# A reference genome assembly for *Quercus canariensis* Willd


## Abstract

We present a genome assembly from a specimen of *Quercus canariensis* (Fagaceae; Fagales; Magnoliopsida). The assembly was generated using PacBio HiFi long reads with an approximate sequencing depth of 39X and scaffolded using a reference-guided approach. The genome sequence has a total length of 816.0 megabases for haplotype 1 and 804.8 megabases for haplotype 2. The two haplotypes are each resolved into 12 chromosomal pseudomolecules, with only 3.48% and 1.36% of sequences remaining unplaced in haplotypes 1 and 2, respectively. Assembly completeness is supported by BUSCO scores of 98.3% and 98.2% complete genes for haplotypes 1 and 2, respectively. Structural annotation identified 51,882 and 46,482 protein-coding genes in haplotypes 1 and 2, respectively. This genome assembly provides the first chromosome-scale reference genome for *Q. canariensis*, laying the base for future genomic and evolutionary studies in this understudied species of the hybridizing white oak species complex.



**Key words :** *Quercus canariensis*, genome assembly, chromosome-scale genome, PacBio HiFi sequencing, European white oak species complex



**Authors : F. Couturier[1], C. Cravero[2], I. Lesur[1], J. Confais[3], E. Belmonte[4], L. Piat[1], W. Marande[2], C. Rellstab[5], M. Valbuena[6], E. Saez-Laguna[1]*, L. Duvaux[1]*++**

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

[2] INRAE, CNRGV, French Plant Genomic Resource Center, Castanet Tolosan 31326, France

[3] Univ. Paris-Saclay, INRAE, BioinfOmics, URGI, 78026 Versailles, France

[4] Univ. Clermont-Auvergne, INRAE, GDEC, 63000 Clermont-Ferrand, France

[5] Swiss Federal Research Institute WSL, 8903 Birmensdorf, Switzerland

[6] Departamento de Sistemas y Recursos Naturales, Research Group FORESCENT, Universidad Politécnica de Madrid, Madrid, Spain

*These authors contributed equally to this work

++Corresponding Author: ludovic.duvaux@inrae.fr

## Taxonomy

Lineage cellular organisms; Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Fagales; Fagaceae; Quercus

EBI:txid568684

*Quercus canariensis* Willd. 1809 (Willdenow)

## Background

European forests are experiencing increasing pressure from recurrent droughts and heatwaves, leading to widespread decline in both coniferous and deciduous tree populations, including European white oaks (Bratu et al., 2025; Knutzen et al., 2025). In the context of ongoing climate warming and northward shifts of suitable climatic envelopes, transferring genetic resources (species or populations) from southern to northern regions may enhance forest adaptive capacity (Kremer et al., 2014). In this context, forest management strategies are increasingly exploring assisted migration/gene flow for the deployment of species/populations better adapted to projected future climatic conditions (Aitken & Bemmels, 2016; Chakraborty et al., 2024). Population and landscape genomic analyses can inform such forest management strategies (Rellstab et al., 2021), but their ability to do so depends on the availability of comprehensive genomic resources, such as high-quality reference genome assembly (Ellegren, 2014).

*Quercus canariensis* Willd. (2n=2x=24) belongs to the section *Quercus* of the *Quercus* genus, which also includes other widespread European white oaks such as *Q. robur, Q. petraea,* and *Q. pubescens* (Hipp et al., 2020). It occupies a rather narrow sub-Mediterranean climatic niche and is naturally distributed across southern Spain and North Africa (Fig. 1). However, given its Mediterranean ecological preferences and its evolutionary history under warmer and drier conditions than those experienced by more northern white oak species (Sánchez-Salguero et al., 2020; Vila-Viçosa et al., 2020), it has recently emerged as a promising candidate for northward reforestation and climate-adaptation programs in France. Moreover, as part of the European white oak species complex, it can hybridize with other white oak species occurring at higher latitudes (Cannon & Petit, 2020), opening the possibility of adaptive introgression that could increase drought and heat tolerance in local populations (Leroy et al., 2020). A detailed characterization of the genome of *Q. canariensis* is therefore key to clarifying evolutionary relationships within the white oak

species complex and to guiding future studies on adaptation and introgression. For instance, the genomic regions underlying drought and heat tolerance remain so far unidentified, and it is unclear whether the associated adaptive alleles are unique to *Q. canariensis* or shared with other white oak species. Furthermore, the distribution and structure of standing genomic variation across its natural range need to be better characterized, together with the detection of signatures of selection across the genome.

To date, genomic resources for this species remain limited (but see Lepais et al., 2022), with no complete reference genome assembly available. Here, we present the first diploid, chromosome-scale reference genome assembly for *Q. canariensis*, providing a foundational resource for evolutionary studies within the European white oaks species complex.

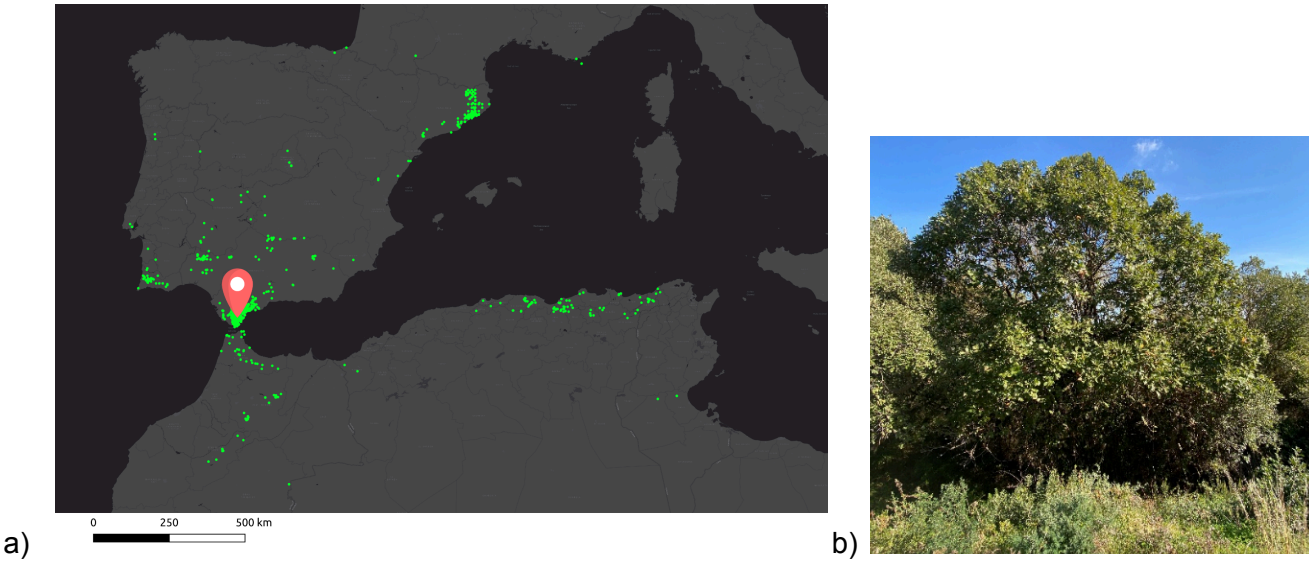


**Figure 1 : a. Geographic distribution of *Quercus canariensis* based on 47,558 georeferenced observations from gbif.org accessed on March 13, 2025 https://doi.org/10.15468/dl.j9e76j and localisation of the specimen used for genome sequencing ; b. Photograph of the sampled *Quercus canariensis* specimen.**

## Results

We generated a PacBio HiFi–based, chromosome-scale diploid reference genome assembly for *Q. canariensis*, comprising two haplotypes scaffolded into 12 chromosomes each, with an average sequencing depth of ~39X and high assembly and annotation completeness (BUSCO scores of 98.3% and 98.2% complete genes).

**Sequencing data, genome assembly and quality assessment**

Sequencing was performed from one individual localized in southern Spain, in the province of Cádiz (Fig. 1). In total, 1,820,767 HiFi reads were produced, yielding 30.7 Gb of sequence data and an approximate sequencing depth of 39X, assuming an haploid genome size of ~783 Mb as estimated by a GenomeScope k-mer analysis. The contig-level assemblies produced 495 contigs for haplotype 1 and 150 contigs for haplotype 2 (Tab. 1). The L50 values were 15 for haplotype 1 and 14 for haplotype 2, while the N50 values reached 20.1 Mb and 16.8 Mb, respectively. Each haplotype was subsequently scaffolded into 12 chromosome-scale sequences using a reference-guided approach. A total of 76 contigs were anchored in haplotype 1 and 87 contigs in haplotype 2. Although a large proportion of contigs remained unplaced (84.6% and 42.0% in haplotypes 1 and 2, respectively), these correspond to only a minor fraction of the assembly length, representing 3.48% and 1.36% of the bases, respectively. Unplaced contigs were gathered into chromosome 0 (chr0). The final assembly sizes reached 816.0 Mb for haplotype 1 (845.4 Mb with chr0) and 804.8 Mb for haplotype 2 (815.9 Mb with chr0).

Both haplotypes showed high contiguity and comparable assembly statistics, indicating consistent assembly quality and a high degree of chromosomal completeness. BUSCO analysis (Table 1) reported 98.3% complete BUSCOs in haplotype 1 (93.4% single-copy and 4.9% duplicated genes), and 98.2% complete BUSCOs in haplotype 2 (92.9% single-copy and 5.3% duplicated genes). The quality of our assemblies was further supported by high long terminal repeat (LTR) Assembly Index (LAI) scores, indicating accurate reconstruction of LTR retrotransposons in repeat-rich regions, reaching 22.99 for haplotype 1 and 24.46 for haplotype 2. Searches for telomeric repeats detected 19 telomeric motifs in haplotype 1 and 16 in haplotype 2, supporting near-complete chromosome assemblies. Consensus accuracy and k-mer completeness were

assessed using Merqury with the PacBio HiFi read dataset. For haplotype 2, Merqury estimated a consensus quality value (QV) of 72.84, corresponding to a base-level error rate of $5.19 \times 10^{-8}$, together with a k-mer completeness of 76.99%. Genome-wide heterozygosity was estimated at 1.9% based on k-mer frequency analysis (GenomeScope), providing an independent measure of allelic divergence between the two haplotypes.

**Structural, functional and transposable element annotation**

A total of 65,629 genes were annotated in haplotype 1 and 55,146 in haplotype 2, including 51,882 and 46,482 mRNA features, respectively. 13,747 and 8,664 non-coding RNA, in haplotype 1 and haplotype 2 respectively, loci were also identified, including rRNA, tRNA and other ncRNA features (Table 2). When considering only the 12 chromosome-scale sequences, 51,343 genes were identified in haplotype 1 and 50,615 in haplotype 2. Annotation completeness was high in both haplotypes, with 96.9% and 97.5% complete BUSCO genes recovered, respectively. Most were single-copy (95.3% and 94.4%), with a small proportion of duplicated genes (1.6% and 3.1%) and few fragmented or missing genes.

Regarding functional annotation (Table 3), performed excluding chr0, InterProScan annotation attributed functional annotations to 46,397 (89.43%) and 42,966 (92.44%) of the *Q. canariensis* proteins for haplotypes 1 and 2, respectively. Additionally, functional assignments were provided by eggNOG for 45,220 (87.16%) and 40,035 (86.13%) haplotypes 1 and 2 proteins, respectively. The E2P2 pipeline further annotated 14,160 haplotype 1 proteins (27.29%) and 13,225 haplotype 2 proteins (28.45%) with enzymatic functions. In summary, a total of 49,252 haplotype 1 (94.93%) and 44,056 haplotype 2 (94.78%) predicted *Q. canariensis* proteins received a functional annotation, leaving only 2,630 (5.07%) and 2,426 (5.22%) proteins unannotated for haplotypes 1 and 2, respectively.

Transposable elements (TEs) account for 54.35% of the genome assembly (haplotype 1). A total of 3,606 TE families and 1,151,899 TE copies were identified, including 3.60% full-length elements. The mean sequence identity across TE families was 81.57% (± 6.82), and the mean length coverage reached 26.19% (± 24.98).

**Table 1 : Genome assembly characteristics and quality indicators for the *Quercus canariensis* reference assembly.**

| | Haplotype 1 | Haplotype 2 |
|---|---|---|
| **Contig-level assembly** | | |
| Contigs | 495 | 150 |
| L50 | 15 | 14 |
| N50 | 20,1 Mb | 16,8 Mb |
| L80 | 33 | 33 |
| N80 | 8,0Mb | 8,2 Mb |
| **Chromosome-level assembly** | | |
| Assembly size (Mb) | 816.0 | 804.8 |
| Number of contigs (assembly) | 76 | 87 |
| Number of chromosomes | 12 | 12 |
| BUSCO completeness | C:98,3%[S:93,4 ; D:4,9%] ; F:1,2% ; M:0,6% ; n:2326 | C:98,2%[S:92,9 ; D:5,3%] ; F:1,2% ; M:0,6% ; n:2326 |
| LAI score | 22.99 | 24.46 |
| Telomeric ends detected | 19 | 16 |

**Table 2 : Structural annotation of the *Quercus canariensis* reference genome assembly.**

| | Haplotype 1 | Haplotype 2 |
|---|---|---|

| | | |
|---|---|---|
| Total genes (mRNA, rRNA, tRNA, ncRNA) | 12 chr<br>51343 (46110, 1306, 579, 3348)<br><br>12 chr + chr0<br>65 629 (51882, 3035, 2047, 8665) | 12 chr<br>50,615 (46196, 552, 595, 3272)<br><br>12 chr + chr0<br>55 146 (46482, 1612, 619, 6433) |
| BUSCO protein | C:96,9%[S:95,3 ; D:1,6%] ; F:2,6% ; M:0,5% ; n:425 | C:97,5%[S:94,4 ; D:3,1%] ; F:2,4% ; M:0,1% ; n:425 |
| Psauron scores | 89.1 | 89.5 |

**Table 3 : Functional annotation of the *Quercus canariensis* reference genome assembly.**

Functional annotation summary for both haplotypes of *Q. canariensis*. Analyses were performed excluding unplaced scaffolds (chr0). Percentages are calculated relative to the total number of predicted proteins per haplotype.

| Annotation method | Haplotype 1 (n, %) | Haplotype 2 (n, %) |
|---|---|---|
| **InterProScan** | 46,397 (89.43%) | 42,966 (92.44%) |
| **eggNOG** | 45,220 (87.16%) | 40,035 (86.13%) |
| **E2P2 (enzymes)** | 14,160 (27.29%) | 13,225 (28.45%) |
| **Total annotated proteins** | 49,252 (94.93%) | 44,056 (94.78%) |

| | | |
|---|---|---|
| **Unannotated proteins** | 2,630 (5.07%) | 2,426 (5.22%) |

## Materials and Methods

### Biological material and sampling

A single adult of *Quercus canariensis* (approximately 40–50 years old) was sampled on January 27, 2024, in southern Spain, in the province of Cádiz (36.17837° N, −5.57642° W), within the Macizo del Aljibe region, at an elevation of 282 m above sea level (Figure 1). This area is characterized by a sub-Mediterranean climate with high annual precipitation (>1000 mm), a prolonged summer drought exceeding three months, and a warm thermal regime with mean annual temperatures above 15 °C and no frost events. The sampling site is located in a humid oak forest dominated by *Q. canariensis*, associated with *Q. suber* (Linnaeus, 1753) and *Olea europaea* var. *sylvestris* (Linnaeus, 1753), developed on clay-rich soils derived from siliceous sandstone substrates. The sampled individual was located at the foot of a stream, in a locally humid microhabitat.

To build the TE library, we sampled an admixed *Q. petraea/pubescens* individual from the mixed stand of Chassagne d'Onnens (beside Lake Neuchâtel, Switzerland), collected in spring 2025 (tree no. 249; Zimmermann et al., 2025).This individual is deposited in public repositories under the BioSample accession SAMEA122115212.

### DNA extraction and sequencing

Single-molecule real-time long reads sequencing was performed at Gentyane Sequencing Platform (Clermont-Ferrand, France) with a PacBio Revio System Sequencer (Pacific Biosciences, Menlo Park, CA, USA). DNA was extracted from leaf buds (1g) with the PacBio® Nanobind® HMW DNA extraction kit using the protocol "Isolating nuclei from plant tissue using TissueRuptor disruption" followed by the protocol "Extracting HMW DNA from plant nuclei using Nanobind® kits". A Femto Pulse (Agilent Technologies, Santa Clara, CA, USA) assay was used to assess the fragments size distribution. The SMRTBell library was prepared using a SMRTbell prep kit 3.0, following the "procedure and checklist -preparing whole

genome and metagenome libraries using SMRTbell prep kit 3.0" protocol. Genomic DNA was sheared using a Megaruptor 3 (Diagenode) generating DNA fragments of approximately 20 kb. A Femto Pulse (Agilent Technologies, Santa Clara, CA, USA) assay was used to assess the fragments size distribution. Sheared genomic DNA were carried into the enzymatic reactions to remove the single-strand overhangs and to repair any damage that may be present on the DNA backbone. An A-tailing reaction followed by the barcoded overhang adapter ligation was conducted to generate the SMRT Bell templates. After nuclease treatment, the sample was then size-selected for fragments above 10kb using the LightBench® (Yourgene Health). The sample was thereafter purified with 1X SMRTbell cleanup beads to obtain the final libraries around 20 kb. The SMRTBell library was quality inspected on a Femto Pulse (Agilent Technologies) and quantified with Qubit dsDNA HS reagent Assay kit (Life Technologies). A ready-to-sequence SMRTBell Polymerase Complex was created using a Revio Polymerase kit (PacBio). The PacBio Revio instrument was programmed to load a 320 pM library and sequenced the sample with Revio Sequencing plate (Pacific Biosciences), acquiring a movie of 24 hours per SMRTcell.

**Genome assembly**

Genome assembly and quality control analyses were performed using the Asm4pg workflow v2.0.0 (https://forge.inrae.fr/asm4pg/GenomAsm4pg/), a fully automated Snakemake-based pipeline designed for efficient long-read genome assembly in pangenomic contexts and the automated generation of R Markdown summary reports. All quality control analyses were performed on the complete assemblies, including unplaced scaffolds (chr0). Genome size and heterozygosity was estimated using Jellyfish v2.3.0 (Marçais & Kingsford, 2011) and GenomeScope v2.0 (Ranallo-Benavidez et al., 2020; Vurture et al., 2017). PacBio HiFi reads were assembled using hifiasm v0.24.0-r703 (Cheng et al., 2021, 2022, 2024), producing a diploid assembly resolved into two haplotypes. Redundant haplotigs and assembly overlaps were identified and removed using the purge_dups python script v1.2.5 (Guan et al., 2020).

Assembly contiguity was evaluated using contig length distributions and N50 statistics computed with GenomeTools v1.5.9 (Gremme et al., 2013) and QUAST v5.2.0 (Mikheenko et al., 2023). Genome-wide k-mer profiles were generated using meryl and Merqury v1.3 (Rhie et al., 2020) to assess assembly completeness and consistency with the sequencing reads. Consensus sequence accuracy and base-level

error rates were also estimated with Merqury based on comparisons between HiFi read k-mers and the assembled genomes. Gene space completeness was evaluated using BUSCO v5.7.1 (Manni et al., 2021) with the embryophyta_odb10.2019-11-20 lineage dataset (n = 2326). Telomeric repeats were identified using the FindTelomeres Python script https://github.com/JanaSperschneider/FindTelomeres. The assembly quality of long terminal repeat retrotransposons was assessed by detecting candidate LTR retrotransposons with LTR_FINDER (Xu & Wang, 2007) followed by filtering and curation using LTR_retriever v3.0.1 (Ou & Jiang, 2018), allowing the calculation of the LTR Assembly Index (LAI) (Ou et al., 2018).

**Reference-guided scaffolding**

Chromosome-scale scaffolding was performed with RagTag v2.1.0 (Alonge et al., 2022) implemented in Asm4pg, using a reference-guided approach. The reference genome required for this step was assembled in this study from publicly available data underlying dhQueRobu3.1, the current NCBI reference assembly for *Q. robur* (Biosample SAMEA9089039, Darwin Tree of Life project). PacBio HiFi reads (ERR8282824, ERR8373983, ERR9682742, ERR9718776) and Hi-C data (ERR8097160) were assembled with Asm4pg in Hi-C mode for scaffolding purposes only. This approach leverages improved Hi-C integration in recent hifiasm versions together with explicit Hi-C scaffolding via yahs (Zhou et al., 2023). The resulting haplotype-resolved assembly (haplotype 1) was used as the reference to order and orient *Q. canariensis* contigs into chromosome-scale scaffolds. Unplaced sequences were assigned to chr0.

**Structural and functional annotation**

Structural gene annotation was performed independently on both haplotypes of *Q. canariensis* using the Eugene annotation pipeline (Carrere et al., 2023; Sallet et al., 2019), based on transcript evidence from publicly available RNA-seq datasets and protein homology to curated plant proteomes, including *Quercus* reference proteomes. Gene prediction models were trained using Helixer. Prior to annotation, genome assemblies were soft-masked for repetitive elements using RED and LTRharvest, as implemented by default in the Eugene pipeline. Annotation outputs included protein-coding genes, rRNA, tRNA, and other non-coding RNA loci. Annotation completeness was assessed at the protein level using BUSCO v5.6.1 with

the *viridiplantae_odb10* dataset (n = 425). Psauron scores were 89.1 and 89.5 for haplotypes 1 and 2, respectively.

Functional annotation was performed using InterProScan v5.64-96.0 (Quevillon et al., 2005) to identify domains and motifs in the 51,882 and 46,482 predicted proteins from haplotypes 1 and 2, respectively, excluding chr0. The eggNOG-mapper V2.1.12 (Cantalapiedra et al., 2021) was used together with the eggNOG v5 genomes and functional databases of orthologs groups (OGs) (Huerta-Cepas et al., 2019). Finally, the Ensembl Enzyme Prediction Pipeline (E2P2) V3c8cbc2 identified the enzyme codes associated with the *Q. canariensis* proteins (https://github.com/carnegie/E2P2) (Chae et al., 2014).

**Transposable elements annotation**

A reference *de novo* transposable element (TE) consensus library was generated from the genome assembly of the admixed *Quercus petraea/pubescens* individual described above (https://doi.org/10.57745/PH1MMQ). This assembly was generated with Asm4pg using the same procedure as for the *Q. canariensis* specimen. The TE library was constructed using the TEdenovo pipeline (Flutre et al., 2011) from REPET package v3.0 (https://urgi.versailles.inra.fr/Tools/REPET). During TEdenovo's step 3, repetitive high-scoring segment pairs (HSPs) identified by BLAST were clustered using Grouper (Quesneville et al., 2003). TE consensus sequences were subsequently classified with PASTEC v2.0 (Hoede et al., 2014), based on the Repbase v30.01 (Jurka et al., 2005), Pfam v37.2 (Finn et al., 2014), GyDB v2.0 (Llorens et al., 2011), RexDBv4.0 (Neumann et al., 2019) and Eukaryota rRNA (GenBank) reference libraries. To reduce redundancy, the TEannot pipeline (Quesneville et al., 2005) was first applied following the procedure described by Jamilloux et al., 2016, using the TEdenovo-derived consensus library to retain only consensus sequences supported by at least one full-length copy (FLC) annotated in the genome. A second TEannot run was then performed using this filtered FLC consensus library to produce the final TE genome annotation. Transposable elements from the *Q. canariensis* genome haplotype 1 were then annotated using the TEannot pipeline (Quesneville et al., 2005) with this *Quercus petraea/pubescens*–derived TE consensus library.

## Data availability

### European Nucleotide Archive

Sequencing data and the genome assembly for *Q. canariensis* are available at the European Nucleotide Archive under project accession PRJEB110096. The sequenced individual is registered under BioSample accession ERS29488839 (sample alias: QcanPM). Raw PacBio HiFi sequencing reads are available under accession ERR16904215. The chromosome-scale genome assemblies for the two haplotypes are available under accession numbers ERZ29250309 (haplotype 1) and ERZ29250310 (haplotype 2).

Sequencing data for *Q. pubescens* are available at the European Nucleotide Archive under project accession PRJEB110096. The sequenced individual is registered under BioSample accession SAMEA122115212 (sample alias: QpubCR). Raw PacBio HiFi sequencing reads are available under accession ERR16980058.

### Annotation data

Structural, functional and transposable element annotations are available at doi: https://doi.org/10.57745/Q5D6UE

The TE reference library used for the TE annotation of *Q. canariensis is available at* https://doi.org/10.57745/PH1MMQ

These data are released prior to the full comparative genomic analyses as part of an ongoing European white oak pangenome project. In accordance with the Toronto Statement on prepublication data sharing, the authors reserve the right to publish the first analyses of the complete dataset.

### Code availability

All scripts, configuration files, and parameters used in this study are available at https://doi.org/10.57745/S0ND7F

## Acknowledgements

We acknowledge the GENTYANE platform (INRAE Clermont-Ferrand; DOI: 10.15454/1.5572409592543596E12) for NGS data production support and the Genotoul bioinformatics platform (Toulouse Midi-Pyrénées, Bioinfo Genotoul, DOI: 10.15454/1.5572369328961167E12) for providing computational resources. We thank Christophe Plomion and Olivier Lepais for careful reading and helpful comments on the manuscript, and René Graf for help in field and laboratory work. TE annotation was carried out using the URGI bioinformatics facility (DOI: 10.15454/1.5572414581735654E12) and was supported by the French government through the National Research Agency (ANR), as part of the France 2030 program for research infrastructure (EQUIPEX+), reference ANR-21-ESRE-0048 as well as Phenome-Emphasis (ANR-11-INBS-0012).

This work was funded by the EVOLTREE network through an Opportunity Call.

## Author contribution

Conceptualisation: M. Valbuena, E. Saez-Laguna, L. Duvaux.

Funding acquisition: C. Rellstab, M. Valbuena, E. Saez-Laguna, L. Duvaux.

Data curation and formal analysis: F. Couturier, C. Cravero, I. Lesur, J. Confais, W. Marande.

Investigation (data production): E. Belmonte.

Resources (sampling): C. Rellstab, M. Valbuena.

Supervision: L. Duvaux.

Writing – original draft, review and editing: All authors.

## References


Aitken, S. N., & Bemmels, J. B. (2016). Time to get moving : Assisted gene flow of forest trees. *Evolutionary Applications*, *9*(1), 271-290. https://doi.org/10.1111/eva.12293

Bratu, I., Dinca, L., Constandache, C., & Murariu, G. (2025). Resilience and Decline : The Impact of Climatic Variability on Temperate Oak Forests. *Climate*, *13*(6), 119. https://doi.org/10.3390/cli13060119

Cannon, C. H., & Petit, R. J. (2020). The oak syngameon : More than the sum of its parts. *New Phytologist*, *226*(4), 978-983. https://doi.org/10.1111/nph.16091

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2 : Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, *38*(12), 5825-5829. https://doi.org/10.1093/molbev/msab293

Carrere, S., Schiex, T., & Gouzy, J. (2023). *Eukaryote EuGene pipeline Version 2* (Version 2.0.2) [Logiciel]. Zenodo. https://doi.org/10.5281/zenodo.7515746

Chae, L., Kim, T., Nilo-Poyanco, R., & Rhee, S. Y. (2014). Genomic Signatures of Specialized Metabolism in Plants. *Science*, *344*(6183), 510-513. https://doi.org/10.1126/science.1252076

Chakraborty, D., Ciceu, A., Ballian, D., Benito Garzón, M., Bolte, A., Bozic, G., Buchacher, R., Čepl, J., Cremer, E., Ducousso, A., Gaviria, J., George, J. P., Hardtke, A., Ivankovic, M., Klisz, M., Kowalczyk, J., Kremer, A., Lstibůrek, M., Longauer, R., … Schueler, S. (2024). Assisted tree migration can preserve the European forest carbon sink under climate change. *Nature Climate Change*, *14*(8), 845-852. https://doi.org/10.1038/s41558-024-02080-5

Cheng, H., Asri, M., Lucas, J., Koren, S., & Li, H. (2024). Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nature Methods*, *21*(6), 967-970. https://doi.org/10.1038/s41592-024-02269-8

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170-175. https://doi.org/10.1038/s41592-020-01056-5

Cheng, H., Jarvis, E. D., Fedrigo, O., Koepfli, K.-P., Urban, L., Gemmell, N. J., & Li, H. (2022). Haplotype-resolved assembly of diploid genomes without parental data. *Nature Biotechnology*, *40*(9), 1332-1335. https://doi.org/10.1038/s41587-022-01261-x

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, *29*(1), 51-63. https://doi.org/10.1016/j.tree.2013.09.008

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2014). Pfam : The protein families database. *Nucleic Acids Research*, *42*(D1), D222-D230. https://doi.org/10.1093/nar/gkt1223

Flutre, T., Duprat, E., Feuillet, C., & Quesneville, H. (2011). Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLoS ONE*, *6*(1), e16526. https://doi.org/10.1371/journal.pone.0016526

Gremme, G., Steinbiss, S., & Kurtz, S. (2013). GenomeTools : A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *10*(03), 645-656. https://doi.org/10.1109/TCBB.2013.68

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896-2898. https://doi.org/10.1093/bioinformatics/btaa025

Hipp, A. L., Manos, P. S., Hahn, M., Avishai, M., Bodénès, C., Cavender-Bares, J., Crowl, A. A., Deng, M., Denk, T., Fitz-Gibbon, S., Gailing, O., González-Elizondo, M. S., González-Rodríguez, A., Grimm, G. W., Jiang, X., Kremer, A., Lesur, I., McVay, J. D., Plomion, C., … Valencia-Avalos, S. (2020). Genomic landscape of the global oak phylogeny. *New Phytologist*, *226*(4), 1198-1212. https://doi.org/10.1111/nph.16162

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., & Quesneville, H. (2014). PASTEC : An Automatic Transposable Element Classification Tool. *PLoS ONE*, *9*(5), e91929. https://doi.org/10.1371/journal.pone.0091929

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0 : A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309-D314. https://doi.org/10.1093/nar/gky1085

Jamilloux, V., Daron, J., Choulet, F., & Quesneville, H. (2016). De Novo Annotation of Transposable Elements : Tackling the Fat Genome Issue. *Proceedings of the IEEE*, 1-8. https://doi.org/10.1109/JPROC.2016.2590833

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, *110*(1-4), 462-467. https://doi.org/10.1159/000084979

Knutzen, F., Averbeck, P., Barrasso, C., Bouwer, L. M., Gardiner, B., Grünzweig, J. M., Hänel, S., Haustein, K., Johannessen, M. R., Kollet, S., Müller, M. M., Pietikäinen, J.-P., Pietras-Couffignal, K., Pinto, J. G., Rechid, D., Rousi, E., Russo, A., Suarez-Gutierrez, L., Veit, S., … Gliksman, D. (2025). Impacts on and damage to European forests from the 2018–2022 heat and drought events. *Natural Hazards and Earth System Sciences*, *25*(1), 77-117. https://doi.org/10.5194/nhess-25-77-2025

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

Lepais, O., Aissi, A., Véla, E., & Beghami, Y. (2022). Joint analysis of microsatellites and flanking sequences enlightens complex demographic history of interspecific gene flow and vicariance in rear-edge oak populations. *Heredity*, *129*(3), 169-182. https://doi.org/10.1038/s41437-022-00550-0

Leroy, T., Louvet, J., Lalanne, C., Le Provost, G., Labadie, K., Aury, J., Delzon, S., Plomion, C., & Kremer, A. (2020). Adaptive introgression as a driver of local adaptation to climate in European white oaks. *New Phytologist*, *226*(4), 1171-1182. https://doi.org/10.1111/nph.16095

Llorens, C., Futami, R., Covelli, L., Dominguez-Escriba, L., Viu, J. M., Tamarit, D., Aguilar-Rodriguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G. P., Maumus, F., Munoz-Pomer, A., Sempere, J. M.,

Latorre, A., & Moya, A. (2011). The Gypsy Database (GyDB) of mobile genetic elements : Release 2.0. *Nucleic Acids Research*, *39*(Database), D70-D74. https://doi.org/10.1093/nar/gkq1061

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update : Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647-4654. https://doi.org/10.1093/molbev/msab199

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k* -mers. *Bioinformatics*, *27*(6), 764-770. https://doi.org/10.1093/bioinformatics/btr011

Mikheenko, A., Saveliev, V., Hirsch, P., & Gurevich, A. (2023). WebQUAST : Online evaluation of genome assemblies. *Nucleic Acids Research*, *51*(W1), W601-W606. https://doi.org/10.1093/nar/gkad406

Neumann, P., Novák, P., Hoštáková, N., & Macas, J. (2019). Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA*, *10*(1), 1. https://doi.org/10.1186/s13100-018-0144-1

Ou, S., Chen, J., & Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky730

Ou, S., & Jiang, N. (2018). LTR_retriever : A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiology*, *176*(2), 1410-1422. https://doi.org/10.1104/pp.17.01310

Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., & Anxolabehere, D. (2005). Combined Evidence Annotation of Transposable Elements in Genome Sequences. *PLoS Computational Biology*, *1*(2), e22. https://doi.org/10.1371/journal.pcbi.0010022

Quesneville, H., Nouaud, D., & Anxolabehere, D. (2003). Detection of New Transposable Element Families in Drosophila melanogaster and Anopheles gambiae Genomes. *Journal of Molecular Evolution*, *57*(0), S50-S59. https://doi.org/10.1007/s00239-003-0007-2

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), 1432. https://doi.org/10.1038/s41467-020-14998-3

Rellstab, C., Dauphin, B., & Exposito-Alonso, M. (2021). Prospects and limitations of genomic offset in conservation management. *Evolutionary Applications*, *14*(5), 1202-1212. https://doi.org/10.1111/eva.13205

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury : Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, *21*(1), 245. https://doi.org/10.1186/s13059-020-02134-9

Sallet, E., Gouzy, J., & Schiex, T. (2019). EuGene : An Automated Integrative Gene Finder for Eukaryotes and Prokaryotes. In M. Kollmar (Éd.), *Gene Prediction : Methods and Protocols* (p. 97-120). Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_6

Sánchez-Salguero, R., Colangelo, M., Matías, L., Ripullone, F., & Camarero, J. J. (2020). Shifts in Growth Responses to Climate and Exceeded Drought-Vulnerability Thresholds Characterize Dieback in Two Mediterranean Deciduous Oaks. *Forests*, *11*(7), 714. https://doi.org/10.3390/f11070714

Vila-Viçosa, C., Arenas-Castro, S., Marcos, B., Honrado, J., García, C., Vázquez, F. M., Almeida, R., & Gonçalves, J. (2020). Combining Satellite Remote Sensing and Climate Data in Species Distribution Models to Improve the Conservation of Iberian White Oaks (Quercus L.). *ISPRS International Journal of Geo-Information*, *9*(12), 735. https://doi.org/10.3390/ijgi9120735

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope : Fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202-2204. https://doi.org/10.1093/bioinformatics/btx153

Xu, Z., & Wang, H. (2007). LTR_FINDER : An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, *35*(Web Server), W265-W268. https://doi.org/10.1093/nar/gkm286

Zimmermann, F., Reutimann, O., Baltensweiler, A., Walthert, L., Olofsson, J. K., & Rellstab, C. (2025). Fine-Scale Variation in Soil Properties Promotes Local Taxonomic Diversity of Hybridizing Oak Species ( *Quercus* spp.). *Evolutionary Applications*, *18*(2), e70076. https://doi.org/10.1111/eva.70076