) for pseudo-molecules m269, m39, m76, m412 for leaves collected on high and low branches. In the eight panels in **(a)**, points correspond to leaves of individual samples in high branches (top row) and low branches (bottom row), grouped by population along the x-axis. Note that a jitter was applied to the location of points along the x-axis for clarity and that populations are ordered by increasing latitude of origin. Points are coloured according to the genotype at the SNP most strongly associated with each pseudo-molecule as described in the legend at the top of the figure. "Hom. ref." (red points) stands for "homozygous for the reference allele", "Het." (Blue points) stands for "heterozygous" and "Hom. alt." (green points) stands for "homozygous for the alternative allele". The coordinate of the SNP is given in the title of each panel. The red horizontal solid line represents the highest value observed in experimental blanks for each pseudo-molecule. Identical plots were produced for all pseudo-molecules for which we found significant associations and are presented in Fig. S5. In the eight panels of **(b)**, violin plots represent the distribution of log intensity values for each genotype (x-axis), in leaves from high (top row) and low branches (bottom row), at the most strongly associated SNP for each pseudo-molecule.

Table 1 Annotations of 5 pseudo-molecules (Mol) and their associations and candidate genes at the highest SNPs associated with the phenotype (when candidate genes were found within 6,000bp of the best association).

Pseudo-molecule ID	**Chemical Class**	**Putative Name**	**Chemical Formula**	**PubChem CID**	**MS-FINDER Total score**	**RT (min)**	**m/z$^+$**	**Chr: Position**	**Association score**	***Quercus robur* candidate gene**	***Quercus robur* gene name**	***A. thaliana* candidate gene**	***A. thaliana* gene name**
m216	Gallic acid and derivatives	Norbergenin	C13H14O9	90476206	6.0479	5.82	315.07	Chr01:36056531	26.32			AT3G16520.2	UDP-GLUCOSYL TRANSFERASE 88A1
m25	cinnamic derivatives; quinic acid derivatives	Caffeoylquinic acid	C16H18O9	1794427	7.8264	4.30	163.04	Chr11:40609749	16.56				
m269	Quinic acid and derivatives	Theogallin, 3-Galloylquinic acid	C14H16O10	442988	6.5784	3.38	345.08	Chr01:19760198	59.74	Qrob_P0270670.2	serine carboxypeptidase like clade I	AT2G22990.6	FPT2,SCPL8
m39		Ferulic acid-like	C10H8O3	445858		5.05	177.05	Chr07:27514866	48.42			AT2G36290.1	alpha/beta-Hydrolases protein superfamily
m514	Flavonoid-3-O-glycosides	Kaempferol-hexoside-rhamnoside	C30H26O13	21606527	7.7587	7.56	595.14	Chr02:63715454	12.39			AT1G28570.3	SGNH hydrolase-type esterase superfamily protein

Based on structural annotations, the pseudo-molecules were assigned to a chemical class and given a putative name. The RT column gives the retention time in min, the m/z+ column gives the mass to charge ratio (in positive mode) for the peak chosen as representative for the pseudo-molecule (which doesn't always correspond to the protonated molecule). The columns "Chr." and "Position" give the coordinates, on the *Q. robur* reference genome, of the genetic marker most strongly associated with the variation of each pseudo-molecule. The column "association score" is the highest association score ($-\log^{10}$(*p-value*)) observed for the genetic marker across both branch heights. The columns "candidate gene" and "gene name" correspond to candidate genes identified within 3 kb of the associated marker, either on the *Quercus robur* reference genome, or based on sequence similarity with *Arabidopsis thaliana* proteins. Structural annotations for the 16 other pseudo-molecules with strong associations are available in Dataset S4. Dataset S3 gives the full list of significant associations detected in our study.

Molecule annotation and candidate genes for leaf specialized metabolites with high association to genetic markers

For the 21 pseudo-molecules with high associations within mQTLs, we sought to confirm their chemical structure using spectral proof from MS^n analyses. Putative structures were proposed for only 12 of these molecules by the automated annotation pipeline (see Material and Methods). Of these, we were able to validate the structures of three, modify the proposed structure of four molecules (Dataset S4), and reject the proposed structures for five molecules. The seven manually validated molecules represented six different chemical structures that belonged to five different ontology chemical classes: flavonoids, quinic and gallic acid derivatives, lignan glycosides and cinnamic acids (Dataset S4).

We aimed to determine if the 21 pseudo-molecules were associated with SNPs located within or near annotated genes in the *Quercus robur* genome. For 9 pseudo-molecules, we found annotated genes in a 6 kb window around the most associated SNP. This allowed identifying 12 candidate genes located in 12 mQTL. In addition, we searched for sequence similarities between the 6 kb regions around associated SNPs and *Arabidopsis thaliana* proteins. This allowed identifying 14 additional candidate genes and confirmed all of the other annotations. For the 7 pseudo-molecules with validated structural annotations, only molecule m269 had significant associations located near genes on chromosome 1 and 2 of the *Quercus robur* genome. For four other annotated pseudo-molecules, m25, m39, m216 and m514, we identified six additional candidate *Arabidopsis* genes. These five metabolites were annotated as galloylquinic acid, norbergenin, ferulic acid like and caffeoylquinic acid as derived compounds of phenylpropanoid pathway and kaempferol-rhamnoside (Table **1**).

Discussion

Oak populations display low genetic differentiation

We observed low genetic differentiation among the nine sessile oak populations , which is consistent with previously published observations, for the same populations, and in oaks more generally (Leroy *et al.*, 2020; Saleh *et al.*, 2022). The lack of differentiation among oak populations likely results from large population sizes, limiting genetic drift, the continuous

distribution of the species and extensive long distance seed and pollen dispersal (Gerber *et al.*, 2014).

Despite very low average genetic differentiation over the genomes, a small number of loci displayed differentiated allele frequency. In a previous study, (Torres-Ruiz *et al.*, 2019) studied the variation of leaf phenology, growth or hydraulic traits in the same populations and identified phenotypic differentiation for phenology and growth traits. It is possible that differentiated loci detected here could contribute to the variation of these traits (including leaf SMs, see next section) and may be involved in the local adaptation of the nine populations to their respective environments.

Phenotypic and genetic variation of most leaf specialized metabolites investigated displayed extensive within population variation and low differentiation among sessile oak populations.

We quantified the variation of 219 pseudo-molecules within the canopies of 225 individuals originating from nine populations. Our first observation was a large effect of branch height (Fig. **2a**). This effect impacted nearly all pseudo-molecules and explained a large fraction of the overall phenotypic variance across the 219 pseudo-molecules. Such variation was previously reported in pedunculate oak for total phenolics (Valdés-Correcher *et al.*, 2020; Volf *et al.*, 2022), suggesting that different environmental factors within the canopy such as the exposure to direct sunlight or to a different set of biotic interactions influences the production of leaf SMs. While sampling height appeared to influence the abundance of nearly all molecules, phenotypic variation among trees in high and low branches was generally highly correlated (Fig. **2c**).

At both branch heights we found very little phenotypic differentiation among populations, both in multivariate analysis and when considering molecules individually (Fig. **2d**). In addition, our genome-wide association analysis detected very strong associations for nearly 10% of the pseudo-molecules we investigated (21 out of the 219), and significant associations for almost 63%, suggesting that the variation of non-volatile leaf SMs within populations was to a large extent genetically determined by a small number of loci. On the one hand, this high heritability and the simple architecture of SMs variation is consistent with estimates in other

species such as rice (Matsuda *et al.*, 2015) or *Arabidopsis thaliana* (Brachi *et al.*, 2015). On the other hand, however, the lack of differentiation among populations is consistent with observations in natural oak stands (Bertić *et al.*, 2021).

Taken together, our results suggest that the variation of leaf SMs we measured was generally not involved in the local adaptation of populations to their respective environments, at least at the geographical scale considered. Previous studies of SMs variation in common gardens observed patterns consistent with local adaptation in plants (Brachi *et al.*, 2015) or in trees (O'Reilly-Wapstra *et al.*, 2013; Meijón *et al.*, 2016). However, differences between environments may also produce patterns of local adaptation as for example in oaks (Bertić *et al.*, 2021). While we quantified the variation of many molecules, using untargeted metabolomics, it is possible we did not capture the variation of locally adaptive molecules. This could have two plausible reasons. First, we explored the variation of leaf SMs between populations grown in a common garden at one time point during the summer. While using a common garden trial allowed to remove the environmental effect, it is possible that phenotypes that would display patterns of variation consistent with local adaptation were not expressed in our phenotyping conditions. Sampling additional time points during the year, or quantifying the variation of leaf in a different environment may reveal more phenotypic variation among populations (Meijón *et al.*, 2016). Second, the phenotypic variation observed within each population may not be representative of the variation in leaf SMs among adult trees in natural populations from which acorns were collected (Ducousso *et al.*, 2022). In each population, acorns were collected, heat treated to avoid infection by pathogens and then raised in controlled conditions for 3 years. This protocol allowed to reduce material loss, however, it may have promoted the development of seedlings that would not have developed under natural conditions.

Consistent with the lack of phenotypic differentiation among populations, the vast majority of markers associated with leaf SMs were not significantly less differentiated between populations (low F_{ST} values) than random loci. One interpretation could be that the variation in leaf SMs evolves neutrally. This is unlikely in our opinion as drift would likely generate differentiation between populations, and leaf SMs are known to impact fitness related traits such as stress resistance and defence against pests (De-la-Cruz *et al.*, 2020). At least three other scenarios could be formulated to explain the very high level of variation within provenances for leaf SMs, and in different scenarios the balanced frequency of phenotypes in

all populations (see Moore *et al.*, 2014 for a review). A first scenario could be that phenotypic variation among adult trees within populations of origin is geographically structured at a fine scale, due to adaptation to micro-local conditions. Locally this would result in a diverse pollen cloud that could generate extensive phenotypic variation among offsprings. This scenario would likely require strong and recurrent selection on seedling populations that vary over relatively small distances (maybe in order of a few kilometres, Gerber *et al.*, 2014) to generate the geographical structure of phenotypic variation in adult populations. A recent study explored the mechanisms that maintain the high variation in leaf glucosinolates across the species range of the perennial wildflower (*Boechera stricta*). It showed that variation for the production of certain glucosinolates was maintained because selection by herbivores and exposure to drought varied over small geographic distances (Carley *et al.*, 2021). The second scenario would be analogous to the first one, but instead of varying over the landscape, selection would vary in time, favouring different phenotypes depending on the year. Investigating allele frequencies at associated markers in trees of different ages could be a way to test this scenario (Saleh *et al.*, 2022). Finally, in a third scenario, variation could be maintained by negative frequency dependent selection. For example, a particular chemotype (or the production of a specific molecule) could be favourable when rare in the population. The rarer chemotypes could for example deter herbivores adapted to the most common chemotype in the population and therefore be advantageous. As the rarer chemotype increases in frequency in the populations, its advantage would decrease as the local insect community adapts. Examples are rare in the literature (Núñez-Farfán *et al.*, 2007), but the mechanism would be analogous to those observed in plant-pathogen interactions (Karasov *et al.*, 2014).

Potential pleiotropic effect of leaf specialized metabolites associated genes

Specialized metabolites, as single molecules or chemotypes, play a key role in multiple stress responses. In oaks, previous studies on leaf SMs abundance variation showed that variation of different SMs, such as quinic acids or quercitols, were shaped by both biotic stresses (Sardans *et al.*, 2014; Bertić *et al.*, 2021) and abiotic stresses (Passarinho *et al.*, 2006; Aranda *et al.*, 2021). Their multiple roles may be reflected by a pleiotropic effect of the biosynthetic genes associated with their variation. As hypothesised in the previous section, selective pressure variation over time and space may act differently on biosynthetic genes and

transcription factors and favour different alleles creating patterns of balancing selection within populations (Carley *et al.*, 2021).

Here, we only discuss possible pleiotropic effects of three metabolites for which we manually validated the structure, identified a relevant candidate gene, and for which literature searches highlighted potential pleiotropic effects on tolerance or resistance to both abiotic and biotic challenges.

The first molecule of interest, m39, was annotated as a ferulic acid like metabolite, derived from the phenylpropanoid pathway, and associated with a single major-effect locus (chromosome 7; position 27,514,866 bp) with a sequence similarity in *Arabidopsis* with an alpha/beta-Hydrolases protein superfamily. Ferulic acids play a key role in plant cell wall biosynthesis and are involved in drought tolerance of cereals (Hura *et al.*, 2007). In addition, ferulic acids may also contribute to herbivore resistance in oaks. Their absolute quantity in leaves, quantified as lignin equivalent, were positively correlated with the growth rate of caterpillars of a generalist herbivore, *Lymantria dispar* (Damestoy *et al.*, 2019).

The second molecule was annotated as a kaempferol derivative (m514) which belongs to the flavonoid class and was associated with another major-effect locus (chromosome 2; position 63,715,454 bp) with high similarity to a hydrolase-type esterase superfamily protein in *Arabidopsis*. Kaempferol based molecules contribute to increased resistance of pedunculate oaks to a specialist herbivore *Tortrix viridana* (Bertić *et al.*, 2021) and to increased drought response of pubescent oaks (*Q. pubescens*) (Saunier *et al.*, 2022). Both alpha/beta-Hydrolases and hydrolase-type esterase protein superfamilies were previously described to be involved in specialized metabolism in plants (review in Mindrebo *et al.*, 2016).

The third molecule was annotated as a galloylquinic acid or theogallin (m269) and we found a very strong association for this theogallin, with a single marker on chromosome 1, at position 19,760,198 bp. We identified one candidate gene in the vicinity of SNPs associated with the variation of this theogallin, annotated as a serine carboxypeptidase like (SCPL) in the oak genome. SCPL genes were previously shown to play a key role in metabolomic biosynthesis pathways of flavonoids in grapevine *Vitis vinifera* (Bontpart *et al.*, 2018) and galloylated catechins in *Camellia sinensis* (Ahmad *et al.*, 2020). In oaks, galloylquinic acids were previously identified within leaves of 12 species of black and white oaks (Yarnes *et al.*, 2006), but, to our knowledge, no study investigated the biological function of this molecule. Galloylquinic acid is composed of a quinic acid and a gallic acid moiety and is a precursor of hydrolysable tannins. Previous studies have investigated the properties of quinic acid and

gallic acid in oaks, however separately. Quinic acid concentrations were shown to increase in response to wounding (Sardans *et al.*, 2014), temperature elevation (Passarinho *et al.*, 2006) and water deprivation (Aranda *et al.*, 2021). In addition, the production of gallic acid was frequently associated with herbivore resistance, and its concentration appears to vary seasonally in oak leaves (Salminen *et al.*, 2004).

In conclusion, we observed extensive variation of leaf SMs within sessile oak populations samples over a latitudinal cline, and few molecules displayed significant differentiation among populations (14%). We found significant associations for 63% of the leaf SM investigated suggesting both that the variation of individual molecules within populations often had high heritability and simple genetic architecture. Thus, our results suggest that most leaf SMs with heritable variation within provenances, were not involved in the local adaptation of populations to their respective environment.

Instead, we found very high levels of genetic variation for leaf SM in all nine European populations investigated. This pattern may be the result of two evolutionary processes. On the one hand, the variation could be mostly selectively neutral. In this case, our results would show the extent of the phenotypic and genetic variation present in oak populations, for specialized metabolites. Hence, this large variability could be an important source of novel phenotypes, affording populations the potential to adapt to extent or emergent changes in their environment. On the other hand, the extensive genetic variation for leaf SMs we observed within all populations could be maintained by natural selection. While we discussed different possible scenarios that could explain the maintenance of this genetic variation, understanding the selective pressures that shape this variation and the ecological role it plays in forests is essential in the context of global change. Our results lay the foundation for detailed analysis of the selective pressures and the ecological consequences of the genetic variation of leaf SMs in oak dominated forests.

Acknowledgments

This work was funded by successive grants to C.P. and B.B. between (2016-2019) from the Idex of the university of Bordeaux, Action Thématique Transversale: METAB-OAK "Natural variation of secondary metabolite production in forest trees: the case of European white oaks". B.B. has received the support of the European Union in the framework of the Marie-Curie FP7 COFUND People Programme, through the award of an

AgreenSkills/AgreenSkills+ fellowship (under Grant Agreement 267196). This work was also made possible by MetaboHUB (ANR-11-INBS-0010). D.C.E. received funding for a PhD from the INRAE department ECODIV and Clermont Auvergne Métropole. Computations and data storage for this work were provided by the Bordeaux Bioinformatics Center at the University of Bordeaux (CBIB, https://www.cbib.u-bordeaux.fr), the Genotoul bioinformatics platform in Toulouse, Occitanie, France (Bioinfo Genotoul, http://bioinfo.genotoul.fr), and Le Mésocentre de Calcul Intensif Aquitain (MCIA), France.

Competing interests

None declared.

Author contributions

B.B. and C.P. designed the study, G.LP., F.B., D.CE. and B.B. collected samples, B.B., D.CE., and C.L. extracted DNA and performed quality controls. B.B., S.B. and A.M. designed and optimised LC-MS methods. B.B., S.B. and D.CE. performed LC-MS analysis. A.K. and A.D. installed the oak common garden and contributed information about provenances and its experimental design. B.B. and D.CE. performed the analyses with guidance from S.B., wrote the manuscript. C.P., A.M., S.B., A.K. provided comments on the manuscript.

Data Availability

Codes and pipelines used for this study is available at:

1. https://forgemia.inra.fr/domitille.coq-etchegaray/gw-oak-snp-pipeline.git
2. https://forgemia.inra.fr/domitille.coq-etchegaray/gwas-gwoak-metaboak-pipeline.git
3. https://forgemia.inra.fr/domitille.coq-etchegaray/metab-oak-pipeline.git

Raw metabolomics data and SNPs matrix will be available upon publication in the Dataverse INRAE Biogeco: doi:10.57745/3CSJZ9

Raw WGS sequencing used for this study will be available upon publication in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB60751. (https://www.ebi.ac.uk/ena/browser/view/PRJEB60751).

References

Ahmad MZ, Li P, She G, Xia E, Benedito VA, Wan XC, Zhao J. **2020**. Genome-Wide

Analysis of Serine Carboxypeptidase-Like Acyltransferase Gene Family for Evolution and Characterization of Enzymes Involved in the Biosynthesis of Galloylated Catechins in the Tea Plant (Camellia sinensis). *Frontiers in Plant Science* **11**: 848.

Aranda I, Cadahía E, Fernández De Simón B. **2021**. Specific leaf metabolic changes that underlie adjustment of osmotic potential in response to drought by four Quercus species. *Tree Physiology* **41**: 728–743.

Bailey JK, Deckert R, Schweitzer JA, Rehill BJ, Lindroth RL, Gehring C, Whitham TG. **2005**. Host plant genetics affect hidden ecological players: links among Populus , condensed tannins, and fungal endophyte infection. *Canadian Journal of Botany* **83**: 356–361.

Bartholomé J, Brachi B, Marçais B, Mougou-Hamdane A, Bodénès C, Plomion C, Robin C, Desprez-Loustau ML. **2020**. The genetics of exapted resistance to two exotic pathogens in pedunculate oak. *New Phytologist* **226**: 1088–1103.

Bednarek P. **2012**. Chemical warfare or modulators of defence responses-the function of secondary metabolites in plant immunity This review comes from a themed issue on Biotic interactions. *Current Opinion in Plant Biology* **15**: 407–414.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. **2015**. The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *genesis* **53**: 474–485.

Bertić M, Schroeder H, Kersten B, Fladung M, Orgel F, Buegger F, Schnitzler JP, Ghirardo A. **2021**. European oak chemical diversity – from ecotypes to herbivore resistance. *New Phytologist* **232**: 818–834.

Bhatia G, Patterson N, Sankararaman S, Price AL. **2013**. Estimating and interpreting FST: The impact of rare variants. *Genome Research* **23**: 1514.

Bontpart T, Ferrero M, Khater F, Marlin T, Vialet S, Vallverdù-Queralt A, Pinasseau L, Ageorges A, Cheynier V, Terrier N. **2018**. Focus on putative serine carboxypeptidase-like acyltransferases in grapevine. *Plant Physiology and Biochemistry* **130**: 356–366.

Brachi B, Meyer CG, Villoutreix R, Platt A, Morton TC, Roux F, Bergelson J. **2015**. Coselected genes determine adaptive variation in herbivore resistance throughout the native range of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 4032–4037.

Bürkner P-C. **2018**. Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal* **10**: 395–411.

Burow M, Halkier BA, Kliebenstein DJ. **2010**. Regulatory networks of glucosinolates

shape *Arabidopsis thaliana* fitness. *Current opinion in plant biology* **13**: 348–353.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. **2009**. BLAST+: Architecture and applications. *BMC Bioinformatics* **10**: 1–9.

Carley LN, Mojica JP, Wang B, Chen C-Y, Lin Y-P, Prasad KVSK, Chan E, Hsu C-W, Keith R, Nuñez CL, *et al.*2021. Ecological factors influence balancing selection on leaf chemical profiles of a wildflower HHS Public Access. *Nat Ecol Evol* **5**: 1135–1144.

Carsten Kuhl A, Tautenhahn R, Treutler H, Neumann S, Steffen Neumann M. **2021**. Package 'CAMERA' Title Collection of annotation related methods for mass spectrometry data.

Chambers MC, MacLean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, *et al.*2012. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology 2012 30:10* **30**: 918–920.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. **2015**. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4**: 7.

Damestoy T, Brachi B, Moreira X, Jactel H, Plomion C, Castagneyrol B. **2019**. Oak genotype and phenolic compounds differently affect the performance of two insect herbivores with contrasting diet breadth. *Tree Physiology* **39**: 615–627.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, *et al.*2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10**: 1–4.

De Luca V, St Pierre B. **2000**. The cell and developmental biology of alkaloid biosynthesis. *Trends in Plant Science* **5**: 168–173.

Dearing MD, Foley WJ, McLean S. **2005**. The Influence of Plant Secondary Metabolites on the Nutritional Ecology of Herbivorous Terrestrial Vertebrates. *https://doi.org/10.1146/annurev.ecolsys.36.102003.152617* **36**: 169–189.

De-la-Cruz IM, Merilä J, Valverde PL, Flores-Ortiz CM, Núñez-Farfán J. **2020**. Genomic and chemical evidence for local adaptation in resistance to different herbivores in Datura stramonium. *Evolution* **74**: 2629–2643.

Ducousso A, Ehrenmann F, Girard Q, Lamy JB, Louvet JM, Reynet P, Musch B, Kremer A. **2022**. Long-term and large-scale Quercus petraea population survey conducted in provenance tests installed in France. *Annals of Forest Science* **79**: 1–10.

Forkner RE, Marquis RJ, Lill JT. **2004**. Feeny revisited: condensed tannins as anti-herbivore defences in leaf-chewing herbivore communities of Quercus. *Ecological Entomology* **29**: 174–187.

Fraisier-Vannier O, Chervin J, Cabanac G, Puech V, Fournier S, Durand V, Amiel A, André O, Benamar OA, Dumas B, *et al.*2020. MS-CleanR: A Feature-Filtering Workflow for Untargeted LC-MS Based Metabolomics. *Analytical Chemistry* **92**: 9971–9981.

Galeotti F, Barile E, Curir P, Dolci M, Lanzotti V. **2008**. Flavonoids from carnation (Dianthus caryophyllus) and their antifungal activity. *Phytochemistry Letters* **1**: 44–48.

Gerber S, Chadoeuf J, Gugerli F, Lascoux M, Buiteveld J, Cottrell J, Dounavi A, Fineschi S, Forrest LL, Fogelqvist J, *et al.*2014. High rates of gene flow by pollen and seed in oak populations across Europe. *PLoS ONE* **9**.

Hura T, Grzesiak S, Hura K, Thiemt E, Tokarz K, Wędzony M. **2007**. Physiological and Biochemical Tools Useful in Drought-Tolerance Detection in Genotypes of Winter Triticale: Accumulation of Ferulic Acid Correlates with Drought Tolerance. *Annals of Botany* **100**: 767–775.

Joshi N, Fass J. **2011**. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files.

Jun G, Wing MK, Abecasis GR, Kang HM. **2015**. An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. *Genome Research* **25**: gr.176552.114.

Karasov TL, Kniskern JM, Gao L, Deyoung BJ, Ding J, Dubiella U, Lastra RO, Nallu S, Roux F, Innes RW, *et al.*2014. The long-term maintenance of a resistance polymorphism through diffuse interactions. *Nature* **512**: 436–440.

Kersten B, Ghirardo A, Schnitzler JP, Kanawati B, Schmitt-Kopplin P, Fladung M, Schroeder H. **2013**. Integrated transcriptomics and metabolomics decipher differences in the resistance of pedunculate oak to the herbivore Tortrix viridana L. *BMC Genomics* **14**: 737.

Köster J, Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, *et al.*2021. Sustainable data analysis with Snakemake. *F1000Research 2021 10:33* **10**: 33.

Kurtzer GM, Sochat V, Bauer MW. **2017**. Singularity: Scientific containers for mobility of compute. *PLOS ONE* **12**: e0177459.

Leroy T, Louvet JM, Lalanne C, Le Provost G, Labadie K, Aury JM, Delzon S, Plomion C, Kremer A. **2020**. Adaptive introgression as a driver of local adaptation to climate in European white oaks. *New Phytologist* **226**: 1171–1182.

Li H. **2013**. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Libiseller G, Dvorzak M, Kleb U, Gander E, Eisenberg T, Madeo F, Neumann S,

Trausinger G, Sinner F, Pieber T, *et al.*2015. IPO: A tool for automated optimization of XCMS parameters. *BMC Bioinformatics* **16**: 1–10.

Martin M. **2011**. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

Maruki T, Lynch M. **2015**. Genotype-frequency estimation from high-throughput sequencing data. *Genetics* **201**: 473–486.

Maruki T, Lynch M. **2017**. Genotype calling from population-genomic sequencing data. *G3: Genes, Genomes, Genetics* **7**: 1393–1404.

Matsuda F, Nakabayashi R, Yang Z, Okazaki Y, Yonemaru J, Ebana K, Yano M, Saito K. **2015**. Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. *The Plant Journal* **81**: 13–23.

McCormick AC, Unsicker SB, Gershenzon J. **2012**. The specificity of herbivore-induced plant volatiles in attracting herbivore enemies. *Trends in Plant Science* **17**: 303–310.

Meijón M, Feito I, Oravec M, Delatorre C, Weckwerth W, Majada J, Valledor L. **2016**. Exploring natural variation of Pinus pinaster Aiton using metabolomics: Is it possible to identify the region of origin of a pine from its metabolites? *Molecular Ecology* **25**: 959–976.

Mindrebo JT, Nartey CM, Seto Y, Burkart MD, Noel JP. **2016**. Unveiling the functional diversity of the Alpha-Beta hydrolase fold in plants. *Current opinion in structural biology* **41**: 233.

Mitchell RJ, Bellamy PE, Ellis CJ, Hewison RL, Hodgetts NG, Iason GR, Littlewood NA, Newey S, Stockan JA, Taylor AFS. **2019**. Collapsing foundations: The ecology of the British oak, implications of its decline and mitigation options. *Biological Conservation* **233**: 316–327.

Moore BD, Andrew RL, Külheim C, Foley WJ. **2014**. Explaining intraspecific diversity in plant secondary metabolites in an ecological context. *New Phytologist* **201**: 733–750.

Nakabayashi R, Yonekura-Sakakibara K, Urano K, Suzuki M, Yamada Y, Nishizawa T, Matsuda F, Kojima M, Sakakibara H, Shinozaki K, *et al.*2014. Enhancement of oxidative and drought tolerance in Arabidopsis by overaccumulation of antioxidant flavonoids. *The Plant Journal* **77**: 367–379.

Newcombe G. **1998**. A review of exapted resistance to diseases of Populus. *European Journal of Forest Pathology* **28**: 209–216.

Núñez-Farfán J, Fornoni J, Valverde P. **2007**. The Evolution of Resistance and Tolerance to Herbivores. *Annual Review of Ecology and Evolution* **38**: 541–566.

Obata T, Witt S, Lisec J, Palacios-Rojas N, Florez-Sarasa I, Yousfi S, Araus JL, Cairns

JE, Fernie AR. **2015**. Metabolite Profiles of Maize Leaves in Drought, Heat, and Combined Stress Field Trials Reveal the Relationship between Metabolism and Grain Yield. *Plant Physiology* **169**: 2665–2683.

O'Reilly-Wapstra JM, Miller AM, Hamilton MG, Williams D, Glancy-Dean N, Potts BM. **2013**. Chemical Variation in a Dominant Tree Species: Population Divergence, Selection and Genetic Stability across Environments. *PLoS ONE* **8**: 58416.

Passarinho JAP, Lamosa P, Baeta JP, Santos H, Ricardo CPP. **2006**. Annual changes in the concentration of minerals and organic compounds of Quercus suber leaves. *Physiologia Plantarum* **127**: 100–110.

Perez De Souza L, Alseekh S, Brotman Y, Fernie AR. **2020**. Network-based strategies in metabolomics data analysis and interpretation: from molecular networking to biological interpretation. *Expert review of proteomics* **17**: 243–255.

Plomion C, Aury J-M, Amselem J, Leroy T, Murat F, Duplessis S, Faye S, Francillonne N, Labadie K, Le Provost G, *et al.*2018. Oak genome reveals facets of long lifespan. *Nature Plants* **4**: 440–452.

Revelle W. **2022**. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois: Northwestern University.

Rozema J, Björn LO, Bornman JF, Gaberščik A, Häder DP, Trošt T, Germ M, Klisch M, Gröniger A, Sinha RP, *et al.*2002. The role of UV-B radiation in aquatic and terrestrial ecosystems-An experimental and functional analysis of the evolution of UV-absorbing compounds. *Journal of Photochemistry and Photobiology B: Biology* **66**: 2–12.

Sáenz-Romero C, Lamy JB, Ducousso A, Musch B, Ehrenmann F, Delzon S, Cavers S, Chałupka W, Dağdaş S, Hansen JK, *et al.*2017. Adaptive and plastic responses of Quercus petraea populations to climate across Europe. *Global Change Biology* **23**: 2831–2847.

Saleh D, Chen J, Leplé JC, Leroy T, Truffaut L, Dencausse B, Lalanne C, Labadie K, Lesur I, Bert D, *et al.*2022. Genome-wide evolutionary response of European oaks during the Anthropocene. *Evolution Letters* **6**: 4–20.

Salminen J-P, Roslin T, Karonen M, Sinkkonen J, Pihlaja K, Pulkkinen P. **2004**. SEASONAL VARIATION IN THE CONTENT OF HYDROLYZABLE TANNINS, FLAVONOID GLYCOSIDES, AND PROANTHOCYANIDINS IN OAK LEAVES. *Journal of Chemical Ecology* **30**.

Sardans J, Gargallo-Garriga A, Pérez-Trujillo M, Parella TJ, Seco R, Filella I, Peñuelas J. **2014**. Metabolic responses of Quercus ilex seedlings to wounding analysed with nuclear magnetic resonance profiling. *Plant Biology* **16**: 395–403.

Sardans J, Peñuelas J, Rivas-Ubach A. **2011**. Ecological metabolomics: Overview of current developments and future challenges. *Chemoecology* **21**: 191–225.

Saunier A, Greff S, Blande JD, Lecareux C, Baldy V, Fernandez C, Ormeño E. **2022**. Amplified Drought and Seasonal Cycle Modulate Quercus pubescens Leaf Metabolome. *Metabolites* **12**.

Schweitzer JA, Bailey JK, Fischer DG, Leroy CJ, Lonsdorf EV, Whitham TG, Hart SC. **2008**. PLANT-SOIL-MICROORGANISM INTERACTIONS: HERITABLE RELATIONSHIP BETWEEN PLANT GENOTYPE AND ASSOCIATED SOIL MICROORGANISMS. *Ecology* **89**: 773–781.

Sebastiana M, Gargallo-Garriga A, Sardans J, Pérez-Trujillo M, Monteiro F, Figueiredo A, Maia M, Nascimento R, Silva MS, Ferreira AN, *et al.*2021. Metabolomics and transcriptomics to decipher molecular mechanisms underlying ectomycorrhizal root colonization of an oak tree. *Scientific Reports* **11**: 8576.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. **2003**. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**: 2498–2504.

Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. **2006**. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry* **78**: 779–787.

Torres-Ruiz JM, Kremer A, Carins Murphy MR, Brodribb T, Lamarque LJ, Truffaut L, Bonne F, Ducousso A, Delzon S. **2019**. Genetic differentiation in functional traits among European sessile oak populations. *Tree Physiology* **39**: 1736–1749.

Valdés-Correcher E, Bourdin A, González-Martínez SC, Moreira X, Galmán A, Castagneyrol B, Hampe A. **2020**. Leaf chemical defences and insect herbivory in oak: accounting for canopy position unravels marked genetic relatedness effects. *Annals of botany* **126**: 865–872.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, *et al.*2013. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics* **11**: 11.10.1.

Vargas P, Farias GA, Nogales J, Prada H, Carvajal V, Barón M, Rivilla R, Martín M, Olmedilla A, Gallegos MT. **2013**. Plant flavonoids target Pseudomonas syringae pv. tomato DC3000 flagella and type III secretion system. *Environmental microbiology reports* **5**: 841–850.

Volf M, Volfová T, Seifert CL, Ludwig A, Engelmann RA, Jorge LRé., Richter R, Schedl A, Weinhold A, Wirth C, *et al.*2022. A mosaic of induced and non-induced branches promotes variation in leaf traits, predation and insect herbivore assemblages in canopy trees. *Ecology Letters* **25**: 729–739.

Walker TS, Bais HP, Grotewold E, Vivanco JM. **2003**. Root exudation and rhizosphere biology. *Plant physiology* **132**: 44–51.

Weir BS, Cockerham CC. **1984**. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**: 1358.

Wink M. **2018**. Plant Secondary Metabolites Modulate Insect Behavior-Steps Toward Addiction? *Frontiers in Physiology* **9**.

Yarnes CT, Boecklen WJ, Tuominen K, Salminen JP. **2006**. Defining phytochemical phenotypes: Size and shape analysis of phenolic compounds in oaks (Fagaceae, Quercus) of the Chihuahuan Desert. *Canadian Journal of Botany* **84**: 1233–1248.

Zust T, Heichinger C, Grossniklaus U, Harrington R, Kliebenstein DJ, Turnbull LA. **2012**. Natural enemies drive geographic variation in plant defenses. *Science* **338**: 116–119.

Supporting Information

Table S1. Geographic origin of the populations and number of oak trees analysed per population.

Table S2. Optimised parameters used for peak picking and peak alignment for metabolomics data treatment with the R package XCMS.

Fig S1. Density of markers genotyped along the 12 pseudo-chromosomes of the *Quercus robur* reference genome.

Fig S2. Decay of linkage disequilibrium in our sample of 225 *Quercus petraea* from 9 European populations.

Fig S3. Phenotypic variation within and among populations for the five molecules displaying the most differentiation among populations.

Fig S4. Correlation networks of pseudo-molecules intensities in leaves sampled on low (A) and high (B) branches.

Fig S5. Phenotypic variation explained by genetic variation at the most associated marker for each pseudo-molecule.

Fig S6. Genome-wide distribution of Weir and Cockerham F_{ST} and F_{ST} values of genetic markers associated with 21 strongly associated pseudo-molecules.

Dataset 1: Pseudo-molecules intensities, sampling and peak descriptions.

Dataset 2: Formula and automatic structural annotation of pseudo-molecules based on MS2 acquisitions.

Dataset 3: GWAs associations results in mQTL clusters.

Dataset 4: Structural annotations of 21 pseudo-molecules, associations with genetic markers and candidate genes.