

Haplotype-resolved PacBio HiFi genome assemblies of the olive cultivar 'Picual'

Martín Moret

Universidad de Jaén

Silvia Manrique

Institute of Conservation and Improvement of Valencian Agrobiodiversity (COMAV)- Polytechnic University of Valencia

Aureliano Bombarely

Consejo Superior de Investigaciones Científicas

Francisco Luque

fjluque@u.jaen.es

Universidad de Jaén

data-descriptor

Keywords:

Posted Date: October 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10127517/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

We present a chromosome level genome assembly for *Olea europaea* L. cultivar 'Picual', as well as the two corresponding haplotypes, obtained using PacBio HiFi sequencing. Four SMRT Cells were used for sequencing, resulting in a 35.5× sequencing depth. The consensus genome displayed a total assembly length of 1.24 Gbp, with an N50 value of 45.8 Mbp, an L50 of 11, and a largest scaffold of 91.4 Mbp. BUSCO (v5.8.2) analysis show a good conserved gene completeness (97.9%), with only 1.0% missing and 1.1% fragmented genes out of the 2,326 conserved eudicot genes analyzed. Regarding the haplotype assemblies, Haplotype 1 and Haplotype 2 had total scaffold lengths of 42.2 and 40.4 Mbp, L50 values of 12 and 11, and largest scaffolds of 86.3 and 80.6 Mbp, respectively. In the primary genome assembly, 35,281 protein-coding genes with 41,252 transcript isoforms were identified. Interestingly, although most genes were equally expressed across both haplotypes, 3% were differentially expressed between them.

Background & Summary

The olive tree (*Olea europaea* ssp. *europaea*) is the most widely cultivated and economically important species from the genus *Olea*¹. With 11.6 million hectares cultivated globally, olive groves constitute 25% of the total permanent worldwide cultivated surface². It has special relevance in the Mediterranean basin, its center of origin³ and where the majority of olive crops are concentrated. Spain leads global olive cultivation with 2.5 million ha, followed by Tunisia (1.75 million ha) and Italy (1.14 million ha)², while Portugal has a smaller area but higher productivity due to its commitment to high-density/super-high-density orchards. This makes the olive tree particularly important in the European Union (EU), which is the world's leading producer of olive oil, accounting for around 60% of global olive oil production, with Spain as the main producer⁴.

However, new varieties able to adapt to new ambient challenges and hedgerow cultivation are needed. Olive breeding faces substantial limitations that constrain genetic improvement. The prolonged juvenile period of 10–15 years, combined with long generation cycles⁵ and the need for multi-year phenotypic evaluations across environments⁶, renders traditional selection methods extremely slow and costly⁷. These limitations highlight the critical need for genomics-assisted breeding strategies, but the application of these approaches requires a deep understanding of a genome that presents considerable structural and functional complexity.

The complexity of the olive genome has deep evolutionary roots. Phylogenomic analyses reveal signatures of two or three ancient whole-genome duplication (WGD) events, including two paleopolyploidizations shared across the Oleaceae family, with evidence suggesting additional lineage-specific duplications before modern variety diversification^{8,9}. These ancestral WGD events generated extensive gene family expansions, with pairs of paralogous genes exhibiting patterns of subfunctionalization and neofunctionalization that have contributed to phenotypic diversification. Post-polyploidization genome restructuring, driven by chromosomal rearrangements, segmental duplications and deletions, and the dynamics of highly active transposable elements (TEs), has further shaped the current genomic architecture of the 2n = 46 chromosome complement¹⁰. This paleopolyploid heritage establishes an intrinsically complex genomic foundation characterized by extensive genic redundancy and structural variation.

This ancient complexity is actively maintained and amplified through distinctive reproductive mechanisms and domestication patterns. The olive tree has an unusual system of homomorphic sporophytic diallelic self-incompatibility^{11,12}, with current proposals suggesting a double successive screen model¹³. This self-incompatibility requires cross-pollination between individuals for reproduction, promoting recombination and preventing inbreeding depression. However, olive trees have the capacity to propagate vegetatively through cuttings. Consequently, many modern varieties consist of trees that have been clonally reproduced^{3,14}. This led to the introduction of a dual propagation strategy during domestication: while sexual reproduction generates novel genetic combinations for varietal innovation, vegetative propagation through cuttings allows indefinite clonal maintenance of superior genotypes. This strategy enabled olive breeders to obtain new varieties, effectively "freezing" favorable heterozygous combinations across centuries^{3,15}. This complex demographic history, combined with clonal propagation that preserves heterozygous genotypes, has resulted in cultivated olives exhibiting exceptionally high heterozygosity levels^{10,16,17}. Unlike annual crops that underwent inbreeding during domestication, olive varieties maintain extensive intra-individual genetic diversity accumulated over millennia of clonal persistence. This intrinsic complexity in the olive tree genome has profound implications for modern genomic studies, presenting substantial technical challenges for genome assembly and functional genomic studies.

Over the past decade, progressive advances have been made in the study of the olive genome through the assembly of genomes of different varieties. The first breakthrough was the sequencing and assembly of the variety 'Farga' (Oe6)¹⁸ that was later improved in a subsequent assembly (Oe9) using genetic map data to anchor scaffolds to chromosomes¹⁹. This was followed by the sequencing of a wild olive genome (*Olea europaea* var. *sylvestris*, Oe451) using Illumina short-read technology, although it may not be a real wild olive. This was the first reference assembly at the chromosome level, and marked an important milestone by providing insights into the general structure of the genome and evolutionary history of the olive tree⁸.

Previous studies enabled the assembly of a genome of paramount importance in olive tree genomics: the 'Picual' variety (Oleu0.6.1). In this case, Illumina paired-end and mate-pair libraries were integrated with PacBio RSII long-read sequencing, resulting in an improved assembly with superior completeness of repetitive elements compared to earlier short-read-based assemblies¹⁰. This assembly enabled the discovery that the activation of transposable elements represents an important factor in the genomic evolution of cultivated olive trees and contributed to phenotypic diversification during olive domestication.

The next major advance in olive tree genomics came with the third-generation sequencing technologies: Oxford Nanopore Technologies (ONT) and Pacific Biosciences High-Fidelity (PacBio HiFi). These technologies generate long reads that span repetitive regions and structural variations, improving contiguity and base-level accuracy. For example, ONT long reads were combined with Hi-C data in the assembly of the 'Arbequina' variety producing a chromosome-scale assembly with high quality, with pseudochromosomes that exhibit high contiguity²⁰.

On the other hand, PacBio HiFi can generate extremely accurate long reads (> 99.9% accuracy, typically ~ 10 kb, which is particularly useful in highly heterozygous and repetitive genomes. As a result, the best olive genome assemblies published to date are based on this technology, such as those published for the varieties 'Frantoio' and 'Leccino', which revealed extensive variety-specific structural variation and the dominant contribution of repetitive sequences (especially LTR retrotransposons and tandem satellite repeats) to genome size differences between varieties¹⁷. The current state-of-the-art in olive genome assemblies is represented by the telomere-to-telomere gap-free assembly of 'Leccino' variety, obtained combining PacBio HiFi, ONT ultra long reads and Hi-C⁹. This study found evidence of three ancestral WGD events shared between olive varieties and other Oleaceae, unlike the two events that were known until now.

The gradual improvement of olive genome assemblies has enabled significant advances that have transformed our understanding of the species' biology. However, despite their high quality, all published olive genome assemblies share a fundamental limitation: they represent consensus sequences that collapse heterozygous variants into single haplotypes, resulting in the loss of critical information about allelic variation and the true complexity of heterozygous regions. In highly heterozygous species such as the olive tree this represents a substantial loss of genomic information. For this reason, we applied a computational phasing approach to obtain haplotypes of the 'Picual' variety, in which the two allelic versions at heterozygous loci are separated.

Methods

Plant Material and DNA Extraction

Leaf tissue was collected from an *Olea europaea* var 'Picual' tree cultivated under natural conditions in the World Olive Germplasm Collection of the Andalusian Institute of Agricultural in Córdoba (Andalusia, southern Spain). High-molecular-weight (HMW) genomic DNA was extracted using the Nucleon PhytoPure Plant Extraction Kit (Cytiva), following the manufacturer's protocol and starting with an initial amount of approximately 500 mg per extraction. DNA was quantified using Qubit dsDNA HS Assay Kits and Qubit ADNbc BR Assay Kits (Thermo Fisher Scientific). DNA quality was assessed using Agilent Bioanalyzer (Agilent Technologies). HMW integrity was assessed by agarose gel electrophoresis and using the Agilent Femto Pulse system (HS Large Fragment 50Kbp method). In some cases, samples were filtered to ensure the integrity of the DNA and treated with RNase to remove contaminants. Only samples containing HMW DNA were selected for long-read sequencing.

PacBio HiFi Sequencing and consensus genome assembly

Sequencing was performed at the SCSIE-Universitat de València. SMRTbell libraries were prepared from purified HMW DNA using the SMRTbell Express Template Preparation Kit v3.0 (Pacific Biosciences). The libraries were sequenced on a PacBio Sequel IIe platform employing Circular Consensus Sequencing (CCS) algorithm, with a minimum predicted accuracy threshold of Q20. Read quality was evaluated using NanoPlot, and only high-quality CCS reads were selected for genome assembly. To ensure both accuracy and comprehensive genome coverage, the variety was sequenced across four SMRT Cells, yielding a total of 7,787,395 HiFi reads, achieving total sequencing depth (35.5x). The high quality of PacBio HiFi reads allowed to generate a de novo genome assembly for 'Picual' variety using Hifiasm (v0.19.8). This assembler generates a consensus genome from raw long reads. Subsequently, raw contigs were polished and scaffolded to chromosome-level using RagTag (v2.1.0) with the 'Arbequina' genome as reference²⁰, which provided optimal synteny for achieving chromosome-scale organization. The resulting 'Picual' assembly was designated OleuPic101. The newly generated OleuPic101 assembly represents a substantial improvement over the previous Oleu0.6.1 assembly across multiple quality metrics, establishing it as one of the highest-quality olive genome assemblies published to date, as shown in Table 1.

Table 1
Comparative metrics of genome assemblies for *Olea europaea* varieties.

Assembly	O_europaea_v1	OLEA9	Oleur0.6.1	Oe_Rao	ASM4816919v1	ASM4816504v1	Leccino_Lv	OleurPic101
Variety	Sylvestris	Farga	Picual	Arbequina	Frantoio	Leccino	Leccino	Picual
Reference	[18]	[29]	[20]	[30]	[27]	[27]	[19]	This work
Number of fragments	41,226	9,753	9,174	962	2,047	124	23	395
Total scaffold length (Gbp)	1.14	1.32	1.68	1.10	1.19	1.43	1.28	1.24
Largest scaffold (Mbp)	46.03	36.44	4.14	68.07	7.52	73.35	118.04	91.41
Smallest scaffold (bp)	271	500	1,017	10,772	5,595	15,279	33,890,297	10,145
GC content %	35.40	35.31	34.62	34.33	35.81	36.87	35.90	36.02
Scaffold N50	12,567,911	734,380	410,451	42,601,851	1,779,939	45,875,387	54,845,850	45,795,949
Scaffold N90	22,678	116,062	86,918	279,924	242,455	6,013,538	39,213,977	1,880,911
Scaffold L50	23	162	1,145	11	189	13	9	11
Scaffold L90	3,383	2,019	4,503	52	850	34	20	29

Our assembly spans 1.24 Gbp across 395 fragments with an N50 of 45.8 Mbp, comprising 23 ultra-long scaffolds corresponding with the full set of olive chromosomes, 15 non-assigned scaffolds and 357 contigs. This represents a 111-fold improvement over the previous Picual assembly (Oleur0.6.1; N50 = 410 kbp) and approaching the quality of the most contiguous olive genome to date (Leccino_Lv; N50 = 54.8 Mbp). The total length of the assembly is 1.24 Gbp, being very close to 1.26 Gb estimated by GenomeScope2 based on our PacBio HiFi reads (Supplementary Figure S1).

Genome completeness assessment using BUSCO (v5.8.2) against the eudicots_odb10 database confirmed the high quality of the OleurPic101 assembly, achieving 97.9% completeness with only 1.0% missing and 1.1% fragmented genes out of 2,326 conserved eudicot genes analyzed. The identification of 418 duplicated BUSCOs (18.0%) likely reflects the complex genomic architecture characteristic of olive varieties, including retained heterozygosity and segmental duplications, rather than assembly artifacts.

Repetitive element landscape.

Repetitive elements of the consensus genome were annotated using a de novo and homology-based approach. RepeatModeler2 (version 2.0.6)²¹ with LTR structural analysis enabled was used to generate a custom repetitive elements library. TESorter (v1.4.7)²² was used for sub-classification of transposable elements. The complete library was used to soft-mask the genome with RepeatMasker (v4.1.5) (<http://repeatmasker.org>). EarlGrey (v5.1.1)²³ was used to automatize the process given LTR-specific annotations. The repeat-masked genome was then used as input for gene prediction. Comprehensive annotation of repetitive elements using EarlGrey (v5.1.1) and TESorter (v1.4.7) revealed that repetitive elements constitute approximately 75% of the genome (927 Mbp of 1.24 Gbp), indicating an exceptionally repeat-rich architecture typical of complex plant genomes (Fig. 1A, Supplementary Table S1).

Among different repetitive element families, Long Terminal Repeat (LTR) retrotransposons dominate the repetitive landscape, accounting for 41% of the total genome (507 Mbp) and representing the most abundant class with 983,119 copies across 488 families. DNA transposons constituted 4.1% of the genome (50 Mbp), while Long Interspersed Nuclear Elements and Short Interspersed Nuclear Elements comprised 0.86% and 0.02% respectively. Additionally, simple repetitions and microsatellites represented 11.6% of the genome (144 Mbp), contributing significantly to the overall repetitive content.

Evolutionary analysis using Kimura 2-parameter distance revealed distinct temporal patterns of LTR retrotransposon activity (Fig. 1B). The distribution shows two prominent peaks of LTR expansion at 0.04 and 0.16, suggesting at least two major bursts of retrotransposon activity during olive genome evolution. Notably, the first LTR burst is preceded by a peak of unclassified elements, which may be related to this event. These bursts could be associated with the two whole genome duplication events in the olive tree and may explain the high activity level of retrotransposon elements in olive genome.

Structural annotation and gene content

Structural annotation was performed using an ab initio and evidence-based approach, using a total of four published RNAseqs (BioProject: PRJNA989401; PRJNA870905; PRJNA1124876; PRJNA256033) from different olive tree tissues and conditions, used as evidence to train gene model. RNAseq reads were aligned to the genome using STAR (v2.7.10b)²⁴. Aligned reads were assembled into transcripts using StringTie (v2.2.1)²⁵, and open reading frames (ORFs) were predicted with TransDecoder (v5.5.0) (<https://github.com/TransDecoder/TransDecoder>). These data, along with plant protein sequences from UniProt and Swiss-Prot databases, were provided as input to BRAKER3 (v3.0.8) pipeline to train gene models with GeneMark-ET (v4.72)²⁶ and predict genes with Augustus (v3.5.0)²⁷. The resulting gene annotations were refined to include untranslated regions (UTRs) using InGenAnnot²⁸, which extends gene models based on RNA-seq alignment patterns.

Predicted protein sequences from the 'Picual' genome annotation were functionally characterized by homology search against the Swiss-Prot, TrEMBL and TAIR11²⁹ databases using DIAMOND (v0.9.24)³⁰. Homology-based functional descriptions were assigned using AHRD (v3.3.3) (<https://github.com/groupschoof/AHRD>), which integrates and prioritizes annotations from curated databases while resolving conflicts among multiple hits. The resulting functional annotations were incorporated into the genome GFF file using AGAT (v1.5.1) (<https://github.com/NBISweden/AGAT>), which also filtered out inconsistent gene models. The consensus genome annotation was then transferred to the haplotype-resolved assemblies using Liftoff (v1.5.1)³¹, ensuring accurate mapping of gene structures and functional annotations across both haplotypes. The final annotation includes all gene predictions with their known functions or both the consensus and haplotype-specific genomes.

Comprehensive structural annotation identified 35,281 protein-coding genes distributed across the 1.24 Gbp OleurPic101 genome, producing 41,252 transcript isoforms. The detection of 1.17 isoforms per gene demonstrates successful identification of alternative splicing events, indicating an increased transcript-level annotation that was not captured in earlier assemblies.

Predicted genes exhibited detailed structural architecture with an average gene length of 4,864 bp containing 5.6 exons per transcript and 302 bp average exon length. These parameters indicate comprehensive detection of complex gene structures, including extended UTRs and alternative splicing variants. The mean CDS length of 1,235 bp represents a substantial improvement over the previous 'Picual' annotation (1,005 bp), reflecting more complete gene models and reduced gene fragmentation.

The haplotype genomes assemblies

Conventional consensus genome assemblies collapse haplotype diversity, presenting significant limitations for highly heterozygous species like olive. The substantial heterozygosity observed in olive genomes results in extensive allelic differences that consensus assembly algorithms cannot adequately represent, leading to loss of biologically relevant genomic variation and potential misassembly of complex regions.

The 'Picual' haploid and haplotype-resolved genomes were *de novo* assembled from PacBio HiFi reads (44 Gbp; mean length: 5.65 Kbp; mean quality: 31,1; all read > Q20). The genome size was inferred with GenomeScope2 (v2.0)³². Hifiasm (v0.19.8)³³ was used as the *de novo* assembler with default parameters to obtain haplotype-resolver genome. Assembly quality was initially assessed using GFASStats (v1.3.11)³⁴ to calculate basic assembly statistic. BUSCO (v5.8.2)³⁵ was used to assess gene space completeness, while K-mers completeness was evaluated with Meryl (v2.2) and Merqury (v1.1)³⁶. Additionally Bandage (v0.8.1)³⁷ was used for graphical representation.

To identify and remove contaminant contigs, Minimap2 (v2.26)³⁸ was used to align assembly with HiFi reads and Diamond (v0.9.24)³⁹ to search homolog sequences in UniProt and Swiss-Prot databases for eudicots. Both results were used to perform a contaminant analysis with Blobtools (v1.1.1)⁴⁰. Contaminant contigs were removed manually. Purge_dups (v1.2.6) (https://github.com/dfguan/purge_dups) was used with default parameters to remove duplications. The polished and contamination-free assembled genome was scaffolded using RagTag (v2.1.0)⁴¹.

This approach yielded two haplotype assemblies with contrasting characteristics: Haplotype 1 (H1) comprised 2,745 scaffolds spanning 1.25 Gbp (N50 = 42.17 Mb), while Haplotype 2 (H2) contained 1,307 scaffolds covering 1.20 Gbp (N50 = 40.4 Mbp) (Table 2). Although these assemblies exhibit higher fragmentation compared to the consensus assembly due to reduce per-haplotype coverage, they preserve authentic haplotype-specific sequences that are lost during consensus generation.

Table 2
General statistics of haplotype 1 and haplotype 2 assembled with Hifiasm.

Haplotype	Haplotype 1	Haplotype 2
Number of scaffolds	2,745	1,307
Total contig length (Gbp)	1.25	1.20
Average scaffold length (Kbp)	456.31	917.12
Scaffold N50 (Mbp)	42.17	40.4
Scaffold L50	12	11
Largest scaffold (Mbp)	86.28	80.56
GC content %	36,31	36,18

K-mer spectrum analysis demonstrates the superior sequence representation achieved through haplotype resolution compared to consensus assembly (Fig. 2). While consensus assembly significant k-mer loss in heterozygous regions with a completeness of 75% (Fig. 2A), haplotype-resolved assembly captures nearly complete k-mer diversity from both parental contributions, showing a combined completeness of 99% (Fig. 3B), validating the preservation of authentic genomic variation. This analysis reveals that consensus assembly fails to represent the total genomic diversity present in this heterozygous variety.

Transcriptomic analysis

To analyze the gene expression in the two assembled haplotypes, first we transferred gene annotations from the consensus genome assembly to each haplotype. Annotation transference identified 35,473 genes in H1 and 34,513 genes in H2, with 33,701 genes shared between both haplotypes. The remaining genes were haplotype-specific: 1,772 genes unique to H1 and 812 genes unique to H2, representing structural variation between the two haploid assemblies according annotation transferring. To perform the transcriptomic analysis, a transcriptome including the H1 and H2 allelic versions of each gene was constructed.

To evaluate haplotype-specific gene expression patterns, we analyzed RNA-seq data from multiple 'Picual' organs and a tissue, mapping reads independently to each haplotype⁴². Although this approach could be limited to distinguish alleles with identical haplotypes by masking haplotypic differences, the high heterozygosity of the olive tree reduces this limitation. In fact, the results obtained show that this approach was successful.

Differential allelic gene expression across haplotypes

The objective of this analysis is to determine whether differences in allelic expression exist between the two haplotypes and to characterize this phenomenon. A prerequisite for this comparison is that gene models are accurately represented in both haplotype assemblies, so that differences in read counts genuinely reflect allele-specific expression. To ensure this, we evaluated the quality of the Liftoff-transferred annotations for all 33,701 shared genes. This evaluation identified 4,702 problematic genes for at least one haplotype: 3,724 lacked a valid ORF, 880 showed strand inversions consistent with assignment to paralogous rather than orthologous loci, and 98 showed a discrepancy more than three times greater in the size of the genetic model. In these cases, Liftoff mapped the gene to a probable non-orthologous genomic region, producing truncated models against which RNA-seq reads cannot align, generating artefactual expression differences unrelated to true allelic regulation. After excluding these cases, 28,999 reliable genes remained, to which a subsequent low expression filter was applied eliminating 1,585 genes. This resulted in 27,414 genes (81.3% of the shared genes) that were included in the differential expression analysis. Principal component analysis (PCA) of normalized expression values showed clear clustering, with samples from each organ/tissue and haplotype grouping together (Fig. 3), indicating consistent expression patterns across corresponding organs/tissues and haplotypes.

For the transcriptomic analysis, Q30-quality reads were used. This, together with the high level of variability among haplotype sequences, allowed an accurate expression analysis of the two haplotype gene copies. On average, the cDNA sequences of the analyzed genes contained 10.3 polymorphic sites per Kb, a value 10-fold higher than the expected sequencing error rate. Differential expression analysis between haplotypes identified 832 genes showing significant expression differences (Fig. 4) with FDR < 0.05 genome-wide (Fig. 5A), representing 3% of

analyzed shared genes. Of these genes, 606 showed moderate expression differences ($|\log_2\text{FC}| \leq 2$) and 170 showed high expression differences ($|\log_2\text{FC}| > 2$), suggesting substantial functional divergence between haplotypes at the transcriptional level (Supplementary Table S1).

Notably, a subset of shared genes exhibited highly asymmetric expression patterns between haplotypes. Applying stringent criteria (mean normalized expression > 10 in one haplotype and < 1 in the other, with $\text{FDR} < 0.05$), we identified 29 genes expressed exclusively in H1 and 27 genes expressed exclusively in H2 (Fig. 5B), despite their presence in both genomic assemblies. These haplotype-exclusive expression patterns indicate strong allele-specific transcriptional regulation, where genes present in both haplotypes are functionally silenced in one while actively transcribed in the other. The statistical robustness of these findings is supported by extremely low adjusted p-values, demonstrating that these expression patterns are not technical artifacts.

The 56 genes with haplotype-exclusive expression (29 in H1 and 27 in H2) showed variable patterns across organs/tissues. Expression levels among H1-exclusive genes ranged from ~ 10.18 to 1,626.10 DESeq2 size-factor-corrected counts, while H2-exclusive genes exhibited expression ranges from ~ 10.52 to 3,742.05 (Supplementary Table S3).

Data Records

The assembled genomes are available from the NCBI Bioproject PRJNA1389873.

Technical Validation

The quality and concentration of extracted DNA were assessed using Fragment Analyzer before the genome sequencing and samples of at least 40 Kbp were used for PacBio HiFi sequencing.

After the genome assembly was completed, the completeness of the genome assemblies was evaluated using the BUSCOs eudicotyledons_odb10 database.

Code availability

No custom code was used to generate or process the data described in the manuscript.

Declarations

Author Contributions

M.M: Methodology, Investigation, Writing—original draft. S.M: Investigation, Writing—review and editing. A.B: Conceptualization, Methodology, Writing—review and editing. F.L: Conceptualization, Writing—original draft, Supervision, Funding acquisition.

Funding

This research was funded by the Spanish Ministry of Science and Innovation (MCIN), grant number PID2020 - 115853RR-C33 and Martin Moret was granted by MCIN "Formación de Profesorado Universitario".

Acknowledgments

The technical and human support provided by CICT of the Universidad de Jaén (UJA, MINECO, Junta de Andalucía, FEDER) is gratefully acknowledged.

Competing interests

The authors declare no competing interests.

References

1. Green, P. S. A Revision of *Olea* L. (Oleaceae). *Kew Bulletin* 57, 91 (2002). <https://doi.org/10.2307/4110824>.
2. Fernández-Lobato, L., Ruiz-Carrasco, B., Tostado-Véliz, M., Jurado, F. & Vera, D. Environmental impact of the most representative Spanish olive oil farming systems: A life cycle assessment study. *Journal of Cleaner Production* 442, 141169 (2024). <https://doi.org/10.1016/j.jclepro.2024.141169>

3. Kaniewski, D. *et al.* Primary domestication and early uses of the emblematic olive tree: palaeobotanical, historical and molecular evidence from the Middle East. *Biological Reviews of the Cambridge Philosophical Society* 87, 885–899 (2012). <https://doi.org/10.1111/j.1469-185x.2012.00229.x>
4. International Olive Oil Council (2025, 12 septiembre) Statistics - International Olive Council. International Olive Council. <https://www.internationaloliveoil.org/what-we-do/statistics/>
5. Wechsler, T. *et al.* Determining Reproductive Parameters, which Contribute to Variation in Yield of Olive Trees from Different Cultivars, Irrigation Regimes, Age and Location. *Plants* 11, 2414 (2022). <https://doi.org/10.3390/plants11182414>
6. Yılmaz-Düzyaman, H., De la Rosa, R. & León, L. Seedling Selection in Olive Breeding Progenies. *Plants* 11, 1195 (2022). <https://doi.org/10.3390/plants11091195>
7. Valverde, P., Diez, C. M., Deger, R. E., Barranco, D. & Trapero, C. Genotypic influence in the juvenile to adult transition in olive seedlings. *Frontiers in Plant Science* 15, 1343589 (2024). <https://doi.org/10.3389/fpls.2024.1343589>
8. Unver, T. *et al.* Genome of wild olive and the evolution of oil biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 114, E9413-E9422 (2017). <https://doi.org/10.1073/pnas.1708621114>
9. Lv, J. *et al.* The gapless genome assembly and multi-omics analyses unveil a pivotal regulatory mechanism of oil biosynthesis in the olive tree. *Horticulture Research* 11, uhae168 (2024). <https://doi.org/10.1093/hr/uhae168>
10. Jiménez-Ruiz, J. *et al.* Transposon activation is a major driver in the genome evolution of cultivated olive trees (*Olea europaea* L.). *Plant Genome* 13, e20010 (2020). <https://doi.org/10.1002/tpg2.20010>
11. Alagna, F. *et al.* The Paradox of Self-Fertile Varieties in the Context of Self-Incompatible Genotypes in Olive. *Frontiers in Plant Science* 10, 725 (2019). <https://doi.org/10.3389/fpls.2019.00725>
12. Mariotti, R. *et al.* Diallelic self-incompatibility is the main determinant of fertilization patterns in olive orchards. *Evolutionary Applications* 14, 983–995 (2020). <https://doi.org/10.1111/eva.13175>
13. Breton, C. M. *et al.* A Dual-Successive-Screen Model at Pollen/Stigma and Pollen Tube/Ovary Explaining Paradoxical Self-Incompatibility Diagnosis in the Olive Tree—An Interpretative Update of the Literature. *Plants* 10, 1938 (2021). <https://doi.org/10.3390/plants10091938>
14. Montemurro, C., Dambruoso, G., Bottalico, G. & Sabetta, W. Self-Incompatibility Assessment of Some Italian Olive Genotypes (*Olea europaea* L.) and Cross-Derived Seedling Selection by SSR Markers on Seed Endosperms. *Frontiers in Plant Science* 10, 451 (2019). <https://doi.org/10.3389/fpls.2019.00451>
15. Besnard, G. *et al.* The complex history of the olive tree: from Late Quaternary diversification of Mediterranean lineages to primary domestication in the northern Levant. *Proceedings of the Royal Society B: Biological Sciences* 280, 20122833 (2013). <https://doi.org/10.1098/rspb.2012.2833>
16. Friel, J., Bombarely, A., Fornell, C. D., Luque, F. & Fernández-Ocaña, A. M. Comparative Analysis of Genotyping by Sequencing and Whole-Genome Sequencing Methods in Diversity Studies of *Olea europaea* L. *Plants* 10, 2514 (2021). <https://doi.org/10.3390/plants10112514>
17. Sebastiani, L. *et al.* PacBio genome assembly of *Olea europaea* L. *subsp. europaea* cultivars Frantoio and Leccino. *Scientific Data* 12, 1095 (2025). <https://doi.org/10.1038/s41597-025-05363-4>
18. Cruz, F. *et al.* Genome sequence of the olive tree, *Olea europaea*. *GigaScience* 5, 29 (2016). <https://doi.org/10.1186/s13742-016-0134-5>
19. Julca, I. *et al.* Genomic evidence for recurrent genetic admixture during the domestication of Mediterranean olive trees (*Olea europaea* L.) *BMC Biology* 18, 148 (2020). <https://doi.org/10.1186/s12915-020-00881-6>
20. Rao, G. *et al.* De novo assembly of a new *Olea europaea* genome accession using nanopore sequencing. *Horticulture Research* 8, 64 (2021). <https://doi.org/10.1038/s41438-021-00498-y>
21. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America* 117, 9451–9457 (2020). <https://doi.org/10.1073/pnas.1921046117>
22. Zhang, R. *et al.* TESorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. *Horticulture Research* 9, uhac017 (2022) <https://doi.org/10.1093/hr/uhac017>
23. Baril, T., Galbraith, J. & Hayward, A. Earl Grey: A Fully Automated User-Friendly Transposable Element Annotation and Analysis Pipeline. *Molecular Biology and Evolution* 41, msae068 (2024). <https://doi.org/10.1093/molbev/msae068>
24. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2012). <https://doi.org/10.1093/bioinformatics/bts635>
25. Shumate, A., Wong, B., Pertea, G. & Pertea, M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLOS Computational Biology* 18, e1009730 (2022). <https://doi.org/10.1371/journal.pcbi.1009730>
26. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research* 42, e119 (2014). <https://doi.org/10.1093/nar/gku557>
27. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24, 637–644 (2008). <https://doi.org/10.1093/bioinformatics/btn013>

28. Lapalu, N. *et al.* Improved Gene Annotation of the Fungal Wheat Pathogen *Zymoseptoria tritici* Based on Combined Iso-Seq and RNA-Seq Evidence. *Molecular Plant-Microbe Interactions* 38, 834–847 (2025). <https://doi.org/10.1094/MPMI-07-25-0077-TA>
29. Cheng, C. *et al.* Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *The Plant Journal* 89, 789–804 (2016). <https://doi.org/10.1111/tpj.13415>
30. Buchfink, B., Reuter, K. & Drost, H. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* 18, 366–368 (2021). <https://doi.org/10.1038/s41592-021-01101-x>
31. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* 37, 1639–1643 (2020). <https://doi.org/10.1093/bioinformatics/btaa1016>
32. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11, 1432 (2020). <https://doi.org/10.1038/s41467-020-14998-3>
33. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* 8, 170–175 (2021). <https://doi.org/10.1038/s41592-020-01056-5>
34. Formenti, G. *et al.* Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics* 38, 4214–4216 (2022). <https://doi.org/10.1093/bioinformatics/btac460>
35. Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution* 35, 543–548 (2017). <https://doi.org/10.1093/molbev/msx319>
36. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* 21, 245 (2020). <https://doi.org/10.1186/s13059-020-02134-9>
37. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 31, 3350–3352 (2015). <https://doi.org/10.1093/bioinformatics/btv383>
38. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100 (2018). <https://doi.org/10.1093/bioinformatics/bty191>
39. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12, 59–60 (2014). <https://doi.org/10.1038/nmeth.3176>
40. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. *F1000Research* 6, 1287 (2017). <https://doi.org/10.12688/f1000research.12232.1>
41. Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology* 23, 258 (2022). <https://doi.org/10.1186/s13059-022-02823-7>
42. Ramírez-Tejero, J. A., Jiménez-Ruiz, J., Leyva-Pérez, M. de la O., Barroso, J. B. & Luque, F. Gene expression pattern in olive tree organs (*Olea europaea* L.). *Genes* 11, 544 (2020). <https://doi.org/10.3390/genes11050544>

Figures

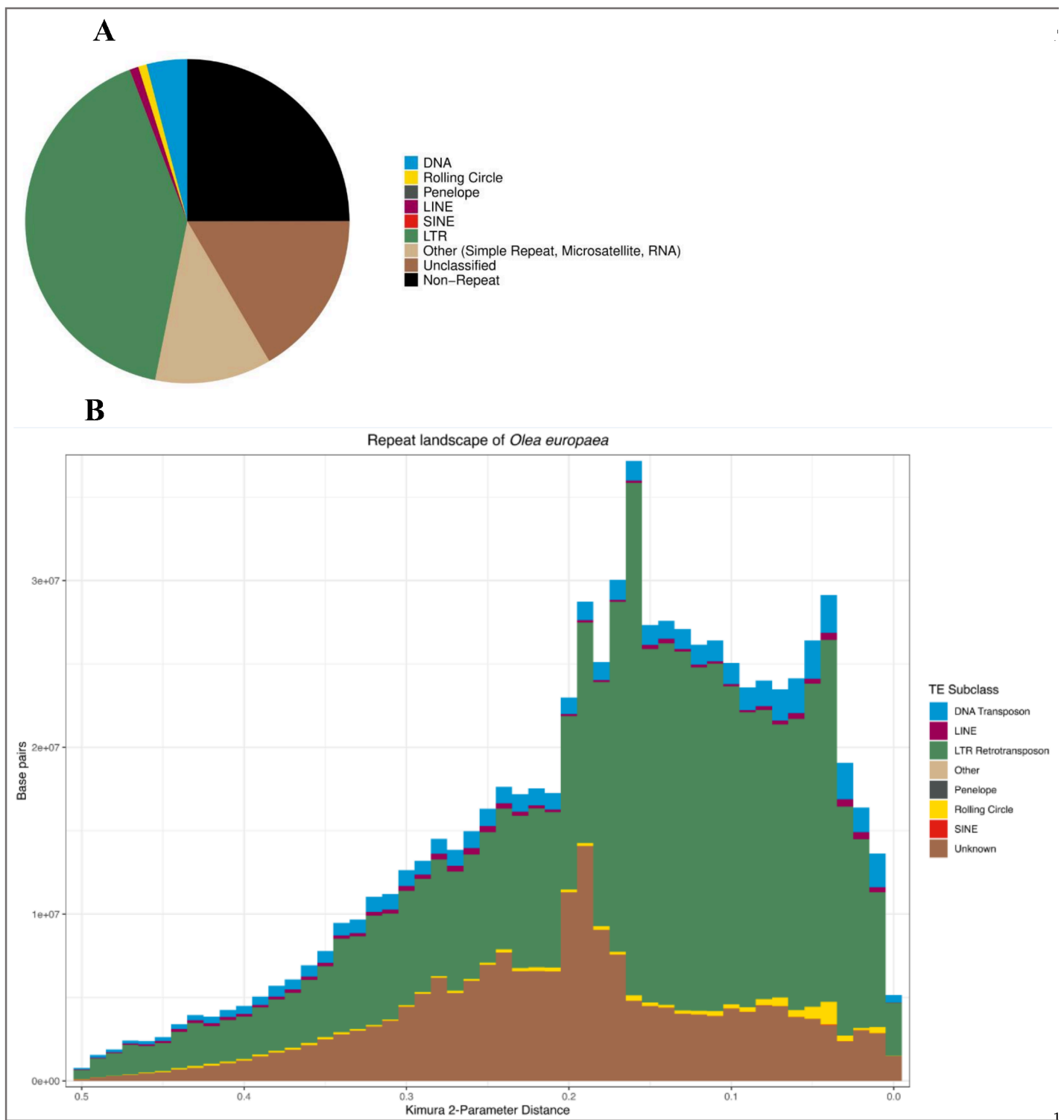


Figure 1

Distribution of different families of repetitive elements in *Olea europaea* subsp. *europaea* var. 'Picual' (A). Distribution of repetitive elements over time based on Kimura 2-Parameter Distance (B).

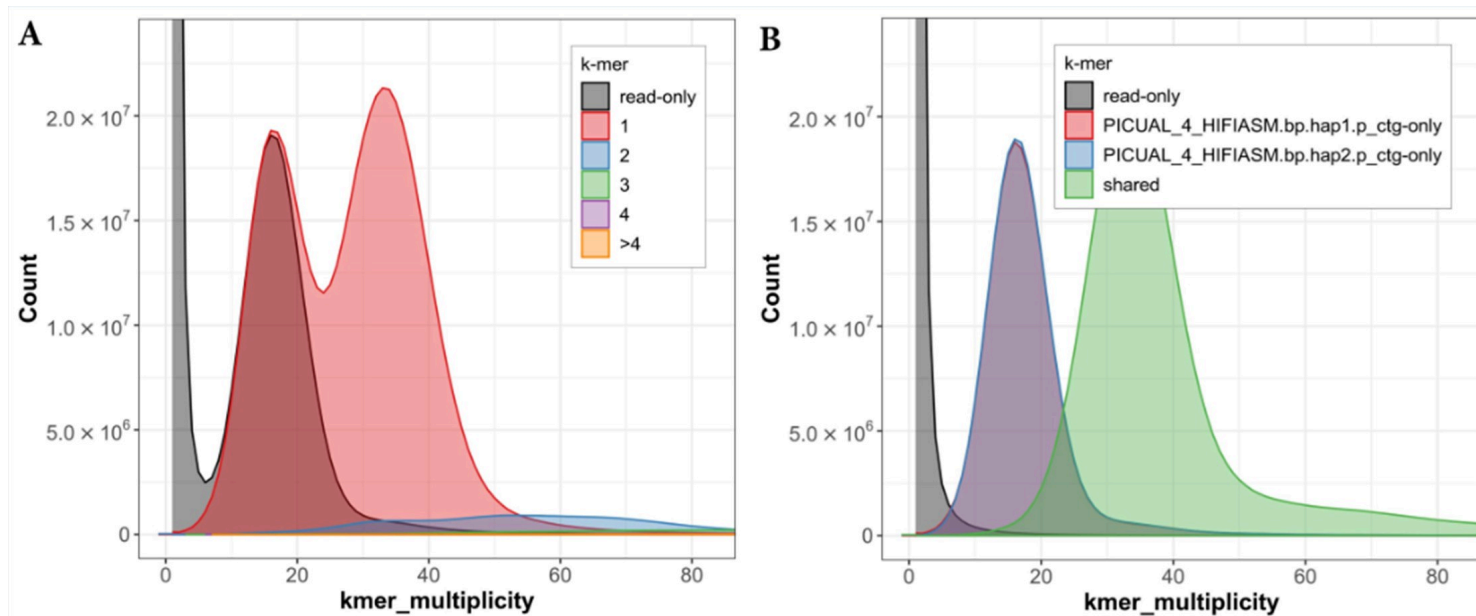


Figure 2

Representation of Mercury K-mers multiplicity respect counts. (A) K-mers multiplicity taking into account the contigs for consensus genome. (B) K-mers multiplicity taking into account the contigs for both assembled haplotypes.

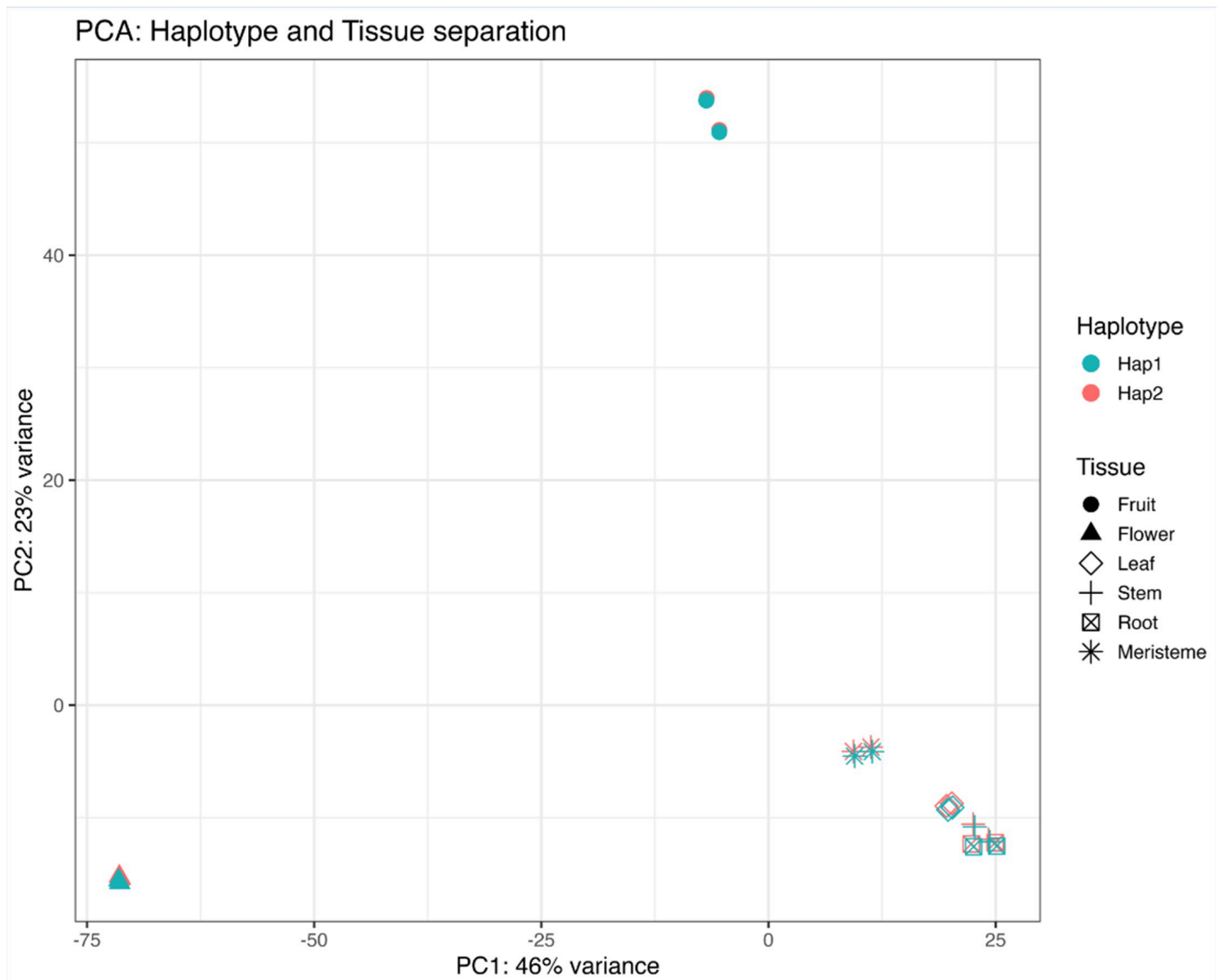


Figure 3

Principal component analysis (PCA) of variance-stabilized gene expression data from 24 samples across 6 tissue types (two biological replicates each). Points are colored by haplotype (H1: red; H2: blue) and shaped by tissue (F: fruit, Flower: flower, H: leaf, M: meristem, R: root, T: stem). Analysis includes 20,654 shared genes (≥ 10 reads).

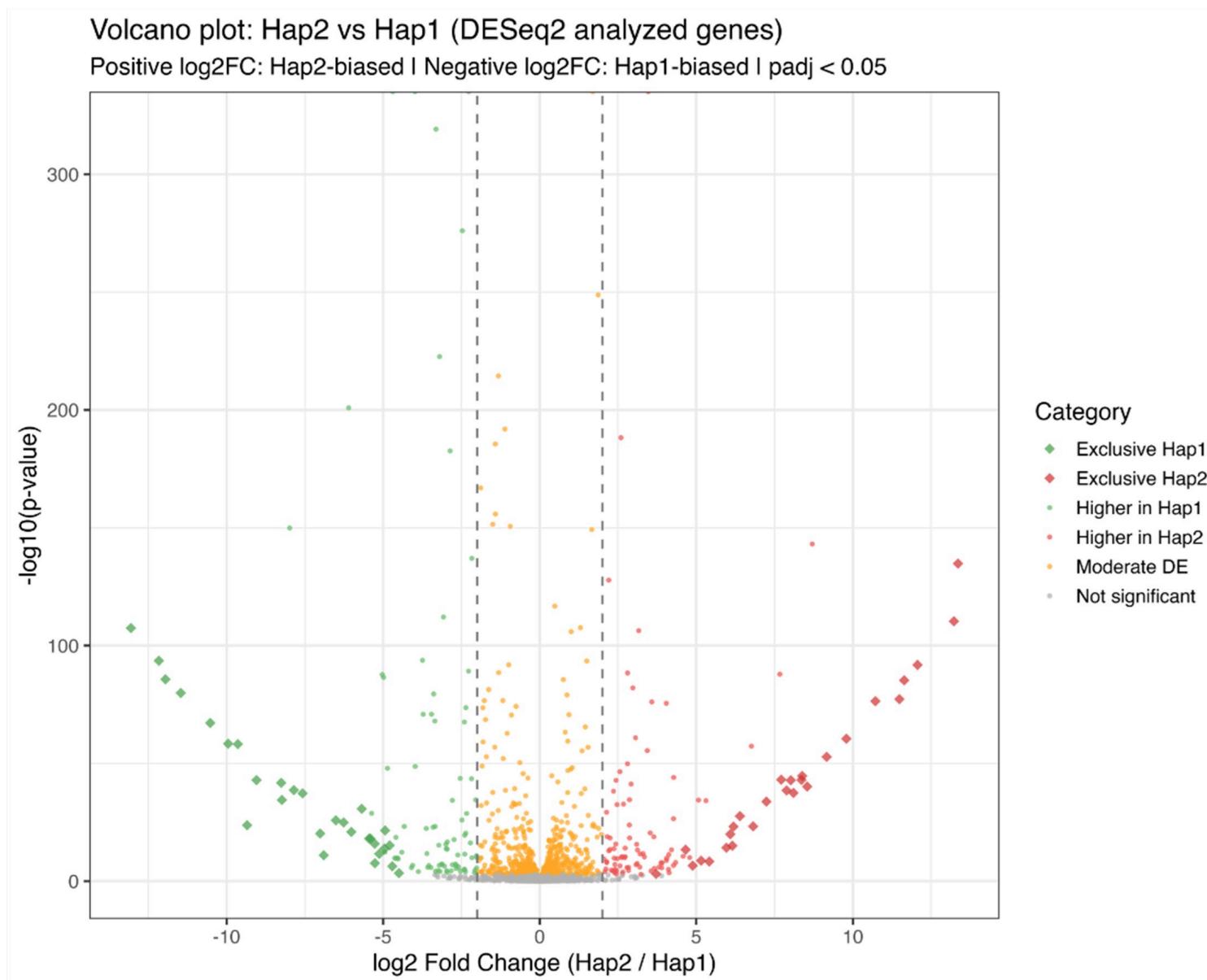


Figure 4

Volcano plot of general results of DESeq2 analysis of RNAseq data from different haplotypes in different tissues.

Chromosomal mapping of differentially expressed genes across haplotypes



Figure 5

Haplotype-chromosomal location of differentially expressed genes (A) and Haplotype-exclusive expressed genes (B).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterials.docx](#)