Journal of Heredity, 2024, **115**, 221–229
https://doi.org/10.1093/jhered/esae002
Advance access publication 2 February 2024
Genome Resources

American Genetic Association

OXFORD

Genome Resources

The genome assembly of Island Oak (*Quercus tomentella*), a relictual island tree species

Alayna Mead[1,2], Sorel T. Fitz-Gibbon[1], Merly Escalona[3], Eric Beraut[4], Samuel Sacco[4], Mohan P.A. Marimuthu[5], Oanh Nguyen[5] and Victoria L. Sork[1,6,*]

[1]Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles CA 90095-7239, United States,
[2]Present address: Department of Ecosystem Science and Management, Pennsylvania State University, State College PA 16803, United States.
[3]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, United States,
[4]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, United States,
[5]DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California, Davis, CA 95616, United States,
[6]Institute of the Environment and Sustainability, University of California Los Angeles, Los Angeles CA 90095, United States

*Corresponding author: Email: vlsork@ucla.edu

Corresponding Editor: Rachel Meyer

Abstract

Island oak (*Quercus tomentella*) is a rare relictual island tree species that exists only on six islands off the coast of California and Mexico, but was once widespread throughout mainland California. Currently, this species is endangered by threats such as non-native plants, grazing animals, and human removal. Efforts for conservation and restoration of island oak currently underway could benefit from information about its range-wide genetic structure and evolutionary history. Here we present a high-quality genome assembly for *Q. tomentella*, assembled using PacBio HiFi and Omni-C sequencing, developed as part of the California Conservation Genomics Project (CCGP). The resulting assembly has a length of 781 Mb, with a contig N50 of 22.0 Mb and a scaffold N50 of 63.4 Mb. This genome assembly will provide a resource for genomics-informed conservation of this rare oak species. Additionally, this reference genome will be the first one available for a species in *Quercus* section *Protobalanus*, a unique oak clade present only in western North America.

Key words: California Channel Islands, California Conservation Genomics Project, island oak, *Quercus tomentella*

Introduction

Island oak (*Quercus tomentella*) is a rare oak that is currently found only on six islands: five of the Channel Islands off the coast of California (Santa Rosa, Santa Cruz, Anacapa, Catalina, and San Clemente) and Guadalupe Island in Baja California, Mexico; however, it was once distributed across mainland California (Axelrod 1939, 1944a, 1944b). Island oak forms oak woodlands and provides habitat for many animal species (Fig. 1), such as the Island Scrub Jay, a rare island endemic (Sawyer et al. 2009; Pesendorfer et al. 2018). These island ecosystems have undergone major changes since settlement by the Spanish throughout the 1800s, including the introduction of non-native plants and herbivores, and as a result island oak is categorized as endangered by the IUCN Red List (Beckman and Jerome 2017; Beckman et al. 2019). In fact, this species may occupy only a small proportion of the potentially suitable habitat on the islands (Kindsvater 2006). Conservation and management of this species would benefit from understanding the genetic structure of populations within and across islands using genomic analyses.

As part of the California Conservation Genomics Project (CCGP; Shaffer et al. 2022), we sequenced and assembled a genome for *Q. tomentella*. This resource will facilitate evaluations of neutral and adaptive genetic diversity across the range of island oak to inform optimal management strategies. Previous genetic work in this species using microsatellite markers suggests that genetic diversity is not low, despite high levels of clonality on some islands (Ashley et al. 2018). Additionally, populations were genetically differentiated across islands, particularly the most isolated population on Guadalupe Island (Ashley et al. 2018). However, it is unclear whether this differentiation is due solely to isolation by distance, or also includes genetic variation involved in adaptation to the differing climate conditions on each island (i.e. isolation by environment). This genome assembly will be valuable for future studies using whole-genome sequences to determine whether populations on each island differ in adaptive genetic variation. It will also be helpful in understanding the influence of a sister species, *Quercus chrysolepis*, which is widespread across California and introgresses with island oak throughout the California Channel Islands (Mead and Sork, unpublished data). We will use this information to determine ideal seed sources for the replanting of *Q. tomentella* groves that are part of ongoing restoration projects on several

Received November 20, 2023; Accepted February 1, 2024

islands. Additionally, this genome assembly is the first for *Quercus* section *Protobablanus*, a small clade restricted to western North America (Nixon 2002; Ortego et al. 2018). The genome assembly described here expands the taxonomic coverage of currently available oak genomic resources, which already include several reference genomes (Bodénès et al. 2016; Plomion et al. 2018; Ramos et al. 2018; Ai et al. 2022; Maldonado-Alconada et al. 2022; Sork et al. 2022; Kapoor et al. 2023; O'Donnell et al. unpublished data), expanding opportunities for further evolutionary studies using oaks as a model clade (Cavender-Bares 2018).

Methods

Biological materials

Young leaf tissue was collected from an adult individual planted at the California Botanic Garden, Claremont, California (coordinates: 34.11613, –117.71630). The accession number in the California Consortium of Herbaria for the planted tree is RSA801373 (https://ucjeps.berkeley.edu/consortium/details.php?aid=RSA801373) and the California Botanic Garden accession number is 12616. The seed for the planted tree was originally collected on Catalina Island by D. Probst on 13 October 1966; the recorded locality was Toyon Canyon junction (33.3744, –118.35377).

Nucleic acid library preparation

High molecular weight (HMW) genomic DNA (gDNA) was extracted from young leaves and catkins (1.6 g) using the Nanobind Plant Nuclei Big DNA Kit (Pacific BioSciences—PacBio, Menlo Park, CA) and Workman et al. (2018) protocol with the following modification. We used nuclear isolation buffer supplemented with 350 mM Sorbitol to resuspend ground tissue and during the first wash of the nuclei pellet. The extracted HMW DNA was further purified using the high-salt-phenol–chloroform method (PacBio). The DNA purity was estimated using absorbance ratios (260/280 = 1.78 and 260/230 = 2.14) on the NanoDrop ND-1000 spectrophotometer. The final DNA yield (15.6 µg) was quantified using the Quantus Fluorometer (QuantiFluor ONE dsDNA Dye assay; Promega, Madison, WI). The size distribution of the HMW DNA was estimated using the Femto Pulse system (Agilent, Santa Clara, CA), and found that 52% of the fragments were 80 Kb or longer.

HiFi library preparation and sequencing

The HiFi SMRTbell library was constructed using the SMRTbell Express Template Prep Kit v2.0 (PacBio, Cat. #100-938-900) according to the manufacturer's instructions. HMW gDNA was sheared to a target DNA size distribution between 15 and 18 kb. The sheared gDNA was concentrated using 0.45× of AMPure PB beads (PacBio, Cat. #100-265-900) for the removal of single-strand overhangs at 37 °C for 15 min, followed by further enzymatic steps of DNA damage repair at 37 °C for 30 min, end repair and A-tailing at 20 °C for 10 min and 65 °C for 30 min, and ligation of overhang adapter v3 at 20 °C for 60 min. The SMRTbell library was purified and concentrated with 1× Ampure PB beads (PacBio, Cat. #100-265-900) for nuclease treatment at 37 °C for 30 min and size selection using the BluePippin/PippinHT system (Sage Science, Beverly, MA; Cat #BLF7510/HPE7510) to collect fragments greater than 7 to 9 kb. The 15 to 20 kb average HiFi SMRTbell library was sequenced at UC Davis DNA Technologies Core (Davis, CA) using one 8M SMRT cell, Sequel II sequencing chemistry 2.0, and 30-h movies each on a PacBio Sequel II sequencer.

Omni-C library preparation and sequencing

The Omni-C library was prepared using the DovetailTM Omni-CTM Kit (Dovetail Genomics, Scorrs Valley, CA) according to the manufacturer's protocol with slight modifications. First, specimen tissue (Sample ID: Qtom.A.LA.223) was thoroughly ground with a mortar and pestle while cooled with liquid nitrogen. Nuclear isolation was then performed using published methods (Workman et al. 2018). Subsequently, chromatin was fixed in place in the nucleus. Fixed chromatin was digested under various conditions of DNase I until a suitable fragment length distribution of DNA molecules was obtained. Chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter containing ends. After proximity ligation, crosslinks were reversed, and the DNA was purified from proteins. Purified DNA was treated to remove biotin that was not internal to ligated fragments. An NGS library was generated using an NEB Ultra II DNA Library Prep kit (New England Biolabs—NEB, Ipswich, MA) with an Illumina compatible y-adaptor. Biotin-containing fragments were then captured using streptavidin beads. The post-capture product was split into two replicates prior to PCR enrichment to preserve library complexity with each replicate receiving unique dual indices. The library was sequenced at Vincent J. Coates Genomics Sequencing Lab (Berkeley, CA) on an Illumina NovaSeq 6000 platform (Illumina, San Diego, CA) to generate approximately 100 million 2 × 150 bp read pairs per GB of genome size.

Nuclear genome assembly

We assembled the genome of the *Q. tomentella* following the CCGP assembly pipeline Version 5.0, as outlined in Table 1, which lists the tools and non-default parameters used in the assembly. The pipeline uses PacBio HiFi reads and Omni-C data to produce high quality and highly contiguous genome assemblies. First, we removed the remnants adapter sequences from the PacBio HiFi dataset using HiFiAdapterFilt (Sim et al. 2022) and generated the initial phased diploid assembly using HiFiasm (Cheng et al. 2021) on Hi-C mode, with the filtered PacBio HiFi reads and the Omni-C dataset. We then aligned the Omni-C data to both assemblies following the Arima Genomics Mapping Pipeline (https://github.com/ArimaGenomics/mapping_pipeline) and then scaffolded them with SALSA (Ghurye et al. 2017, 2019).

Both genome assemblies were manually curated by iteratively generating and analyzing their corresponding Omni-C contact maps. To generate the contact maps we aligned the Omni-C data with BWA-MEM (Li 2013), identified ligation junctions, and generated Omni-C pairs using pairtools (Goloborodko et al. 2018; Abdennur et al. 2023). We generated a multi-resolution Omni-C matrix with cooler (Abdennur and Mirny 2020) and balanced it with hicExplorer (Ramírez et al. 2018). We used HiGlass (Kerpedjiev et al. 2018) and the PretextSuite (https://github.com/wtsi-hpag/PretextView; https://github.com/wtsi-hpag/PretextMap; https://github.com/wtsi-hpag/PretextSnapshot) to visualize the contact maps where we identified misassemblies and

Table 1. Assembly pipeline and software used.

Assembly	Software and any non-default options	Version
Filtering PacBio HiFi adapters	HiFiAdapterFilt	Commit 64d1c7b
K-mer counting	Meryl (k = 21)	1
Estimation of genome size and heterozygosity	GenomeScope	2
De novo assembly (contiging)	HiFiasm (Hi-C Mode, –primary, output p_ctg.hap1, p_ctg.hap2)	0.16.1-r375
Scaffolding		
Omni-C data alignment	Arima Genomics Mapping Pipeline	Commit 2e74ea4
Omni-C Scaffolding	SALSA (-DNASE, -i 20, -p yes)	2
Gap closing	YAGCloser (-mins 2 -f 20 -mcc 2 -prt 0.25 -eft 0.2 -pld 0.2)	Commit 0e34c3b
Omni-C Contact map generation		
Short-read alignment	BWA-MEM (-5SP)	0.7.17-r1188
SAM/BAM processing	samtools	1.11
SAM/BAM filtering	pairtools	0.3.0
Pairs indexing	pairix	0.3.7
Matrix generation	cooler	0.8.10
Matrix balancing	hicExplorer (hicCorrectmatrix correct --filterThreshold -2 4)	3.6
Contact map visualization	HiGlass	2.1.11
	PretextMap	0.1.4
	PretextView	0.1.5
	PretextSnapshot	0.0.3
Manual curation tools	Rapid curation pipeline (Wellcome Trust Sanger Institute, Genome Reference Informatics Team)	Commit 4ddca450
Genome quality assessment		
Basic assembly metrics	QUAST (--est-ref-size)	5.0.2
Assembly completeness	BUSCO (-m geno, -l embryophyta)	5.0.0
	Merqury	2020-01-29
Contamination screening		
Local alignment tool	BLAST+ (-db nt, -outfmt "6 qseqid staxids bitscore std," -max_target_seqs 1, -max_hsps 1, -evalue 1e-25)	2.1
General contamination screening	BlobToolKit (HiFi coverage, BUSCO = embryophyta, NCBI Taxa ID = 60424)	2.3.3
Genome assembly comparisons		
Genome-to-genome sequence alignments	lastz	1.04.18
Calculate mean nucleotide identity	Bedtools	2.30.0

misjoins, and finally modified the assemblies using the Rapid Curation pipeline from the Wellcome Trust Sanger Institute, Genome Reference Informatics Team (https://gitlab.com/wtsi-grit/rapid-curation). Some of the remaining gaps (joins generated during scaffolding and/or curation) were closed using the PacBio HiFi reads and YAGCloser (https://github.com/merlyescalona/yagcloser). Finally, we checked for contamination using the BlobToolKit Framework (Challis et al. 2020).

Genome quality assessment

We generated k-mer counts from the PacBio HiFi reads using meryl (https://github.com/marbl/meryl). The k-mer counts were then used in GenomeScope2.0 (Ranallo-Benavidez et al. 2020) to estimate genome features including genome size, heterozygosity, and repeat content. To obtain general contiguity metrics, we ran QUAST (Gurevich et al. 2013). To evaluate genome quality and functional completeness we used BUSCO (Manni et al. 2021) with the Embryophyta ortholog database (embryophyta_odb10) which contains 1,614 genes. Assessment of base level accuracy (QV) and k-mer completeness was performed using the previously generated meryl database and merqury (Rhie et al. 2020). We further estimated genome assembly accuracy via BUSCO gene set frameshift analysis using the pipeline described in Korlach et al. (2017). Measurements of the size of the phased blocks is based on the size of the contigs generated by HiFiasm on HiC mode. We follow the quality metric nomenclature established by Rhie et al. (2021), with the genome quality code $x.y.P.Q.C$, where, x = log10[contig NG50]; y = log10[scaffold NG50]; P = log10 [phased block NG50]; Q = Phred base accuracy QV (quality value); C = % genome represented by the first "n" scaffolds, following a karyotype of $2n$ = 24, known for the number of chromosomes for this species (Genome on a Tree—GoaT; tax_tree(quercus tomentella), Challis et al. 2023). Quality metrics for the notation were calculated on the assembly for haplotype 1.

Comparison to other genomes

The quality metrics of the *Q. tomentella* genome were compared with that of other oak genomes available on NCBI

(accessed 17 October 2023). The nucleotide identity of *Q. tomentella* was also compared with four other oak genomes: *Quercus aquifolioides*, an Asian species from subgenus *Cerris*, section Ilex (NCBI:GCA_019022515.1); and three species from subgenus *Quercus*, section *Quercus*: *Q. lobata* from North America (NCBI:GCA_001633185.5), *Q. mongolica* from Asia (NCBI:GCA_011696235.1), and *Q. robur* from Europe and Western Asia (NCBI:GCA_932294415.1). Nucleotide identity was determined from chromosome-to-chromosome lastz (Harris 2007) alignments using default (sensitive) parameters. To minimize error from poorly aligned regions and focus on nucleotide identity in clearly cognate regions, the set of longest alignments covering approximately 25% of the reference genome (*Q. tomentella*) were used for the calculation. Bedtools (Quinlan and Hall 2010) and perl scripts were used to collapse overlapping alignments and calculate mean nucleotide identity for overlapping fragments (no weighting).

Results

The PacBio HiFi and Omni-C sequencing libraries generated 91.91 million read pairs and 1.88 million reads respectively. The latter yielded ~40-fold coverage (N50 read length 16,523 bp; minimum read length 157 bp; mean read length 16,218 bp; maximum read length of 54,575 bp) based on the Genomescope 2.0 genome size estimation of 758.51 Mb. Based on PacBio HiFi reads, we estimated 0.16% sequencing error rate and 1.55% nucleotide heterozygosity rate. The k-mer spectrum based on PacBio HiFi reads show (Fig. 2A) a bimodal distribution with two major peaks at ~20 and ~40-fold coverage, where peaks correspond to homozygous and heterozygous states of a diploid species. The distribution presented in this k-mer spectrum supports that of a high heterozygosity profile.

The final assembly (dhQueTom1) consists of two haplotypes, and both genome assemblies are similar but not equal to the estimated value from Genomescope2.0, which has been observed in other taxa (Fig. 2A; for example, see Pflug et al. 2020). Haplotype 1 consists of 304 scaffolds spanning 780.70 Mb with contig N50 of 18.34 Mb, scaffold N50 of 54.38 Mb, longest contig of 47.06 Mb and largest scaffold of 90.51 Mb. On the other hand, haplotype 2 consists of 181 scaffolds, spanning 774.88Mb with contig N50 of 21.98 Mb, scaffold N50 of 48.66 Mb, largest contig 42.19 Mb, and largest scaffold of 81.8 Mb.

During manual curation, we generated a total of 13 breaks, where 8 breaks were made on haplotype one and 6 were made on haplotype two; and 74 joins, 40 joins on haplotype one, and 34 on haplotype two. We were able to close 6 gaps in total, 4 gaps on haplotype one and 2 gaps on haplotype two. Finally, we removed a single contig on haplotype one corresponding to contaminants which matched to an Arthropod. We detected a potential ~14Mb long inversion between haplotype one and two on the 9th largest scaffold (Fig. 3).

The haplotype one has a BUSCO completeness score of 97.3% using the Embryophyta gene set, a per base quality (QV) of 64.22, a kmer completeness of 80.18 and a frameshift indel QV of 49.79. The haplotype two has a BUSCO completeness score of 98.5% using the same gene set, a per base quality (QV) of 64.31, a kmer completeness of 80.23, and a frameshift indel QV of 49.83. The Omni-C contact maps shows that both assemblies are highly contiguous with some chromosome-length scaffolds (Fig. 2C and D). We have deposited scaffolds corresponding to both assemblies to the National Center for Biotechnology Information (NCBI; see Table 2 and Data Availability for details).

Assembly statistics are reported in Table 2, and in Fig. 2B for haplotype 1 (see Supplementary Fig. S1 for haplotype 2). Supplementary Table S1 shows a comparison of assembly statistics, including contig and scaffold N50 values, among other oak genomes. Among the four species used for comparison, *Q. tomentella* had the highest nucleotide identity with *Q. aquifolioides* at 92.2% (Table 3).

Fig. 1. Photos of *Q. tomentella* on Santa Rosa Island, CA. (Left) Image of leaves and flowers. (Right) Grove of trees; the multiple stems growing near each other may be one clonal individual. Both photos were taken by Alayna Mead.

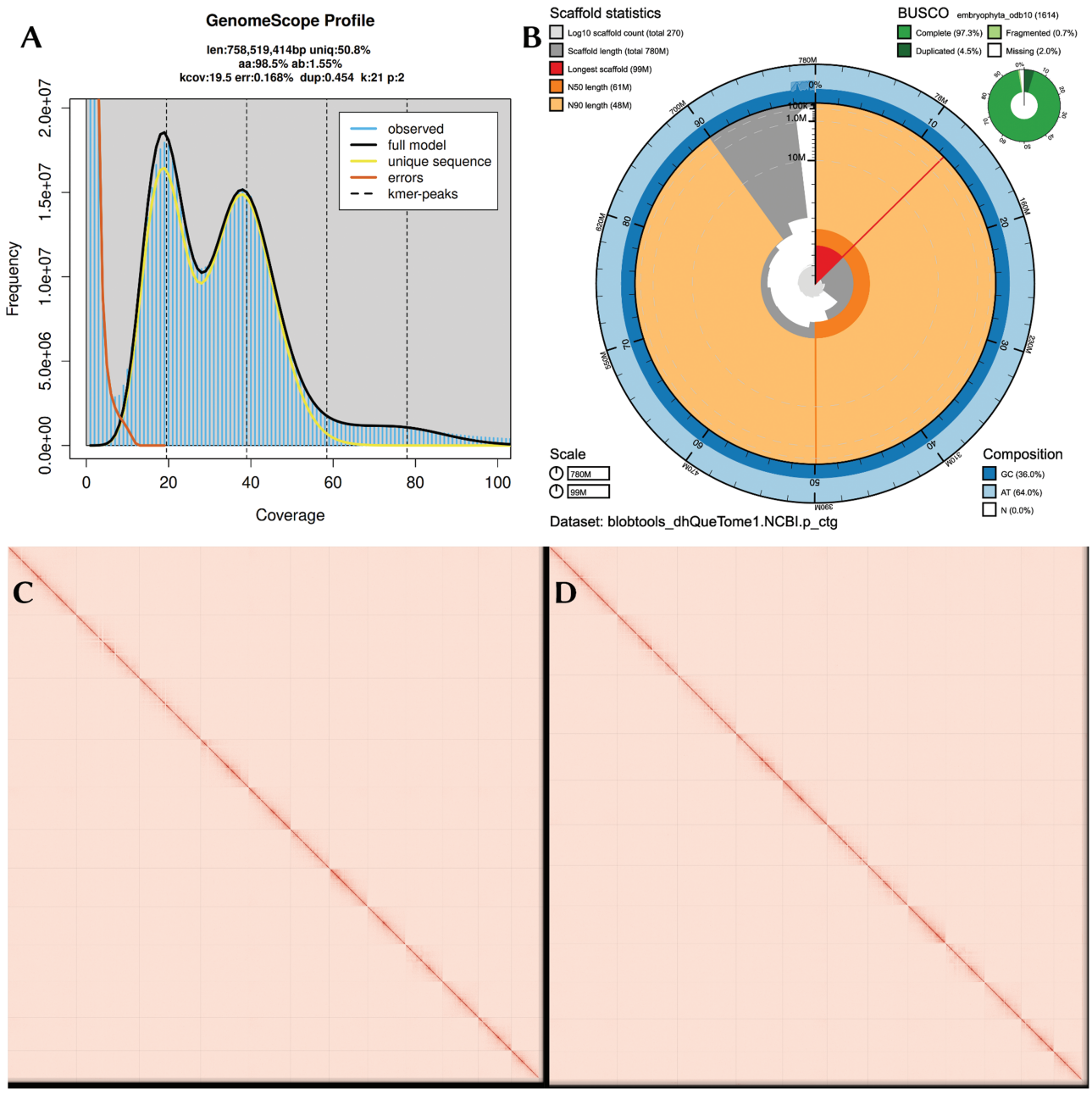

Fig. 2. (A) K-mer spectra output generated from PacBio HiFi data without adapters using GenomeScope2.0. The bimodal pattern observed corresponds to a diploid genome and the k-mer profile matches that of high heterozygosity. K-mers covered at lower coverage and high frequency correspond to differences between haplotypes, whereas the higher coverage and slightly lower frequency correspond to the similarities between haplotypes. (B) BlobToolKit snail plot showing a graphical representation of the quality metrics presented in Table 2 for the *Quercus tomentella* haplotype 1 (dhQueTome1.0.hap1). The first diagonal (~30 degrees) line represents the size of the longest scaffold; all other scaffolds are arranged in size-order moving clockwise around the plot and drawn in gray starting from the outside of the central plot. Dark and light orange arcs show the scaffold N50 and scaffold N90 values. The central light gray spiral shows the cumulative scaffold count with a white line at each order of magnitude. White regions in this area reflect the proportion of Ns in the assembly; the dark vs. light blue area around it shows mean, maximum and minimum GC vs. AT content at 0.1% intervals (Challis et al. 2020). The corresponding plot for the haplotype 2 assembly is in the supplemental material (Supplementary Fig. S1). (C) Haplotype 1 and (D) haplotype 2 assembly Omni-C contact maps generated with PretextSnapshot. Omni-C contact maps translate proximity of genomic regions in 3D space to contiguous linear organization. Each cell in the contact map corresponds to sequencing data supporting the linkage (or join) between two of such regions.

Discussion

Here we present a high-quality genome assembly for *Q. tomentella*, an island relictual tree species of conservation concern, as well as the first species in *Quercus* section *Protobalanus* for which a reference genome is available. The length of haplotype 1 assembly (dhQueTome1.0.hap1) is 781 Mb, which is similar to that of other oak species (720 to 953 MB, Supplementary Table S1). The contig N50 is 18.3 Mb, and the scaffold N50 is 60.6 Mb. There are currently 16 other oak species with an assembled genome available on

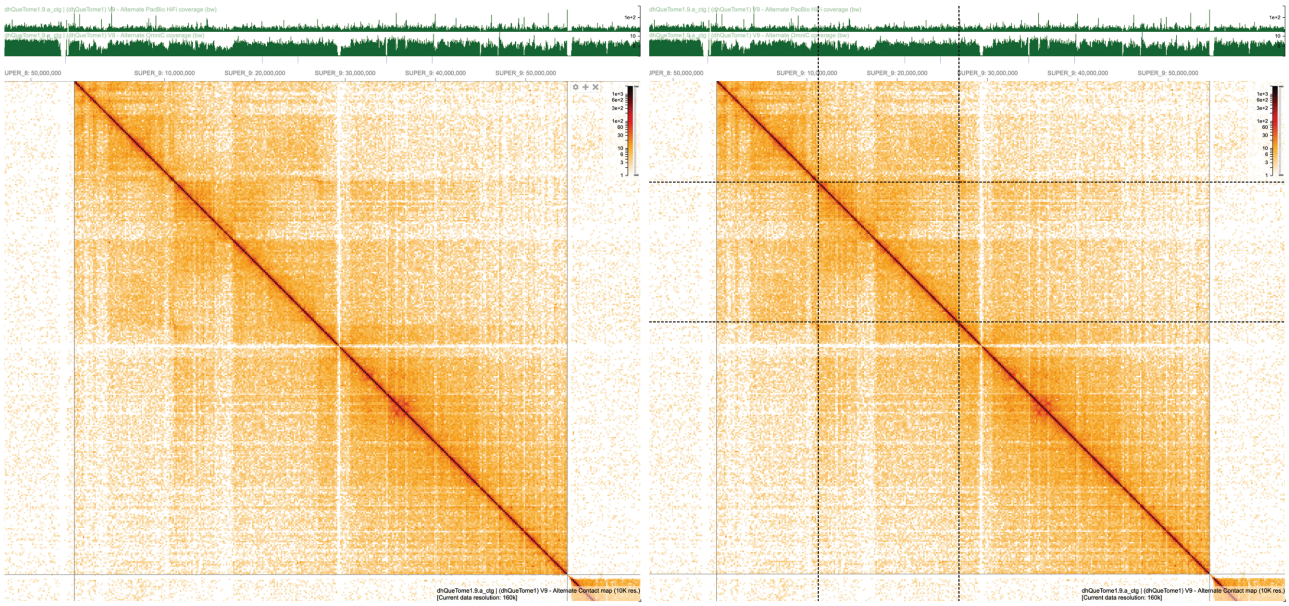

Fig. 3. Contact map (as in Fig. 2C and D) for the 9th largest scaffold, showing a possible inversion between the two haplotypes. The barplots above the contact map show the coverage for the PacBio (top) and OmniC (bottom) assembly. The left image shows the unaltered plot and the right image highlights the inversion block within the dotted lines, identified by the darker regions in the upper right and lower left corner which indicates two spatially distant regions with relatively high linkage, suggesting they may in fact be adjacent to each other.

NCBI (Supplementary Table S1). The contig and scaffold N50 values for our scaffold-level assembly of *Q. tomentella* rank in the upper half of these oak genome assemblies. The scaffold N50 for *Q. tomentella* is the second highest of the other scaffold-level assemblies, following *Quercus berberidifolia*, another CCGP reference genome (O'Donnell et al, unpublished data).

We find evidence for a ~14 Mb long inversion on the 9th largest scaffold (Fig. 3). Future work using whole-genome resequencing data of *Q. tomentella* (generated as part of the CCGP) is necessary to validate this putative inversion and investigate whether it has evolutionary significance. Intraspecific variation in inversions between haplotypes has been found in an increasing number of species, including *Q. rubra* (Kapoor et al. 2023), and these chromosomal inversions may be involved in adaptation by reducing recombination within blocks of locally adapted alleles (Huang and Rieseberg 2020).

Interestingly, *Q. tomentella* shares a greater percent identity with *Q. aquifolioides*, from *Quercus* section *Cerris*, than with any of the species from *Quercus* section *Quercus*, with whom it shares a more recent common ancestor (Hipp et al. 2018). This could be the result of accelerated evolution in section *Quercus* (Hipp et al. 2018), and underscores the importance of having reference genomes spanning the diversity of the oak clade.

As part of the CCGP, this genome assembly will provide a valuable resource for genomics-informed conservation of *Q. tomentella*. Additionally, the growing number of oak reference genomes will enable further study of their diversification and evolutionary success (Kremer and Hipp 2019; Sork et al. 2022).

Supplementary material

Supplementary material is available at *Journal of Heredity* Journal online.

Acknowledgments

The Channel Islands are the ancestral and unceded territories of the Chumash and Gabrielino-Tongva Peoples—past, present, and emerging—and we acknowledge them as the traditional caretakers of the islands and the oak ecosystems sampled for our research. We thank the California Botanic Garden for sampling permission, and James Reed for sampling assistance. PacBio Sequel II library prep and sequencing was carried out at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. Deep sequencing of Omni-C libraries used the Novaseq 6000 sequencing platforms at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. We thank the staff at the UC Davis DNA Technologies and Expression Analysis Cores and the UC Santa Cruz Paleogenomics Laboratory for their diligence and dedication to generating high-quality sequence data.

Funding

This work was supported by the California Conservation Genomics Project, with funding provided to the University of California by the State of California, State Budget Act of 2019 [UC Award ID RSI-19-690224].

Data availability

Data generated for this study are available under NCBI BioProject PRJNA808370. Raw sequencing data for sample Qtom.A.LA.223.w1 (Isolate RSA801373; NCBI BioSample SAMN29046697) are deposited in the NCBI Short Read Archive (SRA) under SRX19356635 for PacBio HiFi sequencing data, and SRX19356636 and SRX19356637

Table 2. Sequencing and assembly statistics, and accession numbers.

Bio Projects and Vouchers	CCGP NCBI BioProject		PRJNA720569				
	Genera NCBI BioProject		PRJNA765841				
	Species NCBI BioProject		PRJNA808370				
	NCBI BioSample		SAMN29046697				
	Specimen identification		RSA801373				
	NCBI Genome accessions		**Haplotype 1**			**Haplotype 2**	
	Assembly accession		JAPQVX000000000			JAPQVY000000000	
	Genome sequences		GCA_028535515.1			GCA_028535165.1	
Genome Sequence	PacBio HiFi reads	Run	1 PACBIO_SMRT (Sequel II) run: 1.9M spots, 30.6G bases, 17.8Gb				
		Accession	SRX19356635				
	Omni-C Illumina reads	Run	2 ILLUMINA (Illumina NovaSeq 6000) runs: 91.6M spots, 27.7G bases, 8.7Gb				
		Accession	SRX19356636, SRX19356637				
Genome Assembly Quality Metrics	Assembly identifier (Quality code*)				dhQueTome1(7.7.P7.Q64.C)		
	HiFi Read coverage§		40.37×				
			Haplotype 1			**Haplotype 2**	
	Number of contigs		327			209	
	Contig N50 (bp)		18,345,613			21,984,091	
	Contig NG50§		18,345,613			21,984,091	
	Longest Contigs		47,067,633			42,190,043	
	Number of scaffolds		274			156	
	Scaffold N50		60,584,938			63,372,236	
	Scaffold NG50§		60,584,938			63,372,236	
	Largest scaffold		98,559,247			97,120,727	
	Size of final assembly		780,705,769			774,886,980	
	Phased block NG50§		18,345,613			21,984,091	
	Gaps per Gbp (# Gaps)		68(53)			68(53)	
	Indel QV (Frame shift)		49.79244778			49.83235261	
	Base pair QV		64.2217			64.3106	
			Full assembly = 64.2657				
	k-mer completeness		80.1833			80.2349	
			Full assembly = 98.9136				
	BUSCO completeness (embryophyta_odb10) *n* = 1,614		C**	S**	D**	F**	M**
		H1‡	97.30%	92.80%	4.50%	0.70%	2.00%
		H2‡	98.50%	94.60%	3.90%	0.70%	0.80%
	Organelles		# Partial/complete mitochondrial sequence				

*Assembly quality code *x.y.P.Q.C* derived notation, from Rhie et al. (2021). *x* = log10[contig NG50]; *y* = log10[scaffold NG50]; *P* = log10 [phased block NG50]; *Q* = Phred base accuracy QV (Value); C = % genome represented by the first "*n*" scaffolds, following a karyotype of 2*n* = 24, known for the number of chromosomes for this species (Genome on a Tree—GoaT; tax_tree(quercus tomentella)). Quality code for all the assembly denoted by primary assembly (dhQueTome1.0.hap1).
**BUSCO Scores. Complete BUSCOs (C). Complete and single-copy BUSCOs (S). Complete and duplicated BUSCOs (D). Fragmented BUSCOs (F). Missing BUSCOs (M).
§Read coverage and NGx statistics have been calculated based on the estimated genome size of 758.51 Mb.
‡(H1) Haplotype 1 and (H2) Haplotype 2 assembly values.

Table 3. Nucleotide percent identity between five *Quercus* species based on the set of largest lastz alignments summing to 25% of the reference genome.

Subgenus	*Quercus*	*Quercus*	*Quercus*	*Quercus*	*Cerris*
Section	*Protobalanus*	Quercus	Quercus	Quercus	Ilex
Query | Ref	Qtom	Qlob	Qmon	Qrob	Qaqu
Qtom	–	83.8	83.6	87.4	91.1
Qlob	85.8	–	89.8	91.1	89.9
Qmon	84.7	89.1	–	93.4	90.5
Qrob	87.4	91.2	92.7	–	91.2
Qaqu	92.2	90.2	91.0	91.0	–

Qtom is *Q. tomentella*, Qlob is *Q. lobata*, Qmon is *Q. mongolica*, Qrob is *Q. robur*, and Qaqu is is *Q. aquilifolioides*. Row labels designate the query genome and column labels designate the reference genome for each pairwise alignment.

for the Omni-C Illumina sequencing data. NCBI GenBank accessions for both haplotypes of the genome assembly are GCA_028535515.1 (for dhQueTome1.0.hap1) and GCA_028535165.1 (for dhQueTome1.0.hap2); and for genome sequences JAPQVX000000000 and JAPQVY000000000. Assembly scripts and other data for the analyses presented can be found at the following GitHub repository: www.github.com/ccgproject/ccgp_assembly.

References

Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics. 2020:36:311–316.

Abdennur N, Fudenberg G, Flyamer IM Galitsyna AA, Goloborodko A, Imakaev M, Venev SV. Pairtools: from sequencing data to chromosome contacts. bioRxiv. 2023: Feb 15.

Ai W, Liu Y, Mei M, Zhang X, Tan E, Liu H, Han X, Zhan H, Lu X. A chromosome-scale genome assembly of the Mongolian oak (*Quercus mongolica*). Mol Ecol Resour. 2022:22:2396–2410.

Ashley MV, Backs JR, Kindsvater L, Abraham ST. Genetic variation and structure in an endemic island oak, *Quercus tomentella*, and mainland canyon oak, *Quercus chrysolepis*. Int J Plant Sci. 2018:179:151–161.

Axelrod DI. A miocene flora from the western border of the Mohave desert. Washington, D.C.: The Carnegie Institution of Washington; 1939.

Axelrod DI. The Mulholland Flora. In: Chaney RW, editor. Pliocene floras of California and Oregon, [Carnegie Institution of Washington] Contributions to paleontology. Washington, D. C.: The Carnegie Institution of Washington; 1944a. p. 103–162.

Axelrod DI. The Sonoma Flora. In: Chaney RW, editor. Pliocene floras of California and Oregon, [Carnegie Institution of Washington] Contributions to paleontology. Washington, D. C.: The Carnegie Institution of Washington; 1944b. p. 167–218.

Beckman E, Jerome D. *Quercus tomentella*. IUCN Red List of Threatened Species. 2017.

Beckman E, Meyer A, Denvir A, Gill D, Man G, Pivorunas D, Shaw K, Westwood M. Conservation gap analysis of native U.S. Oaks. Lisle, IL: The Morton Arboretum; 2019.

Bodénès C, Chancerel E, Ehrenmann F, Kremer A, Plomion C. High-density linkage mapping and distribution of segregation distortion regions in the oak genome. DNA Res. 2016:23:115–124.

Cavender-Bares J. Diversification, adaptation, and community assembly of the American oaks (*Quercus*), a model clade for integrating ecology and evolution. New Phytol. 2018:221:669–692.

Challis R, Kumar S, Sotero-Caio C, Brown M, Blaxter M. Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life. Wellcome Open Res. 2023:8:24.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit—interactive quality assessment of genome assemblies. G3 Genes|Genomes|Genetics. 2020:10:1361–1374.

Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, Heng L. 2021. Robust haplotype-resolved assembly of diploid individuals without parental data, https://www.nature.com/articles/s41587-022-01261-x.

Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. Scaffolding of long read assemblies using long range contact information. BMC Genomics. 2017:18:527.

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019:15:e1007273.

Goloborodko A, Abdennur N, Venev S, Brandao HB, Fudenberg G. mirnylab/pairtools: v0.2.0. Zenodo; 2018.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013:29:1072–1075.

Harris RS. Improved pairwise alignment of genomic DNA [Ph.D. Thesis]. The Pennsylvania State University; 2007.

Hipp AL, Manos PS, González-Rodríguez A, Hahn M, Kaproth M, McVay JD, Avalos SV, Cavender-Bares J. Sympatric parallel diversification of major oak clades in the Americas and the origins of Mexican species diversity. New Phytol. 2018:217:439–452.

Huang K, Rieseberg LH. Frequency, origins, and evolutionary role of chromosomal inversions in plants. Front Plant Sci. 2020:11:1–13.

Kapoor B, Jenkins J, Schmutz J, Zhebentyayeva T, Kuelheim C, Coggeshall M, Heim C, Lasky JR, Leites L, Islam-Faridi N, et al. A haplotype-resolved chromosome-scale genome for *Quercus rubra* L provides insights into the genetics of adaptive traits for red oak species. G3 (Bethesda, Md.). 2023:13:jkad209.

Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A, Kumar N, et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biol. 2018:19:125.

Kindsvater LC. Conservation and restoration of the endemic island oak, Quercus tomentella in Channel Islands National Park using a habitat approach [Ph.D. Thesis]. [Davis]: University of California; 2006.

Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, Cantin L, Jarvis ED. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. GigaScience. 2017:6:gix085.

Kremer A, Hipp AL. Oaks: an evolutionary success story. *New Phytologist*. 2019:226:987–1011.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv, preprint: not peer reviewed.

Maldonado-Alconada AM, Castillejo MA, Rey M-D, Labella-Ortega M, Tienda-Parrilla M, Hernández-Lao T, Honrubia-Gómez I, Ramírez-García J, Guerrero-Sanchez VM, López-Hidalgo C, et al. Multiomics molecular research into the recalcitrant and orphan *Quercus ilex* tree species: why, what for, and how. Int J Mol Sci. 2022:23:9980.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021:38:4647–4654.

Nixon KC. The oak (*Quercus*) biodiversity of California and adjacent regions. In: Standiford RB, Purcell KL, editors. USDA Forest Service Gen. Tech. Rep. Vol PSW-GTR-18. San Diego, CA; 2002. p. 3–20.

Ortego J, Gugger PF, Sork VL. Genomic data reveal cryptic lineage diversification and introgression in Californian golden cup oaks (section *Protobalanus*). New Phytol. 2018:218:804–818.

Pesendorfer MB, Baker CM, Stringer M, McDonald-Madden E, Bode M, McEachern AK, Morrison SA, Sillett TS. Oak habitat recovery on California's largest islands: scenarios for the role of corvid seed dispersal. J Appl Ecol. 2018:55:1185–1194.

Pflug JM, Holmes VR, Burrus C, Johnston JS, Maddison DR. Measuring genome sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles (Coleoptera). G3 (Bethesda, Md.). 2020:10:3047–3060.

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

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010:26:841–842.

Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018:9:189.

Ramos AM, Usié A, Barbosa P, Barros PM, Capote T, Chaves I, Simões F, Abreu I, Carrasquinho I, Faro C, et al. The draft genome sequence of cork oak. Sci Data. 2018:5:180069.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 20 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020:11:1432.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021:592:737–746.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020:21:245.

Sawyer JO, Keeler-Wolf T, Evens JM. CNPS alliance: *Quercus tomentella—Lyonothamnus floribundus*. In: A manual of California vegetation. 2nd ed. Sacramento, CA: California Native Plant Society; 2009. p. 1300.

Shaffer HB, Toffelmier E, Corbett-Detig RB, Escalona M, Erickson B, Fiedler P, Gold M, Harrigan RJ, Hodges S, Luckau TK, et al. Landscape genomics to enable conservation actions: the California conservation genomics project. J Hered. 2022:113:577–588.

Sim SB, Corpuz RL, Simmonds TJ, Geib SM. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. BMC Genomics. 2022:23:157.

Sork VL, Cokus SJ, Fitz-Gibbon ST, Zimin AV, Puiu D, Garcia JA, Gugger PF, Henriquez CL, Zhen Y, Lohmueller KE, Pellegrini, M, Salzberg, SL. High-quality genome and methylomes illustrate features underlying evolutionary success of oaks. Nat Commun. 2022:13:2047.

Workman R, Timp W, Fedak R, Kilburn D, Hao S, Liu K. High molecular weight DNA extraction from recalcitrant plant species for third generation sequencing. Nat Protoc Exchange. 2018.