Research

New Phytologist

In a nutshell: pistachio genome and kernel development

Jaclyn A. Adaskaveg[1]*, Chaehee Lee[1]*, Yiduo Wei[1], Fangyi Wang[1], Filipa S. Grilo[2,3], Saskia D. Mesquida-Pesci[1], Matthew Davis[1], Selina C. Wang[3], Giulia Marino[1], Louise Ferguson[1], Patrick J. Brown[1], Georgia Drakakaki[1], Adela Mena Morales[4], Annalisa Marchese[5], Antonio Giovino[6], Esaú Martínez Burgos[4], Francesco Paolo Marra[5], Lourdes Marchante Cuevas[4], Luigi Cattivelli[7], Paolo Bagnaresi[7], Pablo Carbonell-Bejerano[8], J. Grey Monroe[1] and Barbara Blanco-Ulate[1]

[1]Department of Plant Sciences, University of California, Davis, CA 95616, USA; [2]Corto Olive, Lodi, CA 95212, USA; [3]Department of Food Science and Technology, University of California Davis, Davis, CA 95616, USA; [4]Regional Institute of Agri-Food and Forestry Research and Development of Castilla-La Mancha (IRIAF), IVICAM, CTRA, Toledo-Albacete s/n, 13700, Tomelloso (Ciudad Real), 13700, Spain; [5]Department of Agricultural, Food and Forest Sciences, University of Palermo, Viale delle Scienze – Ed. 4, Palermo, 90128, Italy; [6]CREA for Agricultural Research and Economics (CREA), Research Centre for Plant Protection and Certification (CREA-DC), Viale delle Scienze, Palermo, 90128, Italy; [7]CREA Research Centre for Genomics and Bioinformatics, Fiorenzuola d'Arda, 29017, Italy; [8]Instituto de Ciencias de la Vid y del Vino, ICVV, for Grape and Wine Sciences ICVV, CSIC – Universidad de La Rioja – Gobierno de La Rioja, Logroño, 26007, Spain

Authors for correspondence:
Barbara Blanco-Ulate
Email: bblanco@ucdavis.edu

J. Grey Monroe
Email: gmonroe@ucdavis.edu

Received: *18 September 2024*
Accepted: *19 February 2025*

New Phytologist (2025) **246:** 1032–1048
doi: 10.1111/nph.70060

Key words: chromosome-scale assembly, Kerman, kernel metabolic profile, nut development, nut physiology, pistachio, *Pistacia vera*, reference genome, spatiotemporal transcriptome, tree crop.

Summary

- Pistachio is a sustainable nut crop with exceptional climate resilience and nutritional value. However, the molecular processes underlying pistachio nut development and nutritional traits are largely unknown, compounded by limited genomic and molecular resources.
- To advance pistachios as a future food source and a model system for hard-shelled fruits, we generated a chromosome-scale reference genome of the most widely grown pistachio cultivar (*Pistacia vera* 'Kerman') and a spatiotemporal study of nut development. We integrated tissue-level physiological data from thousands of nuts over three growing seasons with transcriptomic data encompassing 14 developmental time points of the hull, shell, and kernel to assemble gene modules associated with physiological changes.
- Our study defined four distinct stages of pistachio nut growth and maturation. We then focused on the kernel to identify transcriptional and metabolic changes in molecular pathways governing nutritional quality, such as the accumulation of unsaturated fatty acids, which are vital for shelf life and dietary value. These findings revealed key candidate conserved regulatory genes, such as *PvAP2-WRI1* and *PvNFYB-LEC1*, likely involved in oil accumulation in kernels.
- This work yields new knowledge and resources that will inform other woody crops and facilitate further improvement of pistachio as a globally significant, sustainable, and nutritious crop.

Introduction

Tree nuts are the most carbon-efficient protein source of any food (Poore & Nemecek, 2018). Pistachios are also rich in unsaturated fatty acids, antioxidants, and vitamins (Tsantili *et al.*, 2011; Marvinney *et al.*, 2014; Noguera-Artiaga *et al.*, 2019; Polari *et al.*, 2019; Mandalari *et al.*, 2021; Derbyshire *et al.*, 2023). Given that pistachio trees are highly resilient to abiotic stress, particularly drought and salinity, they are projected to be an important source of sustainable nutrition in the face of climate change over the next century (Moazzam Jazi *et al.*, 2016), with global production of pistachios having more than doubled over the past two decades (Food and Agricultural Organization; https://www.fao.org/faostat/en/#search/pistachio; Fig. 1a).

Pistachio (*Pistacia vera*, $2n = 30$) belongs to the Anacardiaceae family, along with cashew and mango, and is the only species in the genus *Pistacia* grown for its edible fruit. Pistachio trees are dioecious and wind-pollinated. Although commonly known as nuts, pistachio fruits are botanically dehiscent drupes consisting of three main tissues: a leathery exo-mesocarp (hull), a stony endocarp (shell), and an edible seed (kernel; Fig. 1b,c). Pistachio nut growth has been previously divided into three stages: (1) growth of the hull and shell; (2) shell lignification; and (3) kernel growth (Lin *et al.*, 1984; Polito & Pinney, 1999; Goldhamer & Beede, 2004; Ferguson *et al.*, 2005; Zhang *et al.*, 2021). These

*These authors contributed equally to this work.

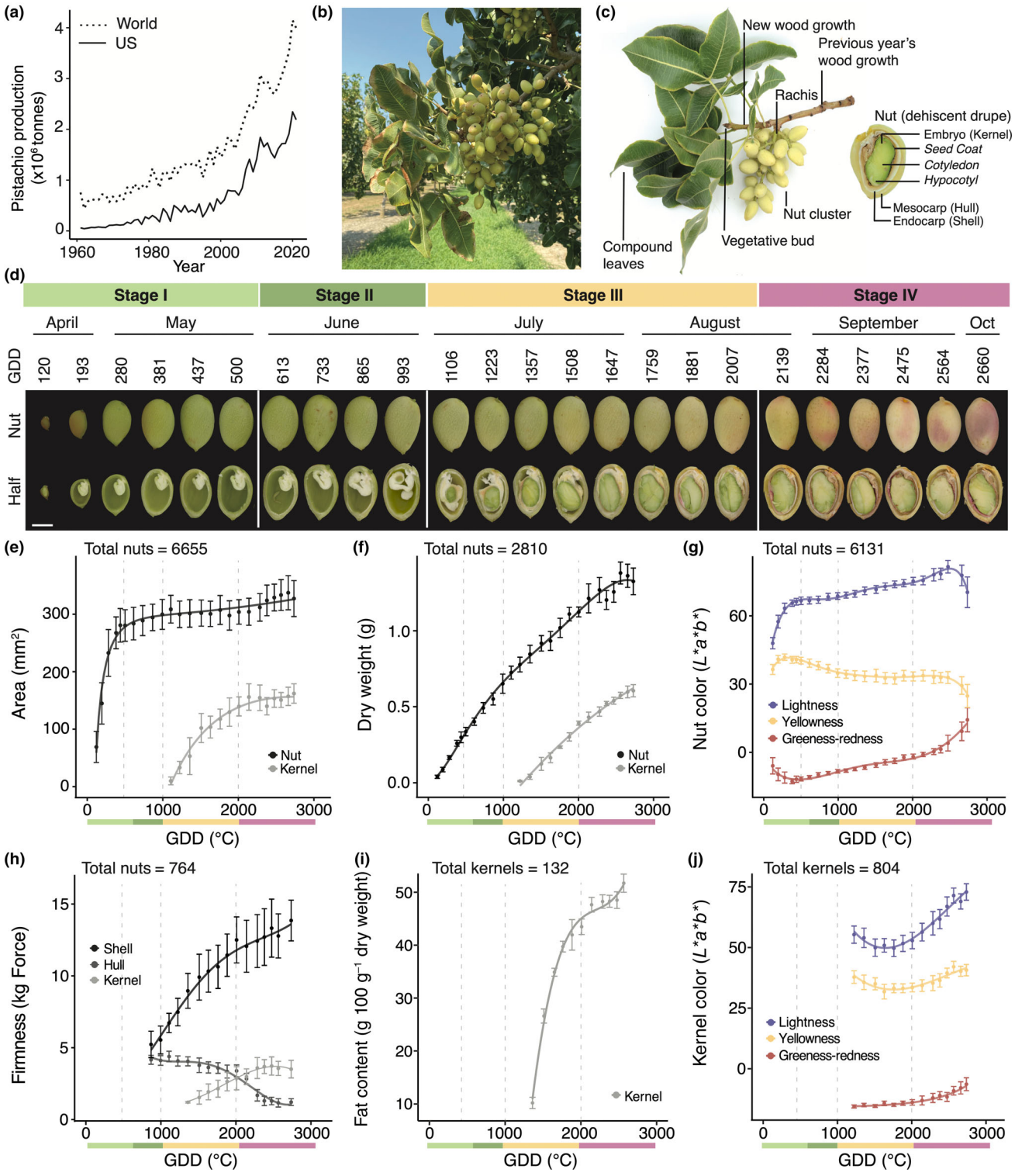


Fig. 1 Pistachio nut development is categorized into four distinct stages. (a) Comparison of United States (US) pistachio (*Pistacia vera*) production to the world in the past 60 years (Food and Agricultural Organization; https://www.fao.org/faostat/en/#search/pistachio). Pistachios ('Kerman') at Stage III on a tree (b) and a branch, (c) with nut and kernel anatomy. (d) Pistachio development (whole nut, halved nut, and kernel) assessed from April to September 2019 in California and categorized into four stages represented by calendar time and accumulated heat units expressed as growing degree days (GDD) in °C. Bar, 1 cm. The new stages were defined by assessing (e) whole nut and kernel area growth (mm^2), (f) dry weight (g) of the whole nut and kernel, (g) color changes in the hull measured in the $L^*a^*b^*$ color space (L^*, or lightness, a^* or redness, and b^* or yellowness), (h) texture changes in the hull, shell, and kernel (kg of Force), (i) fat content in the kernel (g 100 g^{-1} dry weight), and (j) kernel color changes measured in the $L^*a^*b^*$ color space. (e–j) Lines show fitted linear and linear mixed polynomial models as a function of heat accumulation (GDD). Error bars indicate SD from the means. (e–j) The stages are represented in a bar with distinct colors below the *x*-axes. Stage I, light green; Stage II, green; Stage III, yellow; Stage IV, pink.

stages, defined by biometric parameters like nut and kernel size, have guided research and agronomic management in the past; yet, a complete physiological assessment of pistachio development from fruit-set to harvest is lacking.

Also, a better understanding of the molecular mechanisms behind the composition of pistachio kernels can provide a robust foundation for breeding novel pistachio varieties with higher nutritional value and advancing management strategies to boost yield, cut costs, and enhance quality. Such insights provide transferable knowledge that will benefit the development of other hard-shelled nut crops, such as almonds and walnuts.

Here, we present foundational genomic resources for and research into nut development that are critically needed to support the rising demand for pistachio production. We have generated the first chromosome-scale reference-quality genome and annotation of *P. vera* 'Kerman', the most important female cultivar in the United States, leveraging this resource (available at https://pistachiomics.sf.ucdavis.edu/) to address outstanding questions about the molecular genomics of pistachio nut development. We combined transcriptomic and physiological data to uncover pathways and regulators behind the kernel's protein accumulation and high unsaturated fat content. Altogether, this work yields a new model of nut development to significantly expand the knowledge of hard-shelled fruit biology, link key molecular processes with the nutritional profile of pistachios, and support growers in decisions like harvest timing and irrigation.

Materials and Methods

Sample collection for physiological data, RNA-seq experiments, and metabolic analysis

Pistachio (*Pistacia vera* L.) physiological data were collected across three growing seasons (2019, 2020, and 2021) for evaluation from an experimental pistachio orchard ('Kerman' grafted onto UCB1 rootstock) with 30-yr-old trees at the University of California Kearney Agricultural Research and Extension Center (Fresno County) in 2019. The results were validated in 2020 and 2021 in commercial orchards with 10-yr-old trees of the same cultivar located in Woodland, CA (Yolo County), and Three Rocks, CA (Fresno County), respectively. In all the orchards, the pollinizer cultivar was 'Peters'. In each study, homogenous trees ($n_{2019} = 12$, $n_{2020} = 4$, and $n_{2021} = 8$) were randomly selected across the orchard and were continuously sampled throughout the season. In 2019, samplings were conducted weekly after fruit-set on 25 April (120 growing degree days (GDD, in °C)), 10 d postanthesis (dpa) continuing through 14 October (2760 GDD), corresponding to 2 wk after harvest maturity (70% shell split). About 48 clusters were collected per time point from 12 trees. In 2020, samplings occurred weekly from 19 April (24 GDD) to 10 September (2452 GDD, commercial harvest), and in 2021, samplings occurred weekly from 15 July (1531 GDD) to 22 September (2828 GDD). Four whole clusters yielding *c.* 50 nuts of uniform maturity were collected per tree at each sampling. Environmental temperature was recorded using HOBO temperature and light sensors Data Logger (Onset, Bourne, MA, USA) placed in the orchards at three locations. Growing degree days was calculated with the following equation (Zhang *et al.*, 2021):

$$T_{avg} = ((T_{max} + T_{min})/2)\ \mathrm{GDD} = \sum(T_{avg}\text{–Tb})\text{–}7$$

All harvest dates and calculated GDD are summarized in Supporting Information Table S1. A subset of four replicates composed of 12 fruits from three separate trees pooled from the samplings described above were dissected into hull, shell, and kernel tissues and flash-frozen in liquid nitrogen on the day of sampling. The kernel (i.e. embryo) was frozen after removing the seed coat. Frozen tissues were ground into a fine powder with a Retsch Mixer Mill MM 400 for subsequent analysis (Verder Scientific, Haan, the Netherlands).

Physiological measurements and statistical modeling

To assess nut area and color, nuts were imaged longitudinally using a VideometerLab 3 (Videometer, Hovedstaden, Denmark) facilitated by Aginnovation LLC. The VideometerLab 3 software was utilized for image analysis in both 2019 and 2020. Color measurements were taken on 10–30 nuts per tree ($n = 12$) sampled weekly for 25 wk using the $L^*a^*b^*$ color scale across the nut area. Dry weight measurements of the whole nut and the kernels were performed by separating the kernels and the remaining tissues (hull and shell) and drying each at 80°C for 2 d. Dry weight was reported as the average per nut and per kernel weights per cluster ($n = 7$–12 per sampling). Shell split was measured as the proportion of nuts with any degree of separation between sides. Destructive texture measurements were obtained with the TA.XT2i Texture Analyzer (Texture Technologies, Algonquin, IL, USA) using a TA52 2-mm probe with a trigger force of 5 g and a test speed of 2.00 mm s^{-1} with the Exponent software (Texture Technologies Corp) and were reported as kilograms (kg) of force. Peaks of the hull, shell, and kernel tissues were distinguished and recorded in the software. Twenty to sixty nuts were assessed every sampling for 25 wk. Fat content was obtained from oven-dried kernels, as described by Polari *et al.* (2020), and was expressed as grams of fat per 100 g. All physiological trait data across the 3 yr of collections are summarized in Table S1.

Physiological traits, including kernel and nut dry weights, kernel and nut areas, kernel and nut colors, and kernel, shell, and nut textures were modeled against heat accumulation. Various Box–Cox transformations (i.e. square, square root, and logarithm transformation) on traits data using the MASS package (Venables & Ripley, 2003) in R were made before modeling to ensure an approximate normal distribution of traits and roughly equal variance of error terms. Linear and linear mixed models were fitted for each trait with a polynomial of 2 or 3 as a function of heat accumulation, using the LME4 and LMERTest packages in R (Bates *et al.*, 2015; Kuznetsova *et al.*, 2017). Random intercepts were added in linear mixed models. Random effects included cluster, tree, and year, depending on models (Table S2).

Sample collection and preparation for PacBio HiFi, Omni-C, and Iso-Seq sequencing

High molecular weight genomic DNA was extracted from young leaves of 'Kerman' trees located at Finca 'La Entresierra', Centro de Investigación Agroambiental 'El Chaparrillo' (CIAG), Spain, using the Circulomics Nanobind Plant nuclei kit (Pacific Biosciences, Menlo Park, CA, USA) after nuclei isolation according to Workman *et al.* (2018). The library construction and PacBio HiFi sequencing were completed on a Sequel II system at Gentyane facilities in the French National Institute of Agronomy (INRAE). A total of 25.18-Gb PacBio HiFi reads with a mean length of 16.3 kb were generated (Table S3). Omni-C library construction and sequencing were performed by Dovetail Genomics (Cantata Bio LLC, Scotts Valley, CA, USA) with young leaf tissue samples that had been flash-frozen in liquid nitrogen upon collection and stored at −80°C. Samples from various developmental stages were collected for isoform sequencing (Iso-seq) from the same 'Kerman' tree used for genome assembly (dormant buds on 23 March, breaking buds and flowers on 6 April, leaves on 16 April, and fruits on 6 May 2021) and were immediately flash-frozen in liquid nitrogen. One hundred milligrams of material for each sample was pulverized with a mortar and pestle in liquid nitrogen. Total RNA was extracted using the Spectrum Plant Total RNA kit (Sigma-Aldrich), according to the protocol for 'difficult' species. The quantity and purity of the total RNA were checked with the Nanodrop (ND1000; Thermo Fisher Scientific, Carlsbad, CA, USA). Iso-seq library preparation and sequencing were performed using RNA samples with *c.* 400 ng μl^{-1} and were barcoded and sequenced in the same Sequel II SMRT cell at Gentyane facilities in INRAE.

Genome survey

Genome size, heterozygosity, and repeat content were estimated based on k-mer frequency analysis with PacBio HiFi reads. Jellyfish v.2.2.10 41 (Marçais & Kingsford, 2011) was used to count 21-mers with a maximum *k*-mer depth of 1e6, taking repetitive regions into account. The resulting histogram from Jellyfish was subjected to GenomeScope v.1 web (Vurture *et al.*, 2017) to estimate genome size, levels of heterozygosity, and repeat content.

Genome assembly and chromosome construction

A *de novo* assembly of PacBio HiFi reads into contigs was performed using Hifiasm v.0.16.0 (Cheng *et al.*, 2021) with default parameters. Organelle (plastid and mitochondrial) origin contigs were filtered out from the primary contig assembly using 'Kerman' plastid and mitochondrial sequences from PacBio HiFi reads in Geneious Prime (https://www.geneious.com). The high-coverage Omni-C data (99.92 Gb and *c.* 172× depth) was quality-checked with the FastQC toolkit (Andrews, 2010) and aligned to the filtered primary contig assembly using Juicer v.1.6 (Durand *et al.*, 2016). These aligned read pairs were utilized to scaffold the assembly into 15 chromosomes based on chromatin interaction information using 3D-DNA (Dudchenko *et al.*, 2017) with default parameters and manual correction on Juicebox (Robinson *et al.*, 2018). The filtered primary contig assembly was scaffolded again with 3D-DNA scaffolds as a reference using RagTag v.2.1.0 with default parameters (Alonge *et al.*, 2022), followed by manual correction by comparing RagTag and 3D-DNA scaffolds. Chromosome numbers and orientations from previously published pistachio genomes were used (Kafkas *et al.*, 2023). The completeness of assemblies, scaffolds, and 15 chromosomes was assessed using Benchmarking Universal Single-Copy Orthologs (Busco) v.5.4.4 (Simão *et al.*, 2015) with the embryophyta_db10 database. The overall assembly and chromosome construction workflow is provided in Fig. S1.

Genome synteny analysis

The complete genome and annotation of *Mangifera indica* (mango; Wang *et al.*, 2020) in the same family, Anacardiaceae, and *Citrus sinensis* (sweet orange; Wu *et al.*, 2014) in the same order, Sapindales, were downloaded to conduct a comparative analysis. Synteny of the chromosome-level genome of 'Kerman' was compared with that of mango and sweet orange using GENESPACE v.0.9.4 (Lovell *et al.*, 2022), which takes orthology information into account using OrthoFinder v.2.5.4 (Emms & Kelly, 2019). The manual reformation of gene annotation and protein files was needed for GENESPACE input.

Genome annotation

Annotation of transposable elements (TE) was accomplished with primary contig assembly using Extensive de novo TE Annotator (EDTA) v.2.1.0 (Ou *et al.*, 2019). Sequences from the generated nonredundant TE library that overlapped with the filtered plant protein database were filtered out using protExcluder v.1.2 (https://www.canr.msu.edu/hrt/uploads/535/78637/ProtExcluder1.2.tar.gz) to hinder the exclusion of genes in the gene prediction analysis. Output from protExcluder was employed to re-annotate TEs in the primary contig assembly using EDTA.

A softmasked 'Kerman' assembly, RNA sequencing (RNA-seq), and Iso-seq data were used. PacBio Iso-seq data (see 'Sample collection, preparation, and sequencing for PacBio HiFi, Iso-Seq, and Omni-C' section) were assembled from five different tissues using the IsoSeq3 pipeline (Pacific Biosciences), including: demultiplexing and primer removal using lima v.2.0.0; removing poly(A) tails and concatemers; clustering isoforms; and mapping to the assembly using pbmm2 v.1.9.0 and collapsing isoforms. The quality of RNA-seq data from pistachio hull and shell tissues (see 'RNA preparation, RNA-seq, and data processing of nut tissues for spatiotemporal study' section) was assessed using FastQC v.0.11.9 (Andrews, 2010). Raw RNA-seq reads were filtered and adapter-trimmed using Trimmomatic v.0.39 (Bolger *et al.*, 2014) with the following parameters: maximum seed mismatches = 2, palindrome clip threshold = 30, simple clip threshold = 10, minimum leading quality = 3, minimum trailing quality = 3, and minimum length = 36, and clean reads were mapped to the assembly using Hisat2 v.2.2.1 (Kim *et al.*, 2019).

Iso-seq demultiplexed reads were aligned to the assembly using MINIMAP2 (Li, 2018). Aligned RNA- and Iso-seq reads were independently assembled into transcripts using STRINGTIE v.2.2.1 (Pertea *et al.*, 2015). Before *ab initio* gene prediction, the aligned Iso-seq data and softmasked assembly were used as a retraining set for BRAKER2 v.2.1.2 (Brůna *et al.*, 2021). AUGUSTUS v.3.1.0 (Keller *et al.*, 2011) was used to make the *ab initio* prediction with retraining files, softmasked assembly, and exon hints from Iso-seq isoforms produced in the IsoSeq3 pipeline. The assembled RNA- and Iso-seq data and Augustus output gene models were combined to find consensus gene models using EVIDENCEMODELER (EVM) v.2.0.0 (Haas *et al.*, 2008) with different weights for each input data (7, 4, and 1 for Iso-, RNA-seq transcripts, and Augustus gene models, respectively). Finally, 5′ and 3′ untranslated regions and different isoforms were updated from EVM gene models using PASAPIPELINE v.2.5.2 (Haas *et al.*, 2008) based on the alignment of Iso-seq demultiplexed reads. The workflow of the gene annotation pipeline is provided in Fig. S2. The final gene annotation on primary contig assembly was lifted over to 15 chromosomes, and gene IDs were curated using modified python script GFF_RenameThemAll.py (Morales-Cruz *et al.*, 2021). The quality of the final gene models was assessed using BUSCO (Simão *et al.*, 2015; embryophyta_db10 set) with protein sequences extracted by using ANOTHER GTF/GFF ANALYSIS TOOLKIT (AGAT) v.0.9.1 (Dainat *et al.*, 2022). The RIDEOGRAM R package (Hao *et al.*, 2020) was used to plot gene and TE density on the 15 chromosomes in the 'Kerman' genome assembly. The 3D structures of gene models were characterized in an atomic resolution with a language model using ESMFold (Lin *et al.*, 2023) and visualized using UCSF CHIMERAX v.1.6.1 (Pettersen *et al.*, 2021).

tRNAs genes were identified using TRNASCAN-SE v.2.0.12 (Lowe & Chan, 2016) with eukaryote parameters. rRNA genes were predicted using BARRNAP v.0.9 (https://github.com/tseemann/barrnap#barrnap) with Eukaryota database and categorized in 5S, 5.8S, 18S, and 28S rRNAs. INFERNAL v.1.1.4 (Nawrocki & Eddy, 2013) was used to search for miRNAs and snoRNAs based on Rfam14.9 database (Kalvari *et al.*, 2021; Table S4).

Functional annotation and enrichments

Ortholog genes were identified by searching NCBI BLAST (McGinnis & Madden, 2004) with default parameters from the NR database using default parameters. The top BLAST was reported for each gene. The Kyoto Encyclopedia of Genes and Genomes Orthology (KO) and pathways were identified using the Automatic Annotation Server (KAAS, v.2.1; http://www.genome.jp/kegg/kaas/) with bi-directional best hit and BLAST as a search program against closely related organisms, mallow, rose, mustard, and papaya. The plant Transcription factor & Protein Kinase Identifier and Classifier (iTAK) online tool v.1.6 (http://itak.feilab.net/cgi-bin/itak/online_itak.cgi) was used to identify transcription factor (TF) families with default settings. Enrichment analysis for all functional annotations was performed using a Fisher's test, and the resulting *P*-values were adjusted with the Benjamini and Hochberg method (Benjamini & Hochberg, 1995). Enrichments with $P_{adj} < 0.05$ were considered significant.

RNA preparation, RNA sequencing, and data processing of nut tissues for spatiotemporal study

RNA extractions were performed on 2019 samplings from 865 to 2564 GDD for hull tissues, 865 to 2139 GDD for shell tissues, and 1106 to 2564 GDD for kernel tissues, as described in 'Sample collection for physiological data, RNAseq experiments, and metabolic analysis' section. One gram of ground tissue was used for RNA extraction as described in Blanco-Ulate *et al.* (2013). RNA concentrations were quantified with Nanodrop One Spectrophotometer (Thermo Scientific, Lenexa, KS, USA) and Qubit 3 (Invitrogen), and RNA integrity was assessed on an agarose gel.

RNA was extracted and sequenced from at least three biological replicates per time point and tissue. cDNA libraries were prepared with Illumina TruSeq RNA Sample Preparation Kit v.2 (Illumina, San Diego, CA, USA). The quality of the barcoded cDNA libraries was assessed with the High Sensitivity DNA Analysis Kit in the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and then sequenced (150-bp paired-end reads) on the Illumina HiSeq X Ten platform by IDseq Inc. (Davis, CA, USA).

Raw reads were trimmed for quality and adapter sequences using TRIMMOMATIC v.0.39 (Bolger *et al.*, 2014) with the following parameters: maximum seed mismatches = 2, palindrome clip threshold = 30, simple clip threshold = 10, minimum leading quality = 3, minimum trailing quality = 3, window size = 4, required quality = 15, and minimum length = 36. Trimmed reads were then mapped using BOWTIE2 v.2.3.4.1 (Langmead & Salzberg, 2012) to the predicted pistachio 37 955 protein sequences, producing an average of 76.41% mapping percentage across all samples (Table S5). Count matrices were made from the BOWTIE2 results using sam2counts.py v.0.913 (https://github.com/vsbuffalo/sam2counts/blob/master/sam2counts.py). Read counts were normalized with the Bioconductor package DESEQ2 in R (Love *et al.*, 2014). Gene expression visualizations used the median of ratios normalization for cross-sample comparison (Table S6). rRNA genes identified (see 'Genome annotation' section) were excluded from the count matrix for gene expression analysis.

Gene co-expression analysis

Count matrices were normalized for gene co-expression analysis with the variance stabilizing transformation function from the DESEQ2 package in R. After filtering normalized reads with a count of < 10 reads in 90% of each tissue's samples checking resulting subset with the goodSamplesGenes function, weighted gene co-expression network analysis was conducted for each tissue type separately with the WGCNA package in R (Langfelder & Horvath, 2008). Dendrogram plots were made to identify and remove sample outliers. One kernel sample was removed

($h > 200$), and four hull samples were removed ($h > 120$). The network was constructed step by step. The soft thresholding power was selected to obtain the approximate scale-free topology where the index reaches 0.9. The adjacency was calculated with power being equal to the selected soft power and the type being signed type. TOMtype was signed. The initial module eigengenes were detected by default hierarchical clustering function with default parameters other than the appropriate minModuleSize selected for each tissue; then, final module eigengenes were obtained by merging close modules with cutHeight = 0.25. The module eigengenes values were then used to identify modules significantly associated with measured physiological traits by calculating correlations using Pearson's product–moment correlation coefficient.

Macronutrient measurements

Nonstructural carbohydrates (NSC) were analyzed in pistachio kernels. Frozen ground tissue was utilized as described in 'Sample collection for physiological data, RNAseq experiments, and metabolic analysis' section. The concentration of NSC, including soluble sugars and starch, was analyzed as described in previous publications (Leyva *et al.*, 2008; Zwieniecki *et al.*, 2022), with some modifications. For soluble sugar extraction, 25 mg of sample was mixed with 1 ml 0.2 M sodium acetate buffer, pH 5.5, and incubated at 70°C for 15 min. After centrifugation at 21 000 ***g*** for 10 min, the supernatant was diluted in bi-distilled water (1 : 10, v/v). Soluble sugars were then quantified by adding 0.1% anthrone in 98% sulfuric acid (m/v), incubating at 100°C for 20 min, cooling at room temperature for 10 min, and reading absorbance at 620 nm in a photometer (Multiskan FC; Thermo Scientific) and using a glucose standard curve to compare the colorimetric response of the samples. For starch extraction, the remaining buffer and pellet were incubated at 100°C for 10 min and cooled for 20 min. Subsequently, 0.7 U of amylase and 7 U of amyloglucosidase (Sigma-Aldrich) were added. The samples were stirred in a rotary incubator at 37°C for 4 h. Samples were then centrifuged at 21 000 ***g*** for 10 min, and the supernatant was diluted in bi-distilled water (1 : 10, v/v). Starch concentration was determined by subtracting prestarch digestion soluble sugar concentration from total (pre- and poststarch digestion) soluble sugar concentration. Three repetitions per sample were analyzed, and the results were averaged.

Protein content was measured as crude protein with AOAC Official Method 990.03 at the UC Davis Analytical Laboratory (University of California, Davis, CA, USA). Total crude protein was calculated from the nitrogen content, with the protein factor 6.25 applied to the nitrogen result.

Metabolite profiles

Frozen ground kernel samples collected in 2019, as described in 'Sample collection for physiological data, RNAseq experiments, and metabolic analysis' section were used for all metabolic analyses, unless noted. Volatile profiles were extracted and analyzed by gas chromatography following the methodology described in Polari *et al.* (2020). Volatile compounds were identified by their mass spectra and Kovats Retention Index. Results were expressed as nanograms of dodecane per kilogram of sample (ng kg^{-1}). For phenolic profiles, three biological replicates were analyzed for three sampling dates in 2021: 25 August (2162 GDD), 31 August (2279 GDD), and 7 September (2402 GDD). Total phenolics were extracted from the kernels following the protocol detailed in Grilo & Wang (2021). Phenolic extracts were membrane-filtered with 0.45 µm cellulose and subjected to HPLC-DAD analysis as described by Erşan *et al.* (2017) HPLC-DAD model (Agilent G4212-60008, serial no.: DEBAF01604; Agilent Technologies) was used with the following solvents: HPLC grade water (99%, Solvent A) and HPLC grade methanol (99%, Solvent B), with each solvent containing 1% (v/v) formic acid. The column was an Eclipse Plus C18 column (250 mm × 4.6 mm, 5 µm; Agilent Technologies) with a security guard ultra C18 guard column (4.6 mm × 2 mm) of the same material. The gradient included the following: isocratic at 2% B for 10 min, then 2 to 37% B in 27 min, isocratic at 37% B for 5 min, then from 37 to 40% B in 18 min, from 40 to 60% B in 10 min, from 60 to 100% B in 20 min, isocratic at 100% B for 14 min, then from 100 to 2% B in 1 min, and isocratic at 2% B for 7 min. The column temperature was 35°C, and the total run time was 112 min at a flow rate of 1 ml min^{-1} and injection volume of 3 µl. The calibration curves with the external standards were obtained using concentration (mg l^{-1}) with respect to the area obtained from the integration performed at 280 nm (gallotannins), 310 nm (anacardic acids), and 350 nm (flavonols). Results were expressed as milligram per gram of kernel dry weight.

Fatty acids were extracted and analyzed by gas chromatography according to Polari *et al.* (2020), as described. Relative fatty acid proportions were determined by peak area normalization. Individual fatty acids were expressed as the percentage of total fatty acids.

Abscisic acid (ABA) and jasmonic acid (JA) hormones were measured on hull, shell, and kernel tissues collected at 1881 GDD in 2019 (see 'Sample collection for physiological data, RNA-seq experiments, and metabolic analysis' section). Samples were extracted and analyzed by Metware Biotechnology Inc. (Boston, MA, USA) using an UPLC-ESI-MS/MS system.

PvNFYB-LEC1 and PvAP2-WRI1 genes and predicted binding motif analysis

The WRINKLED1 (WRI1; ID: Q6X5Y6) and LEAFY COTYLEDON1 (LEC1) proteins (ID: Q9SFD8) of *Arabidopsis thaliana* from the UniProt (https://www.uniprot.org/) were used to identify homologs of AtWRI1 and AtLEC1 in 'Kerman' and 14 other taxa (Table S7). *Arabidopsis thaliana* WRI1 was queried with BLAST against filtered proteomes of 15 taxa. BLAST hits with similar sizes to *A. thaliana* WRI1 genes were analyzed for the conservativeness of two AP2/EREBP DNA-binding domains. For LEC1, because the sequences outside of the central conserved region showed high divergence between AtLEC1, BnLEC1 (*Brassica napus*, EU371726), and GmLEC1 (*Glycine max*,

Glyma.07G268100), 103 amino acid sequences in the highly conserved region in AtLEC1 were used for the BLAST search with e-value 1e−10. BLAST hits were selected with the highest similarity, and genes were extracted from the proteomes as possible LEC1 genes. Amino acid sequence alignments of all 16 WRI1 and LEC1 genes were performed using MAFFT v.7.505 (Katoh & Standley, 2013). The 3D structure of all WRI1 and LEC1 proteins was predicted using COLABFOLD (Mirdita *et al.*, 2022) and ALPHAFOLD2 (Jumper *et al.*, 2021) and visualized in UCSF CHIMERAX (Pettersen *et al.*, 2021) with coloring based on the predicted local distance difference test (pLDDT) scores. To search for binding sites of WRI1 and LEC1 genes in the 'Kerman' genome, the consensus sequence (CNTNG(N_7)CG) of AW-box binding site for WRI (Maeo *et al.*, 2009; Kuczynski *et al.*, 2022; Sánchez *et al.*, 2022) and 'CCAAT' for LEC1 were examined in nucleotide sequences 1500-bp upstream of the translational initiation site (TIS) of gene models in 'Kerman' genome with a focus on genes associated with fatty acid biosynthesis and the K-III-1 gene module. The consensus sequences of AW-box binding sites were visualized using WebLogo (Crooks *et al.*, 2004). The distribution of WRI1 binding sites was calculated to examine the relative distance from the TISs in genes of fatty acid biosynthesis and the K-III-1 gene module compared with that from other genes with the AW-box binding motif.

Results

Pistachio nuts develop in four distinct stages

Pistachio kernels develop asynchronously from the maternal tissues (hull and shell; Lin *et al.*, 1984; Fig. 1c). To investigate the dynamics of kernel growth in relation to the whole nut, we conducted the most comprehensive study of nut development to date, evaluating physiological traits, such as size and firmness, in pistachio nuts ('Kerman', $n = 663$–6805) for 24 wk from fruit-set to 2 wk after harvest maturity (Materials and Methods section; Table S1; Fig. 1d). The traits were validated across two additional independent field seasons in distinct geographical locations (Materials and Methods section; Tables S1, S2; Fig. S3). We used both GDD (in °C) and calendar-determined time in our analyses because heat accumulation modulates phenology and enzymatic activities that influence metabolism and nut growth (Corelli-Grappadelli & Lakso, 2004; Marino *et al.*, 2018). These data revealed that four stages define nut development: Stage I (from late April through May in California) occurs when the shell and hull tissues grow in a logarithmic growth pattern plateauing at 500 GDD (Fig. 1e,f). During this time, the hull and shell tissues are fused together and display an increasingly green color (Fig. 1g). Stage II (June) is the transition period before the beginning of kernel (embryo) growth, when hull and shell expansion stops but the tissues continue to accumulate dry weight (Fig. 1f). Stage III (late June to mid-August) corresponds to the kernel growth phase, starting at 1000 GDD and reaching its maximum size at 2000 GDD. In contrast to previous reports (Goldhamer & Beede, 2004), we found that shell hardening coincides with kernel growth at Stage III and continues through Stage IV (Fig. 1e,h). Stage IV (from late August to September) marks the onset of kernel maturity, at which the kernels reach their maximum size and fat content despite the continued increase in dry weight (Fig. 1i). Kernel maturity coincides with hull ripening (e.g. tissue softening and color changes) and shell split or dehiscence. At this stage, the kernels start losing their deep green coloration, preparing for seed dormancy (Fig. 1g, h). These findings show that pistachio nut development is closely regulated by heat accumulation, enabling accurate prediction of growth stages across seasons.

A reference-quality genome and annotation of pistachio

Our study aimed to characterize the molecular genomics of pistachio nut development. To enable this, and future pistachio research more generally, we generated a high-quality chromosome-scale genome assembly and annotation for *Pistacia vera* 'Kerman'. This reference genome for pistachio surpasses previous efforts in accuracy and completeness, providing a foundation for breeding and -omics research and facilitating transcriptomic analysis (Table 1). Our assembly illuminates previously undescribed details of the pistachio genome architecture, particularly the presence of significantly large repetitive, knob-like regions on 11 chromosomes (Fig. 2c). These regions, characterized by megabase-scale 178-bp satellite repeats, reveal a remarkable genomic feature that had been observed cytologically (Sola-Campoy *et al.*, 2015), but was previously obscured from genome assemblies. This discovery reshapes our understanding of the pistachio genomic landscape and highlights the value of robust genome assembly for molecular genomic and evolutionary research.

We resolved the 'Kerman' genome to 15 chromosomes (559.11 Mb) with only 12 gaps, by scaffolding and manually curating 102 nonorganelle contigs (579.8 Mb) with high-depth Omni-C data (Fig. 2b,c; Tables 1, S8). The k-mer analysis with PacBio HiFi reads (Table S3) estimated the 'Kerman' genome size to be *c.* 521 Mb, with moderate heterozygosity (0.755%) and repetitiveness (54.1%; Fig. S4). Notably, our assembly has a significant reduction in the number of duplicated complete BUSCO genes compared with previous reports, from between 210 and 286 (previous) to only 45 (this study), indicating the elimination of false duplications (Table S9).

The 'Kerman' genome contains over 0.83 million repetitive elements constituting 376.56 Mb (*c.* 65%) of the assembled genome (Fig. S5; Table S10). This notable increase from previous genomes (53 and 56%; Table S10) implies an accurate recovery of repetitive regions while successfully excluding false-segmental duplications explained by a smaller assembly size (559.11 Mb) than the other assemblies, 'Batoury' (671 Mb; Zeng *et al.*, 2019), and 'Siirt' (596 Mb) and 'Bagyolu' (623.4 Mb; Kafkas *et al.*, 2023). Consistent with other plant species (Hawkins *et al.*, 2009; Bento *et al.*, 2013), long terminal repeat retrotransposons were the most abundant (48.95%) class in the 'Kerman' genome. We also discovered regions containing massive enrichments of 178-bp repeats known as PIVE-180 on one arm of 11 chromosomes in pistachio (e.g. up to *c.* 9 Mb on chromosome 7, where no protein-coding genes were present; Fig. 2c). Importantly, the presence of these regions has been confirmed by

Table 1 Characteristics of the *Pistacia vera* 'Kerman' genome and annotation.

	P. vera 'Kerman'	
Genome assembly	Primary assembly	Chromosomes
No. of Seq.	102	15
Size (bp)	579 837 297	559 112 681
Max length (bp)	49 785 438	62 545 632
N50 (bp)	25 977 840	36 813 375
L50	9	7
L90	21	13
No. of gaps		12
%GC	37	36.67
BUSCO completeness (%)	98.8	98.9
Complete single-copy (%)	95.4	95.8
Complete duplicated (%)	3.3	2.8
Protein-coding gene		Chromosomes
No. of genes		37 063
Total length of genes (bp)		108 162 188
% of genome covered by genes		19.35
Mean gene length (bp)		2918
Mean CDS length (bp)		1073
Mean exon length (bp)		285
Mean intron length (bp)		436
Mean exon number per gene		4.6
BUSCO completeness (%)		98.9
Complete single-copy (%)		97.2
Complete duplicated (%)		1.7

Noncoding loci	Number	Total (bp)
tRNA	514	38 570
rRNA	848	895 039
miRNA	182	24 122
snoRNA	2123	223 030

BUSCO, Benchmarking Universal Single-Copy Orthologs. CDS, coding sequence.

previous cytological work, with the location of these repetitive sequences being distinct from centromeres and corresponding with those observed in our genome assembly (Sola-Campoy *et al.*, 2015). The existence of these extremely repetitive satellite DNAs likely limited the accuracy of previous genome assemblies and chromosome construction. The syntenic analysis showed overall collinearity but a noticeable presence and absence of variation of these and other genomic regions, which implies a significant improvement in the quality of the pistachio genome (Fig. S6).

We annotated 37 955 protein-coding genes in the 'Kerman' genome with multiple lines of evidence (see the Materials and Methods section). Among these, multiple isoforms were characterized in 6938 genes, and 22 942 genes were functionally annotated with multiple databases (see the Materials and Methods section; Table S8). The majority of genes (> 97%, 37 063) were anchored onto the 15 chromosomes (Table 1). The gene annotation demonstrated exceptional quality, reaching nearly 99% completeness in BUSCO scores (Table S11). A complete gene annotation, essential for our study of pistachio nut development and other applied pistachio research, was achieved with high-quality assembly, long- and short-read transcript evidence from various tissue types, and *ab initio* predictions and improved upon previous annotations, which had BUSCO completeness scores under 94% with over 12% duplication (Fig. S7; Tables S11, S12; see the Materials and Methods section). The genome assembly and annotation of *P. vera* 'Kerman' enabled more accurate orthology-constrained macrosynteny analysis between the genomes of pistachio and its close relatives, mango and sweet orange, revealing extensive structural variation even within the family and whole genome duplication in the mango genome (Fig. 2d).

Spatiotemporal gene expression data support the distinct nut developmental stages

Previous transcriptomic analyses of *P. vera* have focused on vegetative tissues. Including fruit tissues in the genome annotation of 'Kerman' provided a resource that enabled the investigation of nut development at the molecular genomic level and increased read mapping coverage (Table S5). We measured tissue-specific whole transcriptome gene expression across 15 wk of nut development encompassing kernel growth and maturation (the end of Stage II through Stage IV; Figs 1d, 2e). These data are available as an atlas of spatiotemporal expression (https://pistachiomics.sf.ucdavis.edu/).

The kernels used in this study resulted from open pollination of 'Kerman' (maternal) with Peters (paternal; the male cultivar used as a pollenizer in 'Kerman' orchards). Based on their transcriptional patterns, we distinguished the maternal tissues (hull and shell) from the hybrid kernels, reflecting differences in genetic backgrounds and ontogenic states (Fig. 3a; PC1). We also observed differences in transcriptional patterns across nut development stages (Fig. 3a; PC2).

We identified groups of genes, *that is* co-expressed gene modules, in each tissue (H-hull, S-shell, and K-kernel) with high expression at specific developmental stages (I–IV) that strongly correlated with the occurrence of physiological changes, providing critical insights into the link between molecular processes and the four nut developmental stages (Figs 3b,c, S8; Table S6). Fat content in the kernel was highly correlated to K-III-1, which was enriched for 'fatty acid biosynthesis' and 'fatty acid elongation' functions. Shell lignin biosynthesis genes were highly expressed and enriched ('phenylpropanoid biosynthesis') in the S-II-1 gene module, correlating to early shell hardening (Stages II to III) when lignin is deposited (Table S13). Similarly, the cell wall-degrading enzymes, pectate lyase and a β-glucosidase, potentially involved in hull softening, were the top expressed genes in their respective modules, H-IV-1 and H-IV-2, coinciding with hull ripening in Stage IV (Fig. 3c).

Gene modules reveal pathways involved in kernel nutritional quality

The molecular events occurring in the kernel during Stages III and IV led to the gain of nutritional traits (Fig. 4a). During Stage III, kernels had a deep green color and grew rapidly (Fig. 1f,j).

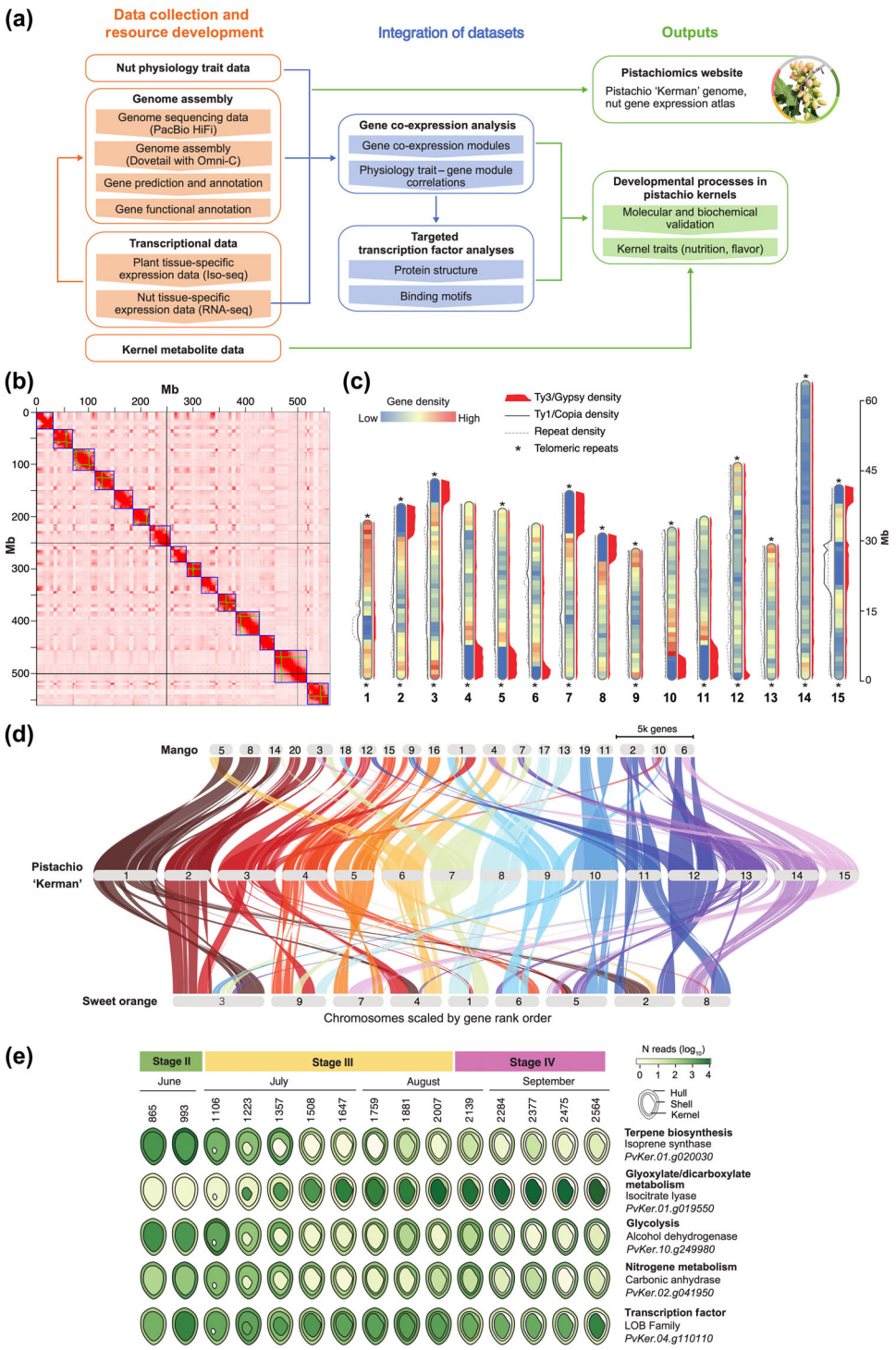


Fig. 2 Chromosome-scale genome assembly of *Pistacia vera* 'Kerman' offers new genetic resources and tools. (a) Overview of current study workflow, including data collection, integration of datasets, and outputs. (b) Heat map of the Omni-C interaction density among 15 chromosomes. The red color indicates the intensity of interactions between genomic regions. Green and blue lines are contigs and chromosomes, respectively. (c) Ideogram with protein-coding gene and repeat density in 1-Mb window size on 15 chromosomes of the 'Kerman' genome. Protein-coding gene density is represented in each chromosome in heatmap style, and repeat density is plotted to the right side of each chromosome in red. The density of long terminal repeat retrotransposons (LTR-RTs) Ty3/Gypsy and Ty1/copia are shown on the left side of each chromosome in normal and dotted lines, respectively. The scale bar for chromosome size is indicated on the right. (d) Macrosynteny analysis of the 'Kerman' 15 chromosomes compared with the mango and sweet orange genomes. (e) Tissue-specific RNA-seq expression of genes highly expressed unique to pistachio nuts identified by comparison of Iso-seq collapsed isoforms demonstrating differential expression patterns across tissues. CDS, coding sequence.

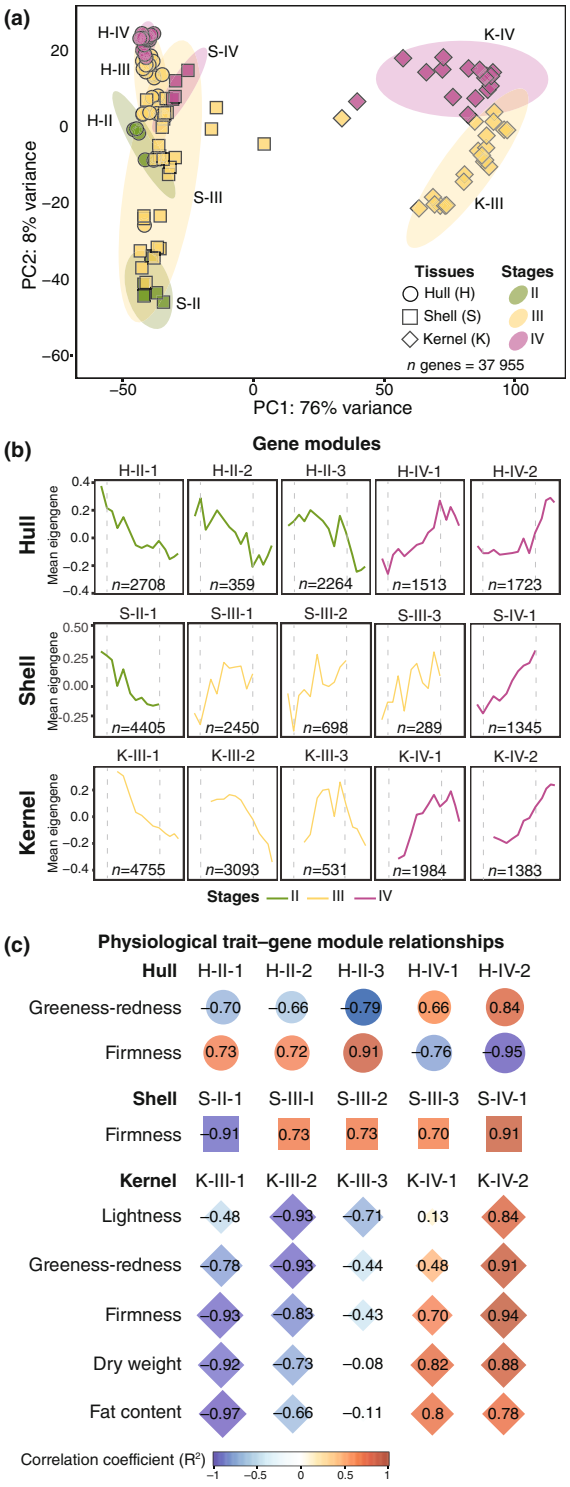


Fig. 3 Gene co-expression patterns confirm pistachio (*Pistacia vera* 'Kerman') developmental stages. (a) Principal component analysis of total gene expression (normalized reads) for all samples, marked by stage (color) and tissue (shape). Then, a weighted gene co-expression network analysis was conducted for each tissue and produced modules of genes with similar expression patterns. (b) Gene modules were selected with high correlations to physiological traits and categorized by stage according to the time points in which expression was elevated (mean eigengene value) in each tissue type (H-hull, S-shell, or K-kernel) and at specific developmental stages (II, III, or IV) for each module (1–X). Each module graph indicates the mean eigengene value at each time point along the *x*-axis. Gray dashed lines indicate the transitions between stages along the *x*-axis. Colors of lines correspond to the stages each module is categorized as (green as II, yellow as III, and magenta as IV). All gene modules can be found in Supporting Information Table S13. (c) Correlation using Pearson's product–moment correlation coefficient between the expression profiles (module eigengene values) of the selected modules and the physiological trait data. Correlations $R^2 > 0.7$ with significance $P < 0.01$ are shown for each tissue. The intensity of the color indicates the strength of the correlation (R^2), and the shape of the icon indicates the tissue type.

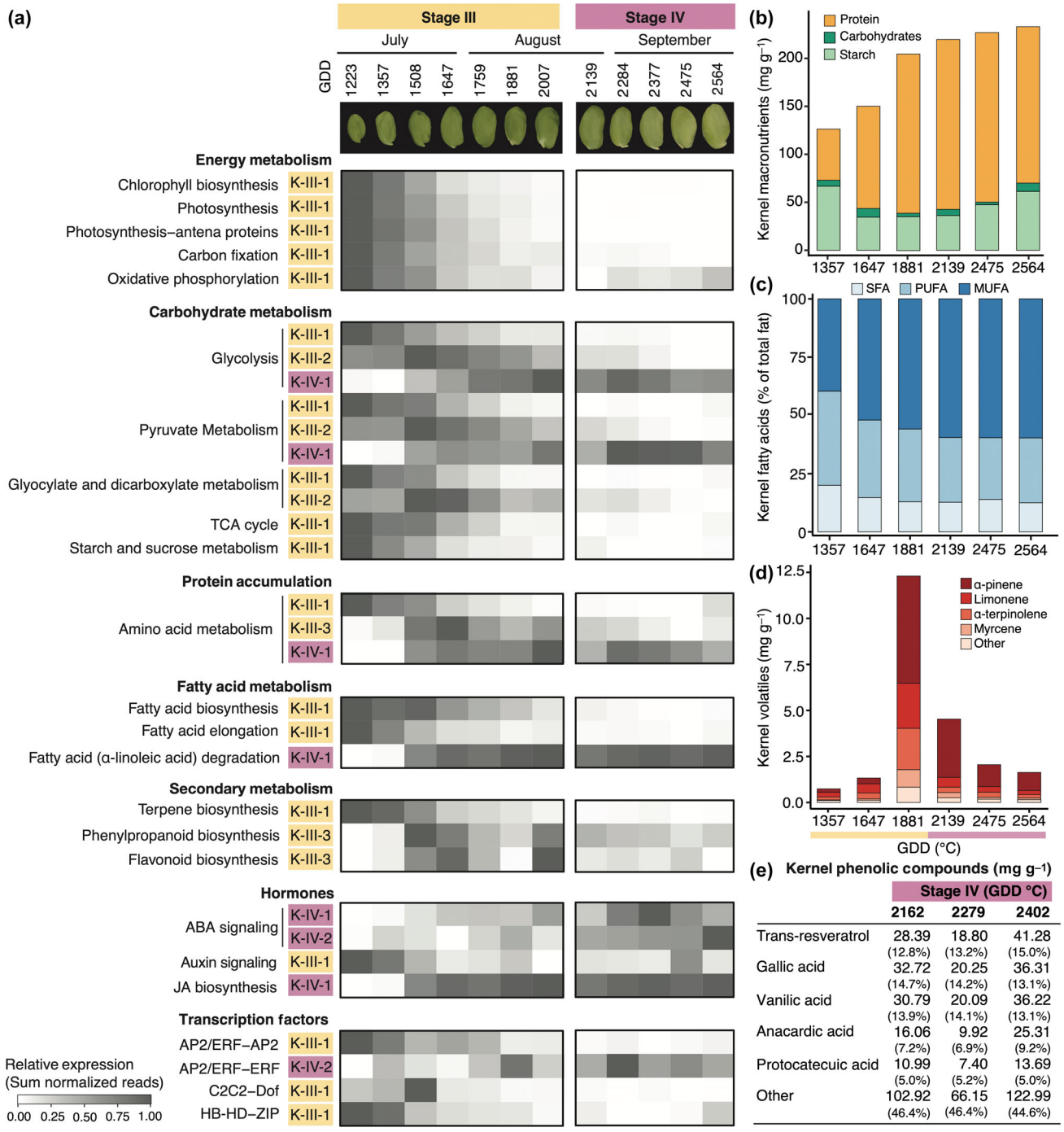
We identified the gene module K-III-1 to be highly correlated with color and the rapid growth during Stage III and found enrichments in functions involved in energy metabolism ('photosynthesis'), carbohydrate metabolism ('starch and sucrose metabolism'), and protein accumulation ('amino acid metabolism'; Table S13). This was supported by increased protein, carbohydrate, and starch content in the kernels during Stage III, with protein making up the highest proportion of these macronutrients (Fig. 4a,b). The enrichment of fatty acid degradation genes, specifically of α-linoleic acid, coincided with the plateau of fatty acid content measured during Stage IV. Furthermore, the ratio of monounsaturated fatty acids (MUFA) to polyunsaturated fatty acids (PUFA) increased during Stages III and IV (Fig. 4a,c; Table S14).

Production of volatile terpenoids and flavonoids in the kernel produces flavor and antioxidant compounds, respectively. We observed a large peak in monoterpenes at the start of Stage IV (1881 GDD), shortly after the peak expression of terpene biosynthesis genes (Fig. 4a,d; Tables S14, S6). The rate-limiting enzyme geranyl-diphosphate synthase and monoterpene synthases were among the genes co-expressed during Stage III (primary K-III-I), which can explain the large increase in volatiles in the *trans*ition to Stage IV (Table S6). At harvest, α-pinene was the main monoterpene present in the kernels, consistent with previous findings (Noguera-Artiaga *et al.*, 2019; Polari *et al.*, 2019). Total phenolic compounds increased during Stage IV, supporting claims and previous reports that pistachio kernels can be an important source of antioxidants (Noguera-Artiaga *et al.*, 2019; Polari *et al.*, 2019). Likewise, phenylpropanoid biosynthesis (K-III-3 and K-IV-6) and flavonoid biosynthesis (K-III-3) were enriched, and genes encoding the rate-limiting steps phenylalanine ammonia-lyase and chalcone synthase were expressed in each respective module. Unlike studies in other varieties, *trans*-resveratrol (41.28 mg g^{-1}) was the most abundant among measured phenols, indicating it may be a major antioxidant in 'Kerman' kernels (Gentile *et al.*, 2007; Tomaino *et al.*, 2010; Liu *et al.*, 2014; Fig. 4e; Table S14).

Conserved transcriptional regulators are associated with fatty acid accumulation in pistachio kernels

Abscisic acid signaling components, including homologs of PYR, SnRK2, ABF, and PP2C, were enriched and highly expressed in

New Phytologist



(e) Kernel phenolic compounds (mg g–1)

	Stage IV (GDD °C)		
	2162	2279	2402
Trans-resveratrol	28.39 (12.8%)	18.80 (13.2%)	41.28 (15.0%)
Gallic acid	32.72 (14.7%)	20.25 (14.2%)	36.31 (13.1%)
Vanilic acid	30.79 (13.9%)	20.09 (14.1%)	36.22 (13.1%)
Anacardic acid	16.06 (7.2%)	9.92 (6.9%)	25.31 (9.2%)
Protocatecuic acid	10.99 (5.0%)	7.40 (5.2%)	13.69 (5.0%)
Other	102.92 (46.4%)	66.15 (46.4%)	122.99 (44.6%)

Fig. 4 Pistachio (*Pistacia vera* 'Kerman') kernels display conserved patterns of seed development and unique metabolite fluctuations. (a) Summary of selected significantly enriched ($P_{adj} < 0.05$ determined by Fisher's exact test) gene functions (*e.g.* Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes, iTAK) in modules highly correlated to relevant kernel traits (all enrichments can be found in Supporting Information Table S13). The sum of gene expression (i.e. normalized reads) at each time point in a given module is shown for enriched ($P_{adj} < 0.05$) functions. Metabolite profiles to confirm the gene expression trends across kernel development were obtained for (b) carbohydrates, starch, and proteins; (c) fatty acids, including monounsaturated fatty acid, polyunsaturated fatty acid, and saturated fatty acid; (d) monoterpene volatiles relevant to flavor; (e) phenolic compounds contributing to flavor and nutrition. Phenolics are reported as mg g^{-1} dry weight and as a percentage of the total amount. Metabolites are reported as averages across kernel samples ($n = 3$–4). The time points of kernel development correspond to the growing degree days (GDD, in °C) at collection.

modules K-IV-1 and K-IV-2 (Fig. 4a). Abscisic acid is a key regulator of seed maturation leading to seed dormancy (Sano & Marion-Poll, 2021). We validated that the concentration of ABA was higher in the kernel (2375.578 mg g^{-1}) relative to the maternal tissues (681.019 mg g^{-1} in the hull and 56.474 mg g^{-1} in the shell) at 1881 GDD when the nut is transitioning to Stage IV

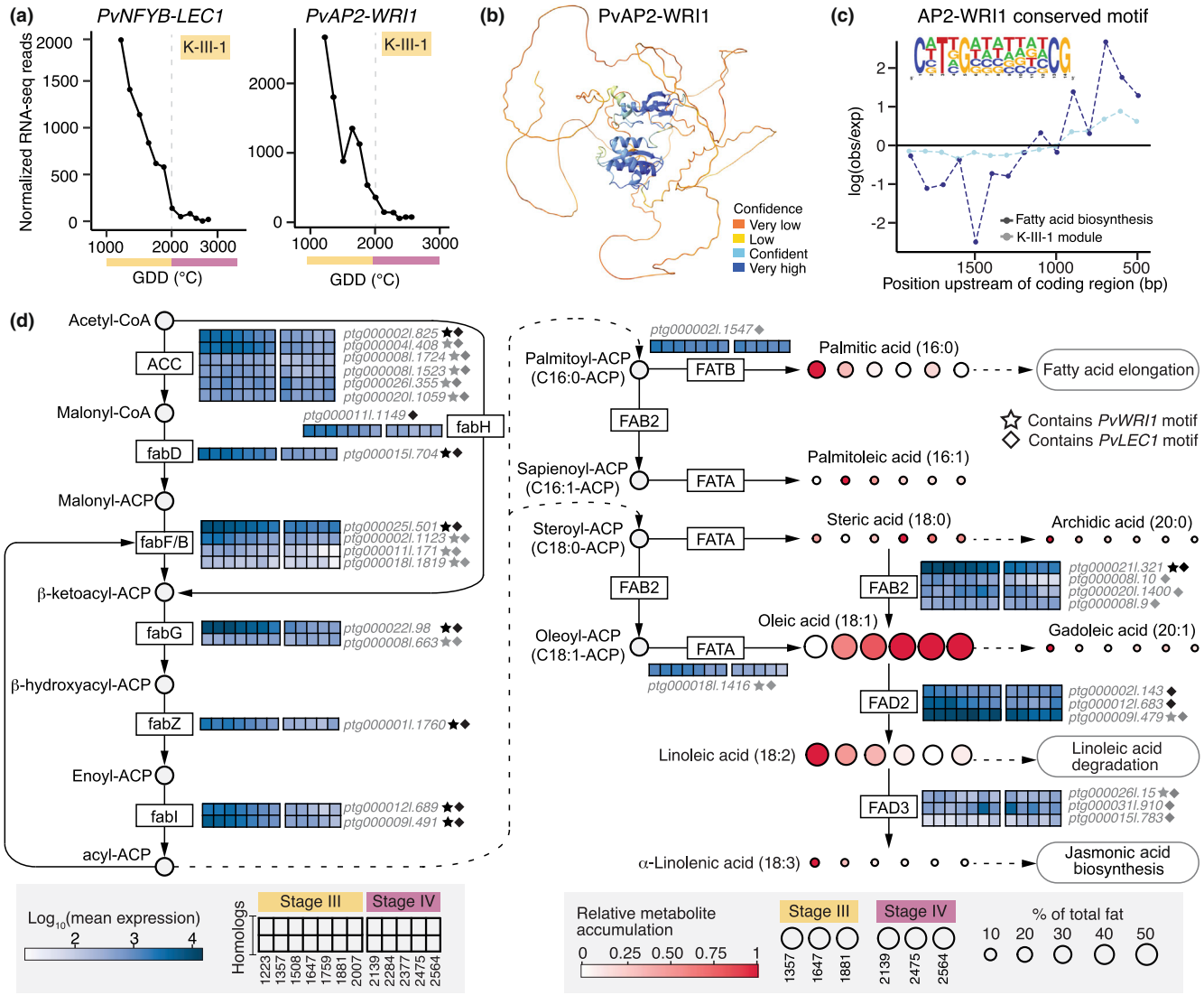


Fig. 5 Fatty acid biosynthesis and potential regulators in pistachio (*Pistacia vera* 'Kerman') kernel development. (a) Expression pattern (in normalized reads) of transcription factors *LEAFY-COTYLEDON1* (*PvNFYB-LEC1*; *PvKer.12.g280410*) and *WRINKLED1* (*PvAP2-WRI1*; *PvKer.03.g083330*) involved in seed development and fatty acid accumulation across timepoints represented in growing degree days (GDD, in °C). The corresponding stages are colored below in yellow (Stage III) and pink (Stage IV). The module each gene is a member of is indicated in each graph of (a). (b) Predicted 3D structure of *P. vera* 'Kerman' PvAP2-WRI1 protein. Colors indicate the folding confidence level as labeled on the right bottom. (c) Relative location of previously reported WRI1 AW-Box binding motifs found in 'Kerman' fatty acid biosynthesis (dark blue) and K-III-1 module genes (light blue) compared with other genes that contained the same binding motif. The consensus sequence of the AW-Box binding motif is shown as a logo. (d) A representation of fatty acid biosynthesis pathways based on Kyoto Encyclopedia of Genes and Genomes pathways (www.genome.jp/kegg/pathway.html, last accessed December 2023). Dashed lines indicate that some steps were omitted. The gene expression levels (i.e. $\log_{10}$ of the mean normalized read count for each sampling) are represented in colored boxes on a white-blue scale with each box representing a sampling point from Stage III and Stage IV determined GDD and each row representing homologous genes of a specific step in the pathway. Samplings for gene expression include dates from 1223 to 2564 GDD. Colored circles represent the accumulation of specific fatty acid compounds. Colors (white-red scale) indicate the relative accumulation of a specific metabolite at each time point, and the size of each circle represents the amount of the metabolite, measured in percentages of total fat (%). Six time points from Stage III (1357, 1647, and 1881 GDD) and Stage IV (2139, 2475, and 2564 GDD) were used for all fatty acid data shown. Fatty acid genes containing the WRI1 AW-box or LEC1 CCAAT target sequences are indicated by a star or diamond shape, respectively. ACA, acetyl-CoA carboxylase; FAB2, acyl-[acyl-carrier-protein] desaturase; fabD, S-malonyltransferase; fabF, 3-oxoacyl-[acyl-carrier-protein] synthase II; fabG, 3-oxoacyl-[acyl-carrier protein] reductase; fabH, 3-oxoacyl-[acyl-carrier-protein] synthase III; fabI, enoyl-[acyl-carrier protein] reductase I; fabZ, 3-hydroxyacyl-[acyl-carrier-protein] dehydratase; FAD2, omega-6 fatty acid desaturase; FAD3, acyl-lipid omega-3 desaturase; FATA, fatty acyl-ACP thioesterase A; FATB, fatty acyl-ACP thioesterase B.

(Table S15). Jasmonic acid biosynthesis, derived from fatty acid α-linoleic acid, was also enriched with increased expression during Stages III and IV (Fig. 4a). Homologs of the JA signaling TFs *JASMONATE-ZIM DOMAIN* (*JAZ*) and *MYC2* were highly expressed in module K-IV-1. Although we detected JA in the kernels (6 mg g^{-1} at 1881 GDD corresponding to late Stage III), the

hull and the shell had higher levels of this hormone at the same stage (1700 and 33 mg g^{-1}, respectively).

We mined TFs with high expression (top 10%) among the gene modules to identify potential regulators of kernel development. In particular, the module K-III-1 enriched in fatty acid biosynthesis and elongation (Fig. 4a) included the homolog of one of the four major seed master TFs, *LEAFY-COTYLEDON1* (*NFYB-LEC1*), known to co-regulate fatty acid biosynthesis across diverse crop species (Mu *et al.*, 2008). This module also contained the homolog of a key regulator of seed oil content, *WRINKLED-1* (*AP2-WRI1*; Kong *et al.*, 2019), which has been reported to be under the control of NFYB-LEC1. The *PvAP2-WRI1* gene in the 'Kerman' genome had a high degree of protein sequence similarity and conservation of DNA-binding domains across 16 representative angiosperm taxa (Sánchez *et al.*, 2022; Kuczynski *et al.*, 2022; Tan *et al.*, 2023; Figs 5b, S9, S10). Fatty acid biosynthesis genes in the K-III-1 module had the putative AW-box binding domain significantly closer to the TIS than the other genes in the genome that contained the same sequence, suggesting that PvAP2-WRI1 may regulate them, but this result will require experimental validation (Lis & Walther, 2016; Fig. 5c). The AW-box and the conserved binding motif (CCAAT) for the PvNFYB-LEC1 predicted protein were prevalent among genes involved in *de novo* fatty acid biosynthesis steps and desaturase enzymes associated with the synthesis of MUFA and PUFA (Fig. 5d).

To further understand the dynamics of fat accumulation in pistachio, we visualized the expression of fatty acid biosynthesis genes across kernel development and measured individual metabolites (Fig. 5d; Table S14). The highest expressed genes in the pathway were the *FAB2* homolog, an acyl-[acyl-carrier-protein] desaturase, and the FAD2 homolog, an omega-6 fatty acid desaturase. These genes represent key steps in unsaturated fatty acid biosynthesis and contain the WRI1 and LEC1 binding motifs (Mu *et al.*, 2008; Bonghi *et al.*, 2011; He *et al.*, 2020). The metabolite data showed that oleic acid, a MUFA, and linoleic acid, a PUFA, were the top two accumulated fatty acids in pistachio kernels. The relative accumulation of linoleic acid decreased over time while oleic acid increased, as was previously observed (Polari *et al.*, 2019), and may be due to fatty acid degradation happening in late kernel development (K-III-3 and K-IV-1; Fig. 4a).

Discussion

This comprehensive study positions pistachio to serve as a model system to investigate the biology of hard-shelled fruits and the asynchronous behavior of seed and fruit development observed in some tree species (Bonghi *et al.*, 2011). The 'Kerman' genome and annotation presented here provide a foundational resource for pistachio breeding and set the stage for further molecular studies. This genome enables the characterization of the notable genomic features discovered, particularly large knob-like satellite repeats found on multiple chromosomes, which do not contain any protein-coding genes. It is important to note, however, that the thorough analysis of these regions, especially focusing on variations between haplotypes and within the species, remains in its preliminary stages. More in-depth, haplotype-level investigations are required to fully understand their effects on genomic structure and function, which will enhance our knowledge of their evolutionary and practical significance in crop improvement. Nonetheless, leveraging this genome, we were able to map molecular processes underlying nut development at high spatiotemporal resolution.

Pistachio kernels do not grow until *c.* 1000 GDD after maternal tissue growth starts (Fig. 1d–j). Energy limitations in the tree may explain why this pattern has evolved in some tree species. Carbohydrates reserved from the previous year are exhausted after developing the hull, shell, and leaves in Stage I (Tixier *et al.*, 2020). The lull in fruit growth identified as Stage II may function to generate resources from photosynthesis that can be used to support kernel growth in Stage III (Marino *et al.*, 2022). This growth pattern may also help explain why some pistachio nuts continue their development without a kernel (i.e. 'blanks') and why the trees drop their buds in early summer (i.e. bud abscission), leading to alternate bearing years (Benny *et al.*, 2020, 2021). Transcriptome and hormone analysis at the transition to seed growth (Stages II to III) will be essential to further elucidate these phenomena.

Transcriptomic analysis during kernel growth and maturation highlighted that pistachio has conserved regulatory mechanisms during seed development consistent with other plant species (Fig. 4). Paralogs of the master TFs critical to seed development known as LAFL (LEC1, ABI2, FUS3, and LEC2) were highly expressed and followed conserved expression patterns during pistachio kernel development (Table S6). Furthermore, we identified homologs of LEC1 and WRI1, known to act together to regulate seed oil content in other plant species (Mu *et al.*, 2008; He *et al.*, 2020), to be candidate regulators of fatty acid biosynthesis in pistachio (Zhai *et al.*, 2018). Benny *et al.* (2020, 2021) found *WRI1* and coregulators *SUCROSE-NON-FERMENTING1-RELATED PROTEIN KINASE1* (*SnRK1*) and *trehalose 6-phosphate* (*T6P*) were co-expressed in early kernel development of 'Bianca' pistachios. Likewise, *WRI1*, *T6P*, and *SnRK1* were in the K-III-1 module associated with increased fat content and should be further explored to define the regulatory network involved.

The knowledge and genomic resources provided here will facilitate molecular-assisted breeding and the study of the biological causes of pistachio quality defects, such as internal kernel discoloration, low shell split, and hull deterioration. Furthermore, our research can be readily applied to pistachio production. The newly defined Stage II, a crucial period of reduced nut growth before kernel initiation, has significant implications for management practices, such as the timing of deficit irrigation (Goldhamer & Beede, 2004). We discovered that shell hardening and kernel growth occur simultaneously, providing a nondestructive benchmark to track kernel development, corroborating an earlier report by Zhang *et al.* (2021). The hull ripening traits observed during Stage IV can give nondestructive indications of the best harvest time. Kernels reach their maximum ratios of macronutrients at the start of Stage IV, while changes to volatiles and phenolics occur later; this should be considered to achieve the best flavor and nutrient content at harvest. Overall, as the demand for pistachios continues to increase, our work provides multiple

avenues for research that can benefit breeders, growers, and consumers.

Acknowledgements

The authors would like to acknowledge and thank Caio Cattai de Andrade for sampling pistachios from the 2019 growing season; Adrian O. Sbodio for sampling nuts from the 2020 season and processing the nut samples in 2020-2021; Jose G. Barquero-Jackson for sampling nuts from the 2021 growing season; Paula Guzman Delgado for measuring starch and carbohydrates in pistachio kernels; Pedro Bello for technical support in capturing images from Fig. 1(d); and collaborators Joseph Coelho and Ian Humrick (Maricopa Orchards), and Joey Thomas and John Thomas (Dewey Farms) for providing access to their pistachio orchards for sampling. This research was funded by the California Pistachio Research Board (grant nos.: HC-2019-15-0, HP-2020-28, and HP-2021-37-0 to BB-U, GM, and SCW; grant no.: HG-2022-35 to GM), USDA-NIFA (grant no.: 108681-Z5327202 to GM and PJB), and the Foundation for Food and Agriculture Research (Rockey Fellowship to MD). Support for sequencing was provided by PacBio (SMART grant). The Department of Plant Sciences, UC Davis, funded by endowments, particularly the James Monroe McDonald Endowment, administered by UCANR, supported JAA and SDMP. Additional funding for SDMP was obtained from the 'La Caixa' Foundation (ID 100010434) under the agreement no.: LCF/BQ/AA19/11720034.

Competing interests

None declared.

Author contributions

JAA, CL, GD, PJB, AMM, AM, AG, EMB, FPM, LMC, LC, PB, PC-B, JGM and BB-U contributed to the conceptualization. JAA, CL, FSG, SDM-P, MD, SCW, GM, LF, JGM and BB-U contributed to the methodology. JAA, CL, YW, FW, FSG, MD, JGM and BB-U contributed to the formal analysis. AMM, AM, AG, EMB, FPM, LMC, LC, PB, PC-B, JGM, BB-U, SW and GM contributed to the resources. JAA, YW, SDM-P and FSG contributed to the investigation. JAA, CL, YW, MD and JGM contributed to the data curation. JAA, CL, JGM and BB-U contributed to the writing – original draft. JAA, CL, YW, MD, JGM and BB-U contributed to the visualization. JAA, CL, PCB, PJB, JGM and BB-U contributed to the supervision. PC-B, JGM and BB-U contributed to the project administration. GM, SCW, AMM, AM, AG, EMB, FPM, LMC, LC, PB, PC-B, JGM, PJB and BB-U contributed to the funding acquisition. All authors contributed to the writing – review and editing. JAA and CL are contributed equally to this work.

ORCID

Jaclyn A. Adaskaveg https://orcid.org/0000-0002-3408-5984
Paolo Bagnaresi https://orcid.org/0000-0001-8466-4805
Barbara Blanco-Ulate https://orcid.org/0000-0002-8819-9207
Patrick J. Brown https://orcid.org/0000-0003-1332-711X
Esaú Martínez Burgos https://orcid.org/0000-0002-5338-5237
Pablo Carbonell-Bejerano https://orcid.org/0000-0002-7266-9665
Luigi Cattivelli https://orcid.org/0000-0002-6067-4600
Lourdes Marchante Cuevas https://orcid.org/0000-0002-0458-8783
Matthew Davis https://orcid.org/0009-0004-8933-349X
Georgia Drakakaki https://orcid.org/0000-0002-3949-8657
Louise Ferguson https://orcid.org/0000-0002-9520-078X
Antonio Giovino https://orcid.org/0000-0001-5501-0204
Filipa S. Grilo https://orcid.org/0000-0003-3535-4776
Chaehee Lee https://orcid.org/0000-0003-3214-2997
Annalisa Marchese https://orcid.org/0000-0002-6816-6184
Giulia Marino https://orcid.org/0000-0002-2577-1974
Francesco Paolo Marra https://orcid.org/0000-0003-1490-0619
Saskia D. Mesquida-Pesci https://orcid.org/0000-0002-4182-5039
J. Grey Monroe https://orcid.org/0000-0002-4025-5572
Adela Mena Morales https://orcid.org/0000-0001-6342-209X
Fangyi Wang https://orcid.org/0000-0003-2687-9909
Selina C. Wang https://orcid.org/0000-0002-9030-837X
Yiduo Wei https://orcid.org/0009-0005-3643-5086

Data availability

All data included in this study can be found under the BioProject accession no.: PRJNA1114109. Under this, all the raw genome sequencing data have been deposited at NCBI (https://www.ncbi.nlm.nih.gov/), with the BioProject accession no.: PRJNA1049825 (PacBio HiFi, Omni-C, and Iso-Seq reads). The genome assembly, annotation, and 3D structure of gene models are available at the Pistachiomics database hosted by UC Davis (https://pistachiomics.sf.ucdavis.edu/). The nut transcriptomic data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus[94] with the BioProject accession no.: PRJNA1110275 and are accessible through GEO Series accession no.: GSE267225 (https://www.ncbi.nlm.nih.gov/geo/).

References

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology* **23**: 258.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models Usinglme4. *Journal of Statistical Software* **67**: 1–48.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **57**: 289–300.

Benny J, Giovino A, Marra FP, Balan B, Martinelli F, Caruso T, Marchese A. 2021. Transcriptomic analysis of the *Pistacia vera* (L.) fruits enable the identification of genes and hormone-related gene linked to inflorescence bud abscission. *Genes* **13**: 60.

Benny J, Marra FP, Giovino A, Balan B, Caruso T, Martinelli F, Marchese A. 2020. Transcriptome analysis of *Pistacia vera* inflorescence buds in bearing and non-bearing shoots reveals the molecular mechanism causing premature flower bud abscission. *Genes* **11**: 851.

Bento M, Tomás D, Viegas W, Silva M. 2013. Retrotransposons represent the most labile fraction for genomic rearrangements in polyploid plant species. *Cytogenetic and Genome Research* **140**: 286–294.

Blanco-Ulate B, Vincenti E, Powell ALT, Cantu D. 2013. Tomato transcriptome and mutant analyses suggest a role for plant stress hormones in the interaction between fruit and *Botrytis cinerea*. *Frontiers in Plant Science* **4**: 142.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bonghi C, Trainotti L, Botton A, Tadiello A, Rasori A, Ziliotto F, Zaffalon V, Casadoro G, Ramina A. 2011. A microarray approach to identify genes involved in seed-pericarp cross-talk and development in peach. *BMC Plant Biology* **11**: 107.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics* **3**: lqaa108.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**: 170–175.

Corelli-Grappadelli L, Lakso AN. 2004. Fruit development in deciduous tree crops as affected by physiological factors and environmental conditions (keynote). *Acta Horticulturae* **636**: 425–441.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Research* **14**: 1188–1190.

Dainat J, Hereñú D, Davis E, Crouch K, Sol L, Agostinho N. 2022. Another GFF analysis toolkit to handle annotations in any GTF/GFF format (v.1.0). *Zenodo*. doi: 10.5281/zenodo.3552717.

Derbyshire E, Higgs J, Feeney MJ, Carughi A. 2023. Believe it or 'nut': why it is time to set the record straight on nut protein quality: pistachio (*Pistacia vera* L.) focus. *Nutrients* **15**: 2158.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP *et al.* 2017. *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**(6333): 92–95.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3**: 95–98.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**: 238.

Erşan S, Güçlü Üstündağ Ö, Carle R, Schweiggert RM. 2017. Determination of pistachio (*Pistacia vera* L.) hull (exo- and mesocarp) phenolics by HPLC-DAD-ESI/MSn and UHPLC-DAD-ELSD after ultrasound-assisted extraction. *Journal of Food Composition and Analysis* **62**: 103–114.

Ferguson L, Polito V, Kallsen C. 2005. *The pistachio tree; botany and physiology and factors that affect yield. Pistachio production manual, 4th edn.* Davis, CA, USA: University of California Fruit & Nut Research Information Center, 31–39.

Gentile C, Tesoriere L, Butera D, Fazzari M, Monastero M, Allegra M, Livrea MA. 2007. Antioxidant activity of Sicilian pistachio (*Pistacia vera* L. var. Bronte) nut extract and its bioactive components. *Journal of Agricultural and Food Chemistry* **55**: 643–648.

Goldhamer D, Beede R. 2004. Regulated deficit irrigation effects on yield, nut quality and water-use efficiency of mature pistachio trees. *The Journal of Horticultural Science & Biotechnology* **79**: 538–545.

Grilo FS, Wang SC. 2021. Walnut (*Juglans regia* L.) volatile compounds indicate kernel and oil oxidation. *Food* **10**: 329.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**: R7.

Hao Z, Lv D, Ge Y, Shi J, Weijers D, Yu G, Chen J. 2020. RIDEOGRAM: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ* **6**: e251.

Hawkins JS, Proulx SR, Rapp RA, Wendel JF. 2009. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proceedings of the National Academy of Sciences, USA* **106**: 17811–17816.

He M, Qin C-X, Wang X, Ding N-Z. 2020. Plant unsaturated fatty acids: biosynthesis and regulation. *Frontiers in Plant Science* **11**: 390.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A *et al.* 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**(7873): 583–589.

Kafkas S, Ma X, Zhang X, Topçu H, Navajas-Pérez R, Wai CM, Tang H, Xu X, Khodaeiaminjan M, Güney M *et al.* 2023. Pistachio genomes provide insights into nut tree domestication and ZW sex chromosome evolution. *Plant Communications* **4**: 100497.

Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z *et al.* 2021. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research* **49**: D192–D200.

Katoh K, Standley DM. 2013. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Keller O, Kollmar M, Stanke M, Waack S. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**: 757–763.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**: 907–915.

Kong Q, Yuan L, Ma W. 2019. WRINKLED1, a "master regulator" in transcriptional control of plant oil biosynthesis. *Plants* **8**: 238.

Kuczynski C, McCorkle S, Keereetaweep J, Shanklin J, Schwender J. 2022. An expanded role for the transcription factor WRINKLED1 in the biosynthesis of triacylglycerols during seed development. *Frontiers in Plant Science* **13**: 955589.

Kuznetsova A, Brockhoff PB, Christensen RHB. 2017. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software* **82**: 1–26.

Langfelder P, Horvath S. 2008. Wgcna: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.

Leyva A, Quintana A, Sánchez M, Rodríguez EN, Cremata J, Sánchez JC. 2008. Rapid and sensitive anthrone-sulfuric acid assay in microplate format to quantify carbohydrate in biopharmaceutical products: method development and validation. *Biologicals: Journal of the International Association of Biological Standardization* **36**: 134–141.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Lin T-S, Polito VS, Crane JC. 1984. Embryo development in 'Kerman' pistachio. *HortScience* **19**: 105–106.

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y *et al.* 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**: 1123–1130.

Lis M, Walther D. 2016. The orientation of transcription factor binding site motifs in gene promoter regions: does it matter? *BMC Genomics* **17**: 185.

Liu Y, Blumberg JB, Chen C-YO. 2014. Quantification and bioaccessibility of california pistachio bioactives. *Journal of Agricultural and Food Chemistry* **62**: 1550–1556.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**: 550.

Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, Emms D, Goodstein DM, Schmutz J. 2022. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife* **11**: e78526.

Lowe TM, Chan PP. 2016. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Research* **44**: W54–W57.

Maeo K, Tokuda T, Ayame A, Mitsui N, Kawai T, Tsukagoshi H, Ishiguro S, Nakamura K. 2009. An AP2-type transcription factor, WRINKLED1, of *Arabidopsis thaliana* binds to the AW-box sequence conserved among proximal upstream regions of genes involved in fatty acid synthesis. *The Plant Journal: For Cell and Molecular Biology* **60**: 476–487.

Mandalari G, Barreca D, Gervasi T, Roussell MA, Klein B, Feeney MJ, Carughi A. 2021. Pistachio nuts (*Pistacia vera* L.): production, nutrients, bioactives and novel health effects. *Plants* **11**: 18.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764–770.

Marino G, Caruso T, Ferguson L, Marra F. 2018. Gas exchanges and stem water potential define stress thresholds for efficient irrigation management in Olive (*Olea europea* L.). *Watermark* **10**: 342.

Marino G, Guzmán-Delgado P, Caruso T, Marra FP. 2022. Modeling seasonal branch carbon dynamics in pistachio as a function of crop load. *Scientia Horticulturae* **296**: 110875.

Marvinney E, Kendall A, Brodt S, Schenck R, Huizen D. 2014. A comparative assessment of greenhouse gas emissions in California almond, pistachio, and walnut production. *Environmental Science, Agricultural and Food Sciences* **20**: 761–771.

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research* **32**: W20–W25.

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. *Nature Methods* **19**: 679–682.

Moazzam Jazi M, Ghadirzadeh Khorzoghi E, Botanga C, Seyedi SM. 2016. Identification of reference genes for quantitative gene expression studies in a non-model tree pistachio (*Pistacia vera* L.). *PLoS ONE* **11**: e0157467.

Morales-Cruz A, Aguirre-Liguori JA, Zhou Y, Minio A, Riaz S, Walker AM, Cantu D, Gaut BS. 2021. Introgression among North American wild grapes (*Vitis*) fuels biotic and abiotic adaptation. *Genome Biology* **22**: 254.

Mu J, Tan H, Zheng Q, Fu F, Liang Y, Zhang J, Yang X, Wang T, Chong K, Wang X-J *et al.* 2008. LEAFY COTYLEDON1 is a key regulator of fatty acid biosynthesis in Arabidopsis. *Plant Physiology* **148**: 1042–1054.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**: 2933–2935.

Noguera-Artiaga L, García-Romo JS, Rosas-Burgos EC, Cinco-Moroyoqui FJ, Vidal-Quintanar RL, Carbonell-Barrachina ÁA, Burgos-Hernández A. 2019. Antioxidant, antimutagenic and cytoprotective properties of hydroSOS pistachio nuts. *Molecules* **24**: 4362.

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T *et al.* 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology* **20**: 275.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**: 290–295.

Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. 2021. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Science* **30**: 70–82.

Polari JJ, Ferguson L, Wang SC. 2020. Pistachio kernel composition of 'Kalehghouchi', 'Pete 1', and 'Lost Hills' in California. *HortScience* **55**: 666–669.

Polari JJ, Zhang L, Ferguson L, Maness NO, Wang SC. 2019. Impact of microclimate on fatty acids and volatile terpenes in "Kerman" and "Golden Hills" pistachio (*Pistacia vera*) kernels. *Journal of Food Science* **84**: 1937–1942.

Polito VS, Pinney K. 1999. Endocarp dehiscence in pistachio (*Pistacia vera* L.). *International Journal of Plant Sciences* **160**: 827–835.

Poore J, Nemecek T. 2018. Reducing food's environmental impacts through producers and consumers. *Science* **360**: 987–992.

Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. 2018. Juicebox.Js provides a cloud-based visualization system for Hi-C data. *Cell Systems* **6**: 256–258.

Sánchez R, González-Thuillier I, Venegas-Calerón M, Garcés R, Salas JJ, Martínez-Force E. 2022. The sunflower WRINKLED1 transcription factor regulates fatty acid biosynthesis genes through an AW box binding sequence with a particular base bias. *Plants* **11**: 972.

Sano N, Marion-Poll A. 2021. ABA metabolism and homeostasis in seed dormancy and germination. *International Journal of Molecular Sciences* **22**: 5069.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.

Sola-Campoy PJ, Robles F, Schwarzacher T, Ruiz Rejón C, de la Herrán R, Navajas-Pérez R. 2015. The molecular cytogenetic characterization of pistachio (*Pistacia vera* L.) suggests the arrest of recombination in the largest heteropycnotic pair HC1. *PLoS ONE* **10**: e0143861.

Tan Q, Han B, Haque ME, Li Y-L, Wang Y, Wu D, Wu S-B, Liu A-Z. 2023. The molecular mechanism of WRINKLED1 transcription factor regulating oil accumulation in developing seeds of castor bean. *Plant Diversity* **45**: 469–478.

Tixier A, Guzmán-Delgado P, Sperling O, Amico Roxas A, Laca E, Zwieniecki MA. 2020. Comparison of phenological traits, growth patterns, and seasonal dynamics of non-structural carbohydrate in Mediterranean tree crop species. *Scientific Reports* **10**: 347.

Tomaino A, Martorana M, Arcoraci T, Monteleone D, Giovinazzo C, Saija A. 2010. Antioxidant activity and phenolic profile of pistachio (*Pistacia vera* L., variety Bronte) seeds and skins. *Biochimie* **92**: 1115–1122.

Tsantili E, Konstantinidis K, Christopoulos MV, Roussos PA. 2011. Total phenolics and flavonoids and total antioxidant capacity in pistachio (*Pistacia vera* L.) nuts in relation to cultivars and storage conditions. *Scientia Horticulturae* **129**: 694–701.

Venables WN, Ripley BD. 2003. *Modern applied statistics with S*. New York, NY, USA: Springer.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**: 2202–2204.

Wang P, Luo Y, Huang J, Gao S, Zhu G, Dang Z, Gai J, Yang M, Zhu M, Zhang H *et al.* 2020. The genome evolution and domestication of tropical fruit mango. *Genome Biology* **21**: 60.

Workman R, Timp W, Fedak R, Kilburn D, Hao S, Liu KJ. 2018. High molecular weight DNA extraction from recalcitrant plant species for third generation sequencing. *Nature Protocol Exchange* **2**: 2402.

Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, Perrier X, Ruiz M, Scalabrin S, Terol J *et al.* 2014. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nature Biotechnology* **32**(7): 656–662.

Zeng L, Tu X-L, Dai H, Han F-M, Lu B-S, Wang M-S, Nanaei HA, Tajabadipour A, Mansouri M, Li X-L *et al.* 2019. Whole genomes and transcriptomes reveal adaptation and domestication of pistachio. *Genome Biology* **20**: 79.

Zhai Z, Keereetaweep J, Liu H, Feil R, Lunn JE, Shanklin J. 2018. Trehalose 6-phosphate positively regulates fatty acid synthesis by stabilizing WRINKLED1. *Plant Cell* **30**: 2616–2627.

Zhang L, Laca E, Allan CJ, Mahvelati NM, Ferguson L. 2021. Nonlinear model selection for fruit and kernel development as a function of heat in pistachio. *HortScience* **56**: 769–779.

Zwieniecki MA, Davidson AM, Orozco J, Cooper KB, Guzman-De. lgado P. 2022. The impact of non-structural carbohydrates (NSC) concentration on yield in *Prunus dulcis*, *Pistacia vera*, and *Juglans regia*. *Scientific Reports* **12**: 4360.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Workflow for PacBio HiFi sequencing, transcript discovery, and *de novo* genome assembly and scaffolding.

Fig. S2 Genome annotation pipeline using both RNA- and Iso-seq data as extrinsic hints.

Fig. S3 Pistachio (*Pistacia vera* 'Kerman') nut physiological patterns are consistent across years and locations.

Fig. S4 Estimation of *Pistacia vera* 'Kerman' genome size, heterozygosity, and repetitiveness with *k*-mer frequency using jellyfish and GenomeScope.

Fig. S5 Repeat analysis in *Pistacia vera* 'Kerman' genome assembly.

Fig. S6 Macrosynteny comparison between genome assemblies of *Pistacia vera* 'Kerman' and three other cultivars ('Batoury', 'Siirt', and 'Bagyolu').

Fig. S7 The number of unique and overlapping transcripts expressed in five different tissue types.

Fig. S8 Weighted gene co-expression network analysis (WGCNA) across all time points and tissue types.

Fig. S9 Comparison of *Pistacia vera* WRINKLED1 (WRI1) protein sequences with 15 representative angiosperm species.

Fig. S10 Comparison of *Pistacia vera* LEAFY COTYLEDON1 (LEC1) protein sequences with 15 representative angiosperm species.

Table S1 Summary of pistachio (*Pistacia vera* 'Kerman') physiological data collected across different locations and harvest years (2019, 2020, 2021).

Table S2 Summary of pistachio (*Pistacia vera* 'Kerman') nut physiological growth models.

Table S3 Statistics of PacBio HiFi sequencing and Dovetail genomics Omni-C data for pistachio (*Pistacia vera* 'Kerman').

Table S4 The statistics of noncoding RNA in *Pistacia vera* 'Kerman' genome.

Table S5 RNA-sequencing read mapping summary for pistachio (*Pistacia vera* 'Kerman') hull, shell, and kernel tissue collected at each time point (growing degree days, GDD) in each stage.

Table S6 *Pistacia vera* 'Kerman' transcriptome functional annotations and gene expression data.

Table S7 List of angiosperm taxa used in comparative genomic analysis.

Table S8 Statistics of genome assembly and annotation of the *Pistacia vera* 'Kerman'.

Table S9 The summary of BUSCO (Benchmarking Universal Single-Copy Orthologs) assessment of *Pistacia vera* 'Kerman' primary contig assembly and 15 chromosomes, and three other pistachio genome assemblies.

Table S10 Summary statistics of repeat content in *Pistacia vera* 'Kerman' primary contig assembly.

Table S11 The summary of BUSCO (Benchmarking Universal Single-Copy Orthologs) assessment of gene annotation of *Pistacia vera* 'Kerman' primary contig assembly and 15 chromosomes.

Table S12 Statistics of the number of transcripts in different tissue types of *Pistacia vera* 'Kerman'.

Table S13 Enrichments of functional terms for *Pistacia vera* 'Kerman'.

Table S14 Summary of metabolite data for *Pistacia vera* 'Kerman'.

Table S15 Hormone measurements in *Pistacia vera* 'Kerman' nut tissues at the end of Stage III.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

Disclaimer: The New Phytologist Foundation remains neutral with regard to jurisdictional claims in maps and in any institutional affiliations.