A chromosome-scale genome assembly of the Mongolian oak (*Quercus mongolica*)

Wanfeng Ai[1], Mei Mei[1], Xiaolin Zhang[2], Lijie Zhang[2], Xiaoyi Han[2], Hao Zhan[1], Xiujun Lu[2]*

[1] College of Horticulture, Shenyang Agricultural University, Shenyang 110866, Liaoning, China.

[2] College of Forestry, Shenyang Agricultural University, Shenyang, 110866, Liaoning, China.

*Corresponding author Xiujun Lu (lxjsyau@syau.edu.cn)

Abstract

Quercus mongolica (Fagaceae) is an important ecological and economic tree species in East Asia. It has excellent biological characteristics, such as hardwood, strong resistance to biotic and abiotic stresses. The availability of a high-quality genome will help to further reveal the underlying mechanisms. Here we assemble the first chromosome-level reference genome of *Q. mongolica*. The final assembled genome was 809.84 Mb with contig and scaffold N50s of 32.64 Mb and 66.74 Mb, respectively. Hi-C scaffolding anchored twelve pseudochromosomes, accounting for 95.65% of the assembled genome. Moreover, 68.5% and 5.4% of the genomic sequence were transposon elements and tandem repeat elements, respectively. A total of 36,553 protein−coding genes were predicted, of which 94.89% were functionally annotated. Comparative genomics analysis indicated that *Q. mongolica* was more closely related to *Q. robur* than to either *Q. lobata* or *Q. suber*. *Q. mongolica* and *Q. robur* diverged ~10.2 Mya. *Q. mongolica* had undergone two whole-genome duplications which occurred earlier than *Q. robur*. We identified multiple genes in 38 positive selection genes, including *pyridoxal reductase 1* (*PLR1*) and *switch subunit 3* (*SWI3B*). In addition, we identified 496 genes

related to wood formation, 88 WRKY genes, and 124 NAC genes in *Q. mongolica*. This genomic information will be an important molecular resource for further exploring the biological characteristics and adaptive evolution of *Q. mongolica*. Meanwhile, the genomic resource from Asian oak will also contribute to the study of the taxonomy, evolution and conservation of *Quercus* species.

KEYWORDS: *Quercus mongolica*, chromosome-scale genome assembly, genome annotation, comparative genomic analysis, PacBio Sequel II

1 | INTRODUCTION

Oaks (genus *Quercus*, Fagaceae) are a major component of forests and savanna ecosystems in the Northern Hemisphere (Cavender-Bares, 2019). They are distributed below 4,000 m above sea level and between the equator and 60 degrees north latitude (Kremer & Hipp, 2019). These long-lived species have high genetic diversity due to extensive gene flow between species (Pang et al., 2019). At present, there are more than 435 *Quercus* species in Asia, North America and Europe (Cannon et al., 2018). Because of their ecological dominance and remarkable diversity, and the increasing phylogenetic, genomic, and ecological data resources that are available, oaks have become an important model for exploring the genomic footprint of evolutionary and ecological changes (Cavender-Bares et al., 2015; Lesur et al., 2015; Cavender-Bares, 2019). To date, the genome sequence of three *Quercus* species from Europe and North America has been published, namely *Q. suber* (Ramos et al., 2018), *Q. robur* (Plomion et al., 2016,2017), *Q. lobata* (Sork et al., 2016). However, the genomic resources for

Asian oak species are still insufficient, which seriously hinders research on the taxonomy, evolution, conservsation, ecology, and genetics of *Quercus* species.

The Mongolian oak, *Quercus mongolica*, is a deciduous tree commonly found in cold temperate zones of Asia. It is mainly distributed in the Russian Far East, Japanese Islands, Korean Peninsula, and Northern China (Chen & Huang, 1998). *Q. mongolica* is highly resistant to pests, diseases, coldness and drought (Hao & Yang, 2016). Moreover, its wood is hard and corrosion resistant, and thus an excellent material for making vehicles, ships, buildings, and furniture (Li, 2003). Its leaves are rich in nutrients, where 17 amino acids are found and contents of crude protein, fat, and fiber and calcium concentration are higher than those of maize, making them a type of high-quality raw material for feeds for deer, cattle, and sheep (Mi et al., 1999). Meanwhile, oak leaf is also used for breeding the Chinese Tussah Silkmoth (*Antheraea pernyi*) (Jiang et al., 2019). Oak fruit contains more than 55% of starch, and its total amount of unsaturated fatty acids can reach 81%, which is close to that of corn, so it is often used in food, feed, and starch industries (Ao et al., 1998). In addition, many triterpenoids, phenolic glycosides, flavonoids, and tannins have been isolated from *Q. mongolica*. These bioactive substances have anti-oxidation, anti-tumor, anti-inflammatory, and anti-fungal activities (Ishimaru et al., 1987; Omar et al., 2013; Kim et al., 2015; Zhou et al., 2017; Yin et al., 2019).

Q. mongolica has shown an excellent potential to be applied in ecology, economy, and medicine. Progress has been made in using *Q. mongolica* to study forest ecology (Zeng et al., 2016; Watanabe et al., 2018; Zhang et al., 2020) (Watanabe et al., 2010; Cannon et al., 2018), evolutionary biology (Liao et al., 2019; Nagamitsu et al., 2019), and bioorganic chemistry

(Yin et al., 2019; Min et al., 2020) in Asian countries like China, Japan, and South Korea. However, the molecular biology of *Q. mongolica* is still in its infancy, and the mechanisms underlying some excellent biological properties (including hard wood, strong resistance to biotic and abiotic stress) are still unclear. The insufficient development of genome resources has seriously restricted the biological research, conservation and genome breeding of this species. In order to solve these problems, a high-quality *Q. mongolica* genome is urgently needed.

In this study, we assembled a chromosome-level reference genome of *Q. mongolica* based on the combination strategy of Illumina short-read, PacBio long-read, and high-throughput chromosome conformation capture (Hi-C) sequencing technology. This is the first report of the high-quality genome of Asian oak. Genomic resources for this species will help to study the taxonomy and evolution of oaks, and at the same time provide valuable genetic information for the protection and utilization of *Q. mongolica* germplasm.

2 | MATERIALS AND METHODS

2.1 | Sample collection

A healthy and mature *Q. mongolica* individual tree was chosen for sampling from the campus of Shenyang Agricultural University (123°34′24″E, 41°49′19″N), Liaoning province, China. The tree was 5 m tall with a chest diameter of 19.7 cm (Figure 1A). Its fruit was collected in September 2018 (Figure 1E). Samples of its roots, twigs, leaves, female and male flowers (Figure 1B, C, D) were collected from the same tree in May 2019. All samples were rinsed

with deionized water, frozen in liquid nitrogen, and stored at -80°C for DNA and RNA extraction.

2.2 | Library construction and sequencing

Genomic DNA of *Q. mongolica* were extracted from leaves and randomly sheared into fragments of ~350 bp in length which was then used for library construction for Illumina paired−end sequencing according to the manufacturer's instructions. Basically the libraries were prepared following these steps: DNA fragmentation by sonication, end-polishing of the DNA fragments, A-tailing and ligation with the full-length adapters for Illumina sequencing, PCR amplification, and purification of PCR products (AMPure XP bead system). The libraries were analyzed for size distribution using an Agilent 2100 Bioanalyzer and were quantified using real-time PCR. Then the libraries were sequenced using the Illumina HiSeq X-ten platform.

After examination of the quality of isolated DNA from the fresh leaves from *Q. mongolica,* the library of 20 kb was constructed using a SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, CA, USA). The construction includes DNA shearing, damage repair, end repair, hairpin adapter ligation, and purification of the library. After quality control test, a SMRTbell library was obtained. The library was sequenced using a single 8 M SMAT Cell on the PacBio Sequel II platform (Pacific Biosciences, CA, USA) (PacBio Sequel II System).

The Hi-C library was prepared using the method described previously (Xie et al., 2015). In short, the library was constructed through the following steps: DNA cross-linking, *Dpn* II digestion, cohesive end repair, DNA cyclization and purification and random shearing into

300-500 bp fragments. Avidin magnetic beads were used to capture labeled DNA. After quality control test of the libraries using Qubit 2.0, an Agilent 2100 instrument (Agilent Technologies, CA, USA), and q-PCR, 150 bp PE sequencing of these libraries were performed on the Illumina HiSeq X Ten platform.

For PacBio Iso-Seq, full-length complementary DNA was synthesized from a total RNA sample where the equal amounts of total RNA from different tissues were mixed. The SMARTer PCR cDNA Synthesis Kit (Takara Bio) was used for cDNA synthesis. The cDNA product was filtered using the BluePippin DNA Size Selection System (Sage Science). The Iso-Seq libraries were constructed following the standard SMRT bell construction protocol (Pacific Biosciences, CA, USA) and sequenced on the PacBio Sequel II platform (Pacific Biosciences, CA, USA).

2.3 | Genome survey and de novo assembly

The genome size of *Q. mongolica* was estimated by the k-mer method (Liu et al., 2013) using sequencing data from Illumina DNA library. Quality-filtered reads were subjected to 17-mer frequency distribution analysis using the GCE program.

The PacBio SMRT-Analysis package (https://www.pacb.com) was used for the quality control of the raw polymerase reads; sequencing adaptors and low-quality short reads were removed. The remaining high-quality subreads of *Q. mongolica* were initially assembled by Falcon v.2.0 (Chin et al., 2013) software with the following parameters: seed_coverage=55 and Length_cutoff_pr=15k. Then the original assembly results were polished with Arrow embed smrtlink 7.0 based on corrected subreads. Finally, the polished sequences were further

corrected with reference to the Illumina reads with Pilon (Walker et al., 2014) using two rounds. The draft genome was obtained by filtering the heterozygous redundant contigs through Purge_haplotigs (Roach et al., 2018).

2.4 | Chromosome-scale assembly with Hi-C data

Low-quality Hi-C reads were filtered and the remaining reads were aligned to the draft genome of *Q. mongolica* by BWA v.0.7.8 (Li & Durbin, 2009) software using default parameters. Reads were excluded from subsequent analysis if they did not align within 500 bp of a restriction site. After assisted assembly of the genome, interaction maps were constructed using Juicer v.1.6.2 (Durand et al., 2016) and visually using JucieBox v.1.8.8 (Durand et al., 2016). The preassembled contigs were clustered, ordered and directed onto the pseudochromosomes with LACHESIS software (Burton et al., 2013). To improve the chromosome-scale assembly quality, manual adjustment of orientation errors with obvious discrete chromatin interaction patterns was performed.

2.5 | Genome assembly quality assessment

To assess the quality of genome assembly, the Continuous Long Reads (CLR) subreads of *Q. mongolica* were selected and aligned back to the assembled genome using minimap2 v.2.5 (Li, 2018). In addition, BUSCO v.3.0.2 (Simão et al., 2015) was used to evaluate the integrity of the gene regions in the whole assembly results.

2.6 | Repeat elements annotation

Annotation of the repetitive sequences in *Q. mongolica* genome was completed using two approaches, homology based and *de novo* prediction. To construct a *de novo* repeat library, the specific transposable elements in our assembly were first screened using RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html), Piler (Edgar & Myers, 2005), RepeatScout (Price et al., 2005), and LTR_FINDER v.1.0.5 (Zhao & Hao, 2007). Then RepeatMasker (Smit & Hubley, 2019) was applied to perform a homology-based repeat search using both the *de novo* repeat library and the known Repbase (Bao et al., 2015) transposable elements database. Tandem repeat sequences were annotated by Tendem Repeat Finder (Gary, 1999).

2.7 | Gene prediction and functional annotation

Three methods (*de novo*, homology-based and Iso-Seq-based predictions) were integrated to annotate protein-coding genes. Augustus v.3.12 (Mario et al., 2004) and GlimmerHMM v.3.03 (Majoros et al., 2004) were utilized for *de novo* prediction of gene structures. For the homology-based prediction, protein sequences from six related plant species were downloaded from public databases, including *Castanea mollissima* (Xing et al., 2019), *Fagus sylvatica* (Mishra et al., 2018), *Q. lobate* (Sork et al., 2016), *Q. suber* (Ramos et al., 2018), *Q. robur* (Plomion et al., 2017), and *Populus trichocarpa* (Tuskan et al., 2006). These data were aligned against the *Q. mongolica* genome using TBLASTN v.2.60 (Gertz et al., 2006). We utilized the method of Minoche et al. (Minoche et al., 2015) to make the prediction based on Iso-Seq. MAKER v.3.0 (Cantarel et al., 2008) was used to generate a non-redundant gene set from the above three approaches. All protein-coding genes were aligned to two integrated

166protein sequence databases: SwissProt and NR. Protein domains were annotated by searching 167against the InterPro v.32.0 using InterProScan (Mulder & Apweiler, 2007), and Pfam v.32.0 168(El-Gebali et al., 2019) databases by HMMER. The Gene Ontology (GO) (Ashburner et al., 1692000) terms for each gene were obtained from the corresponding InterPro or Pfam entries. 170The pathways in which the genes might be involved were assigned by BLAST against the 171KEGG databases (Minoru & Susumu, 2000), with an E-value cutoff of 1e-5. Functional 172annotation results from these two strategies were then merged.

173**2.8 | Non-coding RNA annotation**

174The tRNAscan-SE v.1.3.1 (Peter et al., 2005) was used to evaluate the tRNAs in *Q. mongolica* 175genome. Based on the high degree of conservation of ribosomal RNAs (rRNAs), the rRNAs 176of related species were selected as reference sequences, and the rRNAs in *Q. mongolica* 177genome was searched through BLASTN alignment. Sequences of microRNAs (miRNAs) and 178small nuclear RNAs (snRNAs) in the genome were predicted based on the covariance model 179of the Rfam family, INFERNAL v.1.1 (Nawrocki et al., 2009; Nawrocki & Eddy, 2013) 180software in the Rfam database v.14.1 (Griffiths-Jones et al., 2005).

181**2.9 | Comparative genomics analysis**

182To study the evolutionary relationship between *Q. mongolica* and its related species, we first 183identified and clustered the gene families of these 13 species. The original protein sequences 184of the remaining 12 species (*Q. lobate*, *Q. robur*, *Q. suber*, *Betula pendula* (Salojärvi et al., 1852017), *C. mollissima*, *Eucalyptus grandis* (Alexander A Myburg et al., 2014), *F. sylvatica*,

Juglans regia (Martínez-García et al., 2016), *Malus domestica* (Velasco et al., 2010), *Oryza sativa* (International Rice Genome Sequencing Project, 2005), *P. trichocarpa*, *Vitis vinifera* (Jaillon et al., 2007)) were downloaded from NCBI (Kitts et al., 2016) and GigaDB (Sneddon et al., 2012) databases. Next, the longest transcript protein sequence for each gene was taken, and similarity relationships between these protein sequences across species were determined using BLASTp v.2.2.30 (e-value set to 1e-5) (Camacho et al., 2009). To ensure the comparison quality, those with identities < 30% or coverage < 50% were filtered out. Based on sequence similarity, gene family clustering was completed by the OrthoMCL v.2.0.9 (Li et al., 2003) process (expansion coefficient set to 1.5).

After gene family clustering, single-copy protein sequences were retained if the genes with amino acid length were greater than or equal to 100. MUSCLE v.3.8.31 (Edgar, 2004) was used to perform multiple sequence alignments on the genes in each single-copy homologous gene family. Finally, the multi-sequence alignment results were merged and transformed to a super-gene alignment in PHYLIP format, and RAxML v.8.2.11 (Alexandros, 2002014) was used to construct the evolutionary tree by maximum likelihood method. We also used the constructed evolutionary tree along with the Time Tree (Sudhir et al., 2017) website and associated literatures to obtain time correction points. The software r8s v.1.7.1 (Sanderson, 2003) and the MCMCTREE program (from PAML v.4.9 (Yang, 2007) software packages) were used to estimate the divergence time of these 13 species based on the penalty likelihood method combined with the Bayesian relaxed molecular clock correction method.

CAFE (http://sourceforge.net/projects/cafehahnlab/) (Bie et al., 2006) with the default parameters was used to calculate the expansion and contraction of orthologous gene families

in genomes of the 13 species. Ka and Ks values were calculated for pairs of orthologous genes using the CodeML utility within the PAML software package, and genes under positive selection were detected with the branch-site model.

2.10 | Gene family analysis

BLASTP and HMMER (http://www.hmmer.org/) were used to search for homologous proteins harboring conserved domains of related gene families in the *Q. mongolica* genome using an e-value threshold of <1e-10. The final deduced full-length amino acid sequences were aligned using the MUSCLE program with default parameters. The phylogenetic tree was constructed using RAxML v.8.2.11 and visualized using the Evolview web tool (Subramanian et al., 2019). We performed a pathway enrichment analysis using the KEGG pathway database (http://www.genome.jp/kegg) to identify enriched metabolic or signal transduction pathways associated with plants. The classification of biological terms and the KEGG pathway enrichment analysis were completed with the clusterProfiler R package.

2.11 | Whole genome duplication analyses

Protein sequences of *Q. mongolica*, *Q. robur*, *P. trichocarpa*, and *V. vinifera* were reciprocal aligned using BLASTP with e-value cutoff of 1e-05. For each genome pair, putative paralogous and orthologous genes within and between genomes were searched. 4DTV (fourfold degenerate synonymous sites of the third codons) were extracted from each alignment and concatenated to generate one super-gene for each species. 4DTV values were calculated using in-house Perl scripts. The 4DTV range was determined by plotting the

distribution frequency histogram of 4DTV values.

3 | RESULTS AND DISCUSSION

3.1 | Genome sequencing and assembly

Q. mongolica is an important ecological and garden tree species in East Asia with a karyotype of 2n = 24 chromosomes. We sequenced and assembled its genome using combination of sequencing technologies (FigureS1, Table 1). Based on the 17-Kmer analysis, the *Q. mongolica* genome was estimated to be 842.80 Mb in size with a heterozygosity of 1.09% and repeat of 62.10% (Table S1 and Figure S2), close to that reported for other Fagaceae species (Table S2). A total of 86.9 Gb subreads from PacBio Sequel II were used for the initial contig assembly using FALCON software, which resulted in a total sequence length of 809.83 Mb, with a contig N50 size of 2.64 Mb. The initial contigs were polished with PacBio long reads and Illumina short reads. Subsequently, 102.2 Gb clean reads were obtained from Hi-C sequencing, providing 121.8x coverage of the *Q. mongolica* genome. The polished contigs were assembled into 12 pseudomolecules by LACHESIS software resulting in 774.59Mb (95.65%) sequences distributed on 12 pseudochromosomes with scaffold N50s of 66.74Mb (Figure S3).

To assess the quality of the assembled genome, three indicators were used. First, the final assembled genome size of *Q. mongolica* (809.84 Mb) was similar to the size calculated based on the K−mer frequency distribution (842.80 Mb). Second, the subreads of the Continuous Long Reads (CLR) of *Q. mongolica* were selected and aligned back to the assembled genome

using minimap2 v.2.5. The mapping rate and coverage rate were 97.43% and 99.77%, respectively (Table S3). Third, we performed Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis, and 93.3% of the eukaryotic single-copy genes were detected in the assembled genome which was higher than the most reported genomes of *Quercus* species (Table S4). The high contiguity and quality of the *Q. mongolica* genome will therefore be of great value for further research on the evolution and genomic characteristics of *Quercus* species.

3.2 | Genome annotation

Based on homologous and *de novo* prediction, a total of 554.52 Mb of repetitive elements occupying 68.46% of the *Q. mongolica* genome were annotated including 43.62Mb (5.38% of the genome) of tandem repeat sequences and 531.67 Mb (65.64% of the genome) of transposable elements (Figure S4 and Table S5). The majority of the repeats are long terminal repeats (LTRs) (44.07% of the genome) while the short interspersed nuclear elements (SINEs) made up just 0.03% of the genome. In addition, long interspersed nuclear elements (LINEs) and DNA elements comprised 8.27% and 15.82% of the genome, respectively (Table S5).

We were able to annotate 36,553 protein-coding genes, with an average sequence length of 6,084 bp (Figure S5 and Table S6) based on three methods (*de novo*, homology-based, and Iso-Seq-based predictions). On average, each predicted gene contains 4.82 exons with an average sequence length of 268 bp. Approximately 94.89% of the genes were functionally annotated by similarity searches against homologous sequences and protein domains (Table S7). In addition, we identified noncoding RNA (ncRNA) genes, including 4,896 rRNA, 768

tRNA, 129 miRNA, and 371 snRNA genes (Table S8).

3.3 | Gene family clustering analysis

We performed orthologous clustering of the genes from *Q. mongolica* and 12 related species using OrthoMCL. A total of 453 gene families were specific to the *Q. mongolica* genome, which were related to cellular processes, metabolism and signal transduction based on KEGG enrichment analysis (Figure S6 and Table S9). The complexity of gene families of *Q. mongolica* was then compared with three other *Quercus* species, *Q. lobata*, *Q. robur*, and *Q. suber*. A total of 9,312 gene families were shared by the four *Quercus* species and most of genes were involved in glycan biosynthesis, metabolism, environmental adaptation, and other processes (Figure 2A and Table S10). Only 1,089 gene families were specific to the *Q. mongolica* genome compared with the other three species and these were involved in cellular and environmental information processing and other roles (Figure 2A and Table S11).

3.4 | Phylogenetic analysis

The phylogenetic relationship among *Q. mongolica* and 12 related species was determined using a set of 242 single-copy genes. The results showed that *Q. mongolica* was more closely related to *Q. robur* than to either *Q. lobata* or *Q. suber*. The divergence time between *C. mollissima* and *Quercus* species was estimated to be 20.6 million years ago, while *Q. mongolica* and *Q. robur* separated 10.2 million years ago (Figure S6).

3.5 | Gene family expansion and contraction

During the process of plant evolution, the expansion and contraction of gene families plays an important role in driving phenotypic diversification and enhancing adaptability. To further understand the evolutionary dynamics of *Q. mongolica* genes, the expansion and contraction of orthologous gene families in the genomes of *Q. mongolica* and 12 related species was compared using CAFE based on the default parameters. Based on this analysis, we detected 1,057 gene families that have undergone expansion and 706 gene families that have contracted in *Q. mongolica* (Figure 2B). The expanded gene families were mainly involved in plant-pathogen interactions, linoleic acid metabolism, ABC transporters, and the MAPK signaling pathway (Table S12), while the gene families that contracted were mainly involved in plant-pathogen interactions, phenylpropanoid biosynthesis, Sesquiterpenoid and triterpenoid biosynthesis, and the MAPK signaling pathway (Table S13).

3.6 | Positive selection analysis

To study the adaptive evolution of *Q. mongolica*, we identified positively selected genes based on the results of gene family clustering. Evidence for positive selection was found for 238 genes (FDR < 0.05), of which 34 genes were annotated with potential functions using the Swissprot database (Table S14). Among them, we identified the *LpxB* and *LpxC5* genes encoding key enzymes in the biopolysaccharide biosynthesis pathway (Li et al., 2011), which may be involved in the regulation of plant immune response to pathogens (Newman et al., 1997; Dow et al., 2000; Zeidler et al., 2004; Shang-Guan et al., 2018); the *pyridoxal reductase 1* (*PLR1*) gene encoding a key enzyme in the vitamin B6 metabolic pathway, which plays a key role in resistance to osmotic stress (Herrero et al., 2011); and the *switch subunit 3* (*SWI3*)

gene encoding a positive regulator of ABA signaling (Saez et al., 2008) that participates in resistance to abiotic stress. The positive selection of these genes may reflect *Q. mongolica*'s outstanding resistance to disease, cold, and drought, and accordingly, this knowledge may be helpful in guiding future improvements in related species.

3.7 | Wood formation genes in *Q. mongolica*

Wood formation is an important biological process occurring in woody plants, so here we focused on analyzing the genes involved. Wood mainly consists of cellulose, hemicellulose, and lignin. In *Q. mongolica*, we annotated 30 gene families involved in cell wall formation, including 19 gene families involved in cellulose and hemicellulose biosynthesis, and 11 gene families involved in lignin synthesis (Table S15). According to statistics, a total of 403 genes were related to cellulose and hemicellulose biosynthesis, including 34 *cellulose synthase-like* (*CSL*), 247 *glycoside hydrolase* (*GH*), 20 *glucan synthase−like* (*GSL*), 38 *glycosyltransferase family 8* (*GT8*), 9 *reversibly glycosylated polypeptides* (*RGP*), 5 *xyloglucan fucosyltransferase* (*XFT*), 45 *xyloglucan glucosyltransferase* (*XGT*) and 5 *xyloglucan xylosyltransferase* (*XXT*) genes. In addition, a total of 93 genes were involved in lignin synthesis, including 4 *4-coumarate:CoA ligase* (*4CL*), 2 *p-coumarate 3-hydroxylase* (*C3H*), 3 *Trans-cinnamate-4-hydroxylase* (*C4H*), 17 *cinnamyl alcohol dehydrogenase* (*CAD*), 16 *caffeoyl-CoA 3-O-methyltransferase* (*CCoAOMT*), 4 *cinnamoyl CoA reductase* (*CCR*), 11 *caffeic acid O-methyltransferase* (*COMT*), 4 *ferulate 5-hydroxylase* (*F5H*), 1 *hydroxycinnamoyl-Coenzyme A shikimate/quinate hydroxycinnamoyltransferase* (*HCT*), 26 *Laccase* and 5 *phenylalanine ammonia-lyase* (*PAL*) genes. Simultaneously, we also annotated

330403 genes and 402 genes related to cellulose and hemicellulose biosynthesis in *Q. robur* and
331*P. trichocarpa*, as well as 87 and 86 genes involved in lignin biosynthesis. Compared with *A.*
332*thaliana,* these three woody plants had more genes involved in lignin biosynthesis and fewer
333genes involved in cellulose and hemicellulose biosynthesis, which was consistent with the
334botanical classification and use of these species. In these three woody plants, the total number
335of genes in gene families was similar, but the number of genes in each gene family was
336different. Phylogenetic analysis showed that there were more *GSL*, *XGT* genes in *Q.*
337*mongolica* than that in *Q. robur* and *P. trichocarpa*, more *CSL* and *RGP* genes in *Q.robur*
338than in *Q. mongolica* and *P. trichocarpa* (Figure 3 and Table S15). For the number of genes
339in the *GT8* gene family, more in *P. trichocarpa* than in *Q. mongolica* and *Q.robur* (Figure
340S8). For *GH* orthologs, *Q. mongolica* contained the largest number of *GH3*, *GH5*, *GH10* and
341*GH35* genes, *Q.robur* contained more *GH18* and *GH27* genes, while *P. trichocarpa* had more
342*GH9*, *GH16*, *GH17* and *GH28* genes than *Q. mongolica* and *Q.robur* (Figure S9). The wood
343of *Q. mongolica* and *Q. robur* is harder than that of *P. trichocarpa*, which are excellent
344materials for making vehicles, ships, buildings and furniture. The total number of genes
345involved in lignin synthesis of *Q. mongolica* and *Q. robur* was slightly more than that of *P.*
346*trichocarpa*, especially *CCoAOMT* and *COMT* genes, while the opposite trend was observed
347for *CCR* genes (Figure S10). In conclusion, although wood formation processes are common
348in woody plants, each species has its own unique formation process, which also leads to
349differences in final products.

350**3.8 | Transcription factor analysis**

Transcription factors (TFs) are molecules involved in regulating gene expression which is vital for the normal development of an organism, as well as for routine cellular functions and response to disease. The WRKY transcription factors (TFs) are one of the largest families in higher plants and are found throughout the green lineage (Ulker & Somssich, 2004). They are involved in plant defense regulatory networks, including response to various biotic and abiotic stresses. A total of 88 WRKY genes were annotated in *Q. mongolica*, which was more than the ones found in *Arabidopsis thaliana.* The phylogenetic analysis (Figure 4A) indicated that these genes can be divided into three large groups corresponding to the group I, II and III, as was first defined in *Arabidopsis thaliana* by Eulgem et al. (Eulgem et al., 2000).

NAC proteins constitute one of the largest families of plant-specific transcription factors, and the family is present in a wide range of land plants. The structure of NAC transcription factors is distinct and their functions are diverse. NAC transcription factors have a variety of important functions not only in plant development but also in abiotic stress tolerance. A total of 124 genes were annotated in *Q. mongolica,* which is more than the ones found in *A. thaliana.* Phylogenetic analyses indicate that six major groups of NAC transcription factors were already established in *Q. mongolica* (Figure 4B).

3.9 | Whole-genome duplication analysis

Paralogous gene pairs within the genomes of *Q. mongolica*, *Q. robur*, *P. trichocarpa* and *V. vinifera* were detected based on their protein sequences and the fourfold synonymous third-codon transversion (4DTV) value for each gene pair was calculated. Plotting the 4DTV values for the paralogous gene pairs revealed two peaks in both *Q. mongolica* and *Q. robur* (Figure

3725) indicating that *Q. mongolica* and *Q. robur* not experienced lineage-specific whole-genome 373duplication expect for the ancestral triplication shared among the eudicots (γ) which were 374consistent with a previous report (Plomion et al., 2017). The peak appeared at 4DTV values of 3750.4770, 0.4375, 0.0758 and 0.4186 in *Q. mongolica*, *Q. robur*, *P. trichocarpa* and *V. vinifera*, 376respectively. The results show that the time of whole-genome duplication in *Q. mongolica* 377was earlier than those in *Q. robur*, *P. trichocarpa* and *V. vinifera*.

3784 | CONCLUSIONS

379Here we assembled a chromosome-level reference genome of *Q. mongolica* based on the 380combination strategy of Illumina short-read, PacBio long-read, and Hi-C sequencing 381technology. This is the first report of the high-quality genome of Asian oak. Details of genome 382structure and function provide further insights into the phylogenetic diversity of oaks. This 383genome not only provides an important resource for revealing the biological characteristics 384and evolutionary adaptability of *Q. mongolica*, but also contributes to the study of the 385taxonomy, evolution and conservation of *Quercus* species.

386ACKNOWLEDGEMENTS

387The project was funded by the national key research and development program of Ministry of 388Science and Technology, PRC (2017TFD0600602). We also appreciate Frasergen for the 389technical support of this project.

CONFLICT OF INTEREST

The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

X.J.L. conceived and designed this study. M.M. collected the samples. W.F.A., X.Y.H. and H.Z. performed laboratory work (DNA and RNA extraction, library construction and sequencing). W.F.A. and M.M. performed bioinformatics analyses. W.F.A. wrote the manuscript and X.L.Z., L.J.Z. and X.J.L. revised it. All authors reviewed and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The genome assembly of the Mongolian oak (*Quercus mongolica*) has been submitted to GenBank under the accession JAAMOV000000000. Raw sequencing reads and genome assembly are available at GenBank as BioProject PRJNA609556 and PRJNA607679. Raw sequencing data (Illumina, PacBio, Hi-C and Iso-Seq data) have been deposited in SRA (Sequence Read Archive) database as SRR11093900 *[dataset]* (SYAU, 2020a), SRR11119092 *[dataset]* (SYAU, 2020b), SRR11096897 *[dataset]* (SYAU, 2020c), and SRR11119237 *[dataset]* (SYAU, 2020d). Genome assembly has been deposited in assembly database as GCA_011696235.1 *[dataset]* (SYAU, 2020e). Annotation files of repetitive sequences, protein-coding genes, non-coding RNA genes, and protein functions were

availabled within FigShare *[dataset]* (SYAU, 2020f) (https://doi.org/10.6084/m9.figshare.11888118.v4).

ORCID

Xiujun Lu https://orcid.org/0000-0003-1806-6414

Wanfeng Ai https://orcid.org/0000-0001-6074-0543

SUPPORTING INFORMATION

Figure S1. A workflow for the genome assembly, annotation and comparative genomics analysis.
Figure S2. K-mer analysis for estimating the genome size of *Q. mongolica*.
Figure S3. Hi-C contact data mapped on the *Q. mongolica* genome showing genome-wide all-by-all interactions.
Figure S4. Characteristics of repetitive elements in the *Q. mongolica* genome.
Figure S5. Gene structure prediction results statistics.
Figure S6. Orthologous genes in *Q. mongolica* and 12 species.
Figure S7. Differentiation time between *Q. mongoica* and 12 species
Figure S8 Phylogenetic analysis of *glucosyltransferase 8* (*GT8*) in *Q. mongolica*, *Q. robur*, *P. trichocarpa* and *A.thaliana.*
Figure S9. Phylogenetic analysis of *glycoside hydrolase* (*GH*) genes in *Q. mongolica*, *Q. robur*, *P. trichocarpa* and *A.thaliana.*
Figure S10. Phylogenetic analysis of genes involved in lignin synthesis in *Q. mongolica*, *Q. robur*, *P. trichocarpa* and *A.thaliana.*
Table S1. Survery statistic results of *Q. mongolica* genome.
Table S2. Summary of genome survey information for Fagaceae species.
Table S3. Coverage statistics of *Q. mongolica* genome with continuous long reads (CLR) subreads.
Table S4. Comparison of genome assembly quality for four *Qucrcus* species.
Table S5. Summary of repeat contents in the assembled *Q. mongolica* genome.
Table S6. Basic statistical results of gene structure prediction of *Q. mongolica* genome.
Table S7. The statistical results of gene function annotation of *Q. mongolica* genome.
Table S8. The statistical results of non-coding RNA in *Q. mongolica* genome.
Table S9. KEGG pathway enrichment of species-specific genes in *Q. mongolica* compared with 12 related species.
Table S10. KEGG pathway enrichment of shared genes in four *Quercus* species.
Table S11. KEGG pathway enrichment of species-specific genes in *Q. mongolica* compared

with other three *Quercus* species.

Table S12. KEGG pathway enrichment of significant expansion *Q. mongolica* of gene families.

Table S13. KEGG pathway enrichment of significant contraction *Q. mongolica* of gene families.

Table S14. The function of positive selection genes of *Q. mongolica*.

Table S15. Gene families related to wood formation in *Q. mongolica*, *Q. robur*, *P. trichocarpa* and *A. thaliana*.

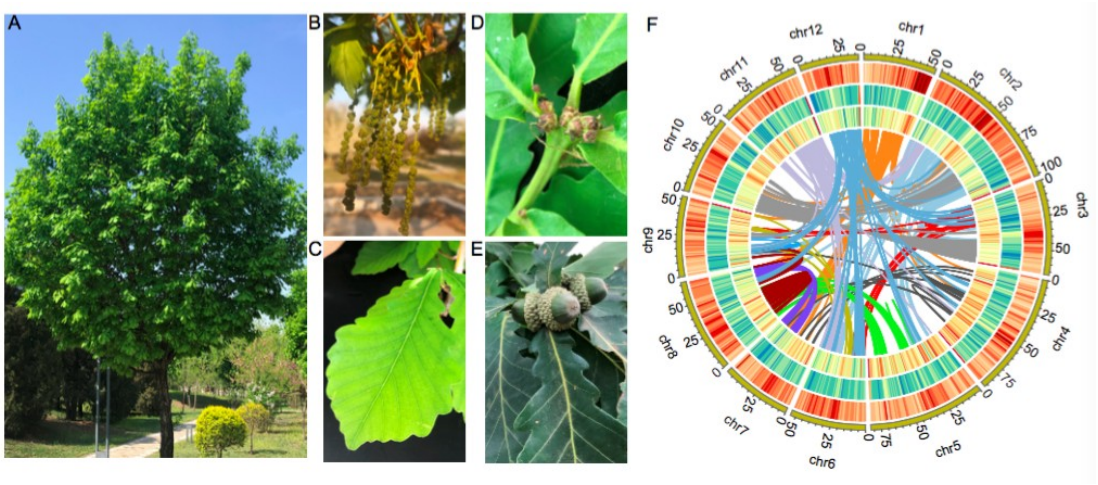
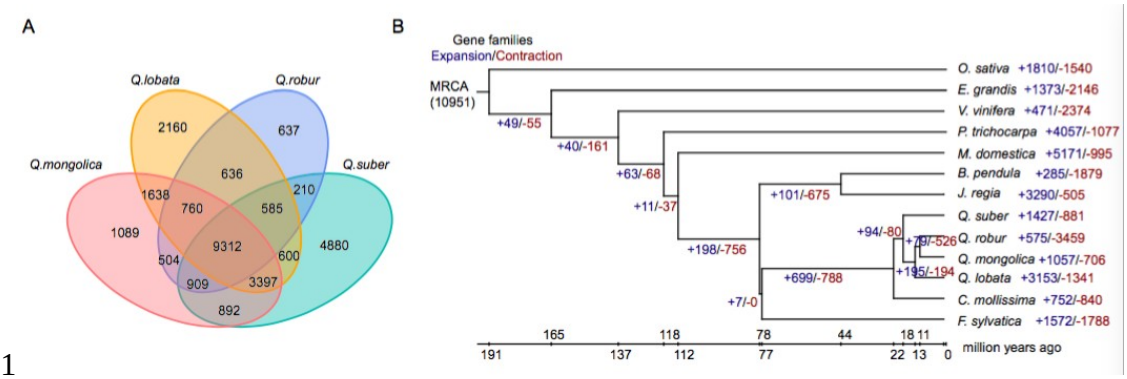


FIGURE 1 Photograph and genome features of the *Quercus mongolica.* Photograph of the *Q. mongolica* (A) and its male flowers (B), leaf (C), female flowers (D) and fruits (E) (photo: Wanfeng Ai). (F) The chromosomal features of the *Quercus mongolica* were Chromosome size with units in Mb, GC content, TE density and gene density from outer to inner rings. Lines in the center linking different chromosomal regions show the syntenic blocks on homologous chromosomes.



461

462**FIGURE 2** The Gene family clustering in *Quercus* species and the expansion and contraction
463of gene families. (A) Venn diagram of the protein-coding orthologues shared among *Quercus*
464*lobata*, *Quercus robur*, *Quercus suber* and *Quercus mongolica*. Each number represents the
465number of gene families; (B) The expansion and contraction of gene families for 13 plants.
466The blue number indicates the number of expanded gene families, while the red number
467indicates the number of contracted gene families.

468

469

45
46

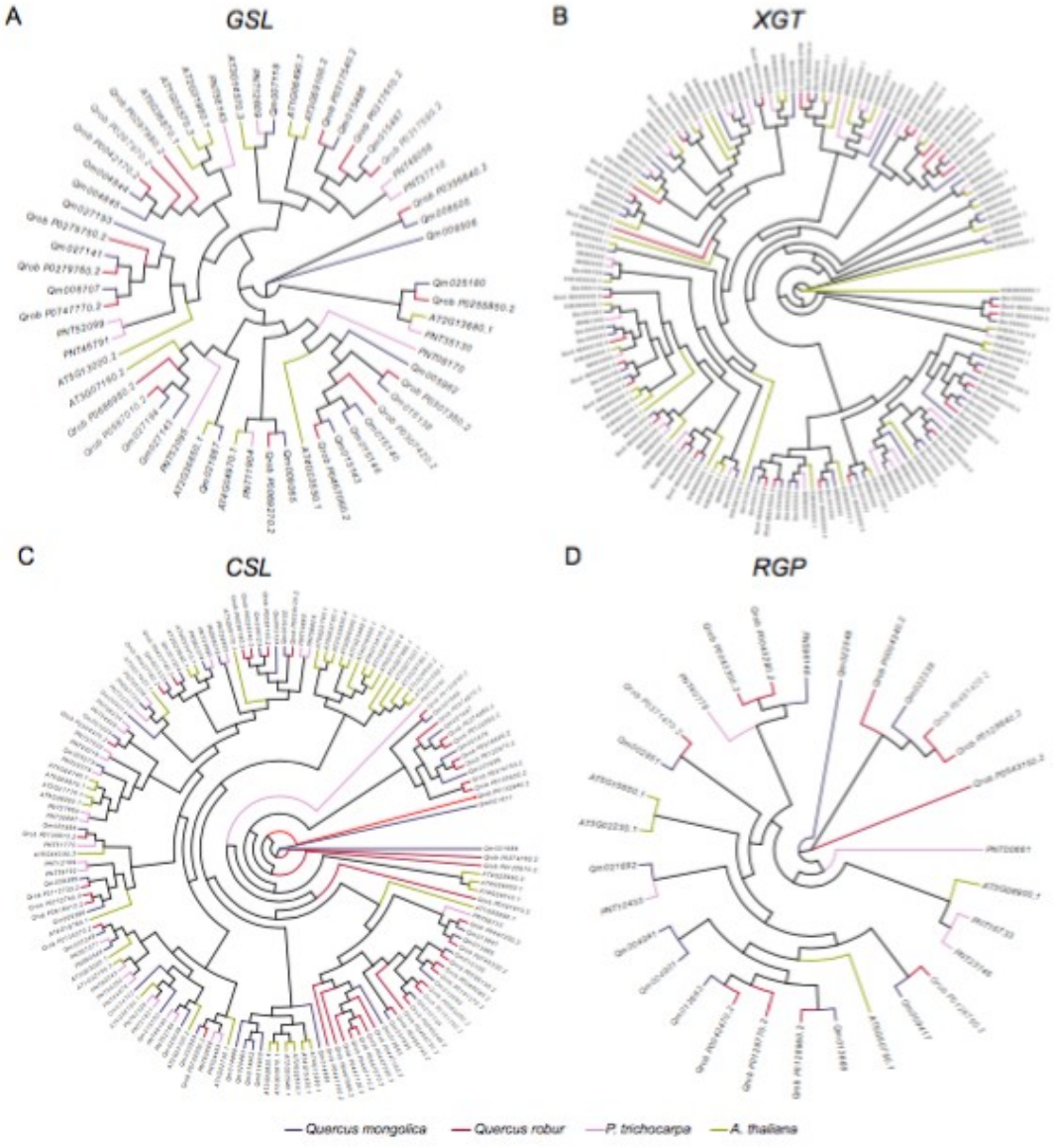
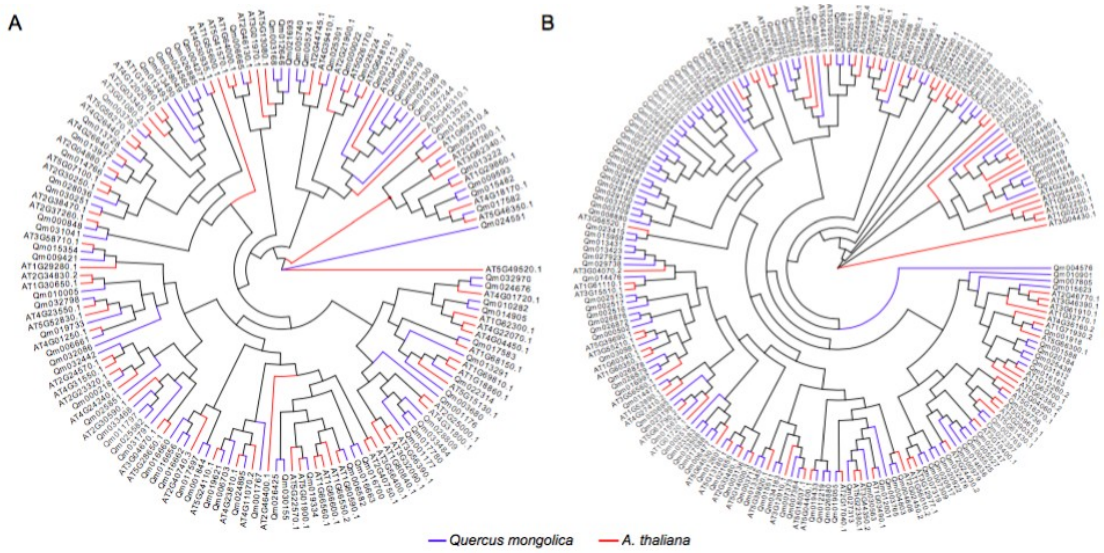


FIGURE 3 Phylogenetic analysis of genes involved in cellulose and hemicellulose biosynthesis in *Quercus mongolica*, *Quercus robur*, *Populus trichocarpa* and *Arabidopsis thaliana*. (A) *Glucan synthase-like* (*GSL*); (B) *Xyloglucan galactosyltransferase* (*XGT*); (C) *Cellulose synthase-like* (*CSL*); (D) *Reversibly glycosylated polypeptides* (*RGP*).



477

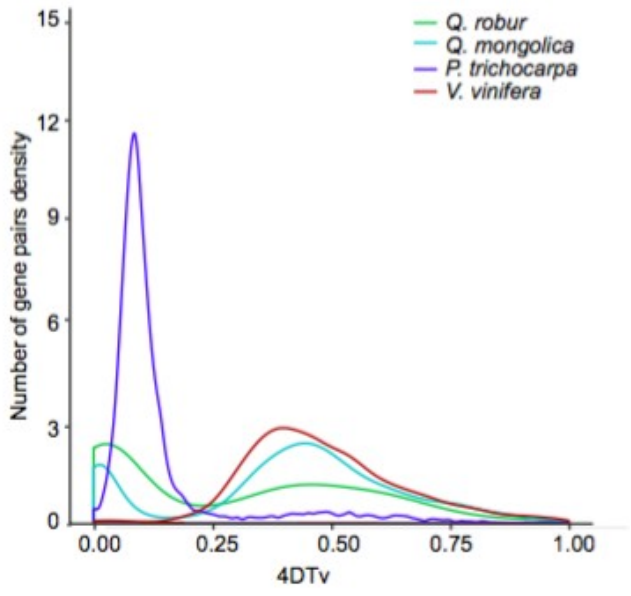
478**FIGURE 4** Phylogenetic tree showing the evolutionary relationship of WRKY and NAC
479genes in *Quercus mongolica* (blue) and *Arabidopsis thaliana* (red). (A) WRYK; (B) NAC.

480

481

482

483



484

485**FIGURE 5** Four-fold synonymous third-codon transversion (4DTV) for each homologous
486gene pair for *Quercus mongolica*, *Quercus robur*, *Populus trichocarpa* and *Vitis vinifera*.

487

49
50

Table 1 Summary of genome and transcriptome sequencing information for *Quercus mongolica.*

Library type	Sequencing platform	Insert size	Clean data (Gb)	Sequence coverage (×)
DNA library	Illumina HiSeq X Ten	350bp	105.20	124.80
	PacBio Sequel II	20kb	86.90	103.10
Hi-C	Illumina HiSeq X Ten	300-500bp	102.20	121.80
RNA library	PacBio Sequel II	4kb	20.70	25.00

Table 2 Statistics for assembled genome information of *Quercus mongolica*.

Category	*Q. mongolica* genome
Estimate of genome size(Mb)	842.80
Assembly size (Mb)	809.83
Total number of contigs	735
Contig N50 length (bp)	2,446,788
Total number of scaffold	321
Scaffold N50 length(bp)	66,735,633
GC content (%)	35.84
% sequence anchored on chromosome	95.65
Number of protein-coding genes	36,533
Repeat content (%)	68.46

References:

Alexander A Myburg D. G. G. A., van der Merwe K., Singh P., van Jaarsveld I., Silva-Junior O. B., Togawa R. C., ... Schmutz J. (2014). The genome of Eucalyptus grandis. *Nature, 510* (7505), 356-362. doi: 10.1038/nature13308

Alexandros S. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30* (9), 1312-1313. doi: 10.1093/bioinformatics/btu033

Ao T., Yang Q., Mi L., Shi H. & Zhao Z. (1998). An Analysis of Vitamin Contents and Fatty Acid Composition of Acorn of Quercus Mongolia Fisch. *Journal of Inner Mongolia Institute of Agriculture & Animal Husbandry,* (3), 3-5.

Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., ... Sherlock G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics, 25* (1), 25-29. doi: 10.1038/75556

Bao W., Kojima K. K. & Kohany O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA, 6* 11. doi: 10.1186/s13100-015-0041-9

Bie T. D., Cristianini N., Demuth J. P. & Hahn M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics, 22* (10), 1269-1271. doi: 10.1093/bioinformatics/btl097

Burton J. N., Adey A., Patwardhan R. P., Qiu R., Kitzman J. O. & Shendure J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology, 31* (12), 1119. doi: 10.1038/nbt.2727

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., ... Madden T. L. (2009). BLAST+: architecture and applications. *Bmc Bioinformatics, 10* 421. doi: 10.1186/1471-2105-10-421

Cannon C. H., Brendel O., Deng M., Hipp A. L., Kremer A., Kua C., ... Sork V. L. (2018). Gaining a global perspective on Fagaceae genomic diversification and adaptation. *New Phytologist, 218* (3), 894-897. doi: 10.1111/nph.15101

Cantarel B. L., Korf I., Robb S. M. C., Parra G., Ross E., Moore B., ... Yandell M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research, 18* (1), 188-196. doi: 10.1101/gr.6743907

Cavender-Bares J. (2019). Diversification, adaptation, and community assembly of the American oaks (Quercus), a model clade for integrating ecology and evolution. *New Phytologist, 221* (2), 669-692. doi: 10.1111/nph.15450

Cavender-Bares J., Gonzalez-Rodriguez A., Eaton D. A. R., Hipp A. A. L., Beulke A. & Manos P. S. (2015). Phylogeny and biogeography of the American live oaks (Quercus subsection Virentes): a genomic and population genetics approach. *Molecular Ecology, 24* (14), 3668-3687. doi: 10.1111/mec.13269

Chen H. & Huang C. (1998). *Flora Reipublicae Popularis Sinicae*. Beijing: Science Press.

Chin C. S., Alexander D. H., Marks P., Klammer A. A., Drake J., Heiner C., ... Korlach J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods, 10* (6), 563-569. doi: 10.1038/nmeth.2474

Dow M., Newman M. A. & Roepenack E. V. (2000). The Induction and Modulation of Plant Defense Responses by Bacterial Lipopolysaccharides. *Annual Review of Phytopathology, 38* 241-261. doi: 10.1146/annurev.phyto

Durand N. C., Robinson J. T., Shamim M. S., Machol I., Mesirov J. P., Lander E. S., ... Aiden E. L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems, 3* (1), 99-101. doi: 10.1016/j.cels.2015.07.012

Durand N. C., Shamim M. S., Machol I., Rao S. S. P., Huntley M. H., Lander E. S., ... Aiden E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems, 3* (1), 95-98. doi: 10.1016/j.cels.2016.07.002

Edgar R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32* (5), 1792-1797. doi: 10.1093/nar/gkh340

Edgar R. C. & Myers E. W. (2005). PILER: identification and classification of genomic repeats. *Bioinformatics, 21 Suppl 1* i152-i158. doi: 10.1093/bioinformatics/bti1003

El-Gebali S., Mistry J., Bateman A., Eddy S. R., Luciani A., Potter S. C., ... Finn R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research, 47* (D1), D427-D432. doi: 10.1093/nar/gky995

Eulgem T., Rushton P. J., Robatzek S. & Somssich I. E. (2000). The WRKY superfamily of plant transcription factors. *Trands in plant science, 5* (5), 199-206. doi: 10.1016/s1360-1385(00)01600-9

Gary B. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research, 27* (2), 573-580. doi: 10.1093/nar/27.2.573

Gertz E. M., Yu Y., Agarwala R., Schäffer A. A. & Altschul S. F. (2006). Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *Bmc Biology, 4* 41. doi: 10.1186/1741-7007-4-41

Griffiths-Jones S., Moxon S., Marshall M., Khanna A. & Bateman A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research, 33* D121-D124. doi: 10.1093/nar/gki081

Hao Y. & Yang C. (2016). The Value and Effects of Quercus mongolica. *Forest By-Product and Speciality in*

China, (5), 97-98.

Herrero S., González E., Gillikin J. W., Vélëz H. & Daub M. E. (2011). Identification and characterization of a pyridoxal reductase involved in the vitamin B6 salvage pathway in Arabidopsis. *Plant Molecular Biology, 76* (1-2), 157-169. doi: 10.1007/s11103-011-9777-x

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature, 436* (7052), 793-800. doi: 10.1038/nature03895

Ishimaru K., Nonaka G. I. & Nishioka I. (1987). Phenolic glucoside gallates from quercus mongolica and q. acutissima. *Phytochem, 26* (4), 1147-1152.

Jaillon O., Aury J. M., Noel B., Policriti A., Clepet C., Casagrande A., ... French-Italian Public Consortium For Grapevine Genome Characterization (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature, 449* (7161), 463-467. doi: 10.1038/nature06148

Jiang Y., Liu W., Wang G., Zhou X. & Qin L. (2019). Research Advances in Germplasm Resource and Utilization of Quercus L. *Ence of Sericulture,* (4), 577-585.

Kim H. H., Kim D. H., Oh M. H., Park K. J., Heo J. H. & Lee M. W. (2015). Inhibition of matrix metalloproteinase-1 and type-I procollagen expression by phenolic compounds isolated from the leaves of Quercus mongolica in ultraviolet-irradiated human fibroblast cells. *Archives of Pharmacal Research, 38* (1), 11-17. doi: 10.1007/s12272-014-0329-1

Kitts P. A., Church D. M., Thibaud-Nissen F., Jinna C., Vichet H., Victor S., ... Avi K. (2016). Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Research, 44* (D1), D73-D80. doi: 10.1093/nar/gkv1226

Kremer A. & Hipp A. L. (2019). Oaks: an evolutionary success story. *New Phytologist, 226* (4), 987-1011. doi: 10.1111/nph.16274

Lesur I. L. P. G., Léger V., Amselem J., Belser C., Quesneville H., Stierschneider M., ... Plomion C. (2015). The oak gene expression atlas: insights into Fagaceae genome evolution and the discovery of genes regulated during bud dormancy release. *Bmc Genomics, 16* (1), 112. doi: 10.1186/s12864-015-1331-9

Li C., Guan Z., Liu D. & Raetz C. R. H. (2011). Pathway for lipid A biosynthesis in Arabidopsis thaliana resembling that of Escherichia coli. *PNAS, 108* (28), 11387-11392. doi: 10.1073/pnas.1108840108

Li H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34* (18), 3094-3100. doi: 10.1093/bioinformatics/bty191

Li H. & Durbin R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics, 25* (14), 1754-1760. doi: 10.1093/bioinformatics/btp324

Li L., Jr C. J. S. & Roos D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research, 13* (9), 2178-2189. doi: 10.1101/gr.1224503

Li W. (2003). *Study on genetic diversity of natural populations in Quercus mongolica.* (Doctoral thesis, Beijing Forestry University, Beijing, China).

Liao W. J., Zhu B. R., Li Y. F., Li X. M., Zeng Y. F. & Zhang D. Y. (2019). A comparison of reproductive isolation between two closely related oak species in zones of recent and ancient secondary contact. *Bmc Evolutionary Biology, 19* (1), 70. doi: 10.1186/s12862-019-1399-y

Liu B., Shi Y., Yuan J., Hu X., Zhang H., Li N., ... Fan W. (2013). Estimation of genomic characteristics by analyzing k-mer
frequency in de novo genome projects. *arXiv, 1308.2012*

Majoros W., Pertea M. & Salzberg S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics, 20* (16), 2878-2879. doi: 10.1093/bioinformatics/bth315

Mario S., Rasmus S., Stephan W. & Burkhard M. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research,* (32), W309-W312. doi: 10.1093/nar/gkh379

Martínez-García P. J., Crepeau M. W., Puiu D., Gonzalez-Ibeas D., Whalen J., Stevens K. A., ... Neale D. B. (2016). The walnut (Juglans regia) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *Plant Journal, 87* (5), 507-532. doi: 10.1111/tpj.13207

Mi L., Su R., Zhang J. & Ao T. (1999). Quercus Mongolica Fisch. Nutrient Composition and Deprivation of Its Poison. *Journal of Neimenggu Forestry College, 2* (1), 72-75.

Min K., Yin J., Hwang I. H., Park D. H., Lee E. K., Kim M. J., ... Lee M. W. (2020). Anti-Acne Vulgaris Effects of Pedunculagin from the Leaves of Quercus mongolica by Anti-Inflammatory Activity and 5α-Reductase Inhibition. *Molecules, 25* (9), 2154. doi: 10.3390/molecules25092154

Minoche A. E., Dohm J. C., Schneider J., Holtgräwe D., Viehöver P., Montfort M., ... Himmelbauer H. (2015). Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. *Genome Biology, 16* (1), 184. doi: 10.1186/s13059-015-0729-7

Minoru K. & Susumu G. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research, 28* (1), 27-30. doi: 10.1093/nar/28.1.27

Mishra B., Gupta D. K., Pfenninger M., Hickler T., Langer E., Nam B., ... Thines M. (2018). A reference genome of the European Beech (Fagus sylvatica L.). *GigaScience, 7* (6), y63. doi: 10.1093/gigascience/giy063

Mulder N. & Apweiler R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods in Molecular Biology, 396* 59-70. doi: 10.1007/978-1-59745-515-2_5

Nagamitsu T., Uchiyama K., Izuno A., Shimizu H. & Nakanishi A. (2019). Environment-dependent introgression from Quercus dentata to a coastal ecotype of Quercus mongolica var. crispula in northern Japan. *New Phytologist, 226* (4), 1018-1028. doi: 10.1111/nph.16131

Nawrocki E. P. & Eddy S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics, 29* (22), 2933-2935. doi: 10.1093/bioinformatics/btt509

Nawrocki E. P., Kolbe D. L. & Eddy S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics, 25* (10), 1335-1337. doi: 10.1093/bioinformatics/btp157

Newman M. A., Daniels M. J. & Dow J. M. (1997). The activity of lipid A and core components of bacterial lipopolysaccharides in the prevention of the hypersensitive response in pepper. *Molecular plant-microbe interactions, 10* (7), 926-928. doi: 10.1094/MPMI.1997.10.7.926

Omar M., Matsuo Y., Maeda H., Saito Y. & Tanaka T. (2013). New ellagitannin and galloyl esters of phenolic glycosides from sapwood of Quercus mongolica var. crispula (Japanese oak). *Phytochemistry Letters, 6* (3), 486-490. doi: 10.1016/j.phytol.2013.06.004

Pang X., Liu H., Wu S., Yuan Y., Li H., Dong J., ... Li B. (2019). Species Identification of Oaks (Quercus L., Fagaceae) from Gene to Genome. *International Journal of Molecular Sciences, 20* (23), 5940. doi: 10.3390/ijms20235940

Peter S., Brooks A. N. & Lowe T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research, 33* W686-W689. doi: 10.1093/nar/gki366

Plomion C., Aury J., Amselem J., Alaeitabar T., Barbe V., Belser C., ... Kremer A. (2016). Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. *Molecular Ecology Resources, 16* (1), 254-265. doi: 10.1111/1755-0998.12425

Plomion C., Aury J., Amselem J., Leroy T., Murat F., Duplessis S., ... Salse J. (2017). Oak genome reveals facets of long lifespan. *Nature Plants, 4* (7), 440-452. doi: 10.1038/s41477-018-0172-3

Price A. L., Jones N. C. & Pevzner P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics, 21 Suppl 1* i351-i358. doi: 10.1093/bioinformatics/bti1018

Ramos A. M., Usié A., Barbosa P., Barros P. M., Capote T., Chaves I., ... Gonçalves S. (2018). The draft genome sequence of cork oak. *Scientific Data, 5* 180069. doi: 10.1038/sdata.2018.69

Roach M. J., Schmidt S. A. & Borneman A. R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *Bmc Bioinformatics, 19* (1), 460. doi: 10.1186/s12859-018-2485-7

Saez A., Rodrigues A., Santiago J., Rubio S. & Rodriguez P. L. (2008). HAB1-SWI3B interaction reveals a link between abscisic acid signaling and putative SWI/SNF chromatin-remodeling complexes in Arabidopsis. *Plant Cell, 20* (11), 2972-2988. doi: 10.1105/tpc.107.056705

Salojärvi J., Smolander O. P., Nieminen K., Rajaraman S., Safronov O., Safdari P., ... Kangasjärvi J. (2017). Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature Genetics, 49* (6), 904-912. doi: 10.1038/ng.3862

Sanderson M. J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics, 19* (2), 301-302. doi: 10.1093/bioinformatics/19.2.301

Shang-Guan K., Wang M., Myint Phyu Sin Htwe N., Li P., Li Y., Qi F., ... Liang Y. (2018). Lipopolysaccharides Trigger Two Successive Bursts of Reactive Oxygen Species at Distinct Cellular Locations. *Plant Physiology, 176* (3), 2543-2556. doi: 10.1104/pp.17.01637

Simão F. A., Waterhouse R. M., Ioannidis P., Kriventseva E. V. & Zdobnov E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31* (19), 3210-3212. doi: 10.1093/bioinformatics/btv351

RepeatMasker (Version 4.09). Seattle: Institute for Systems Biology. Retried from http://www.repeatmasker.org/.

Sneddon T. P., Li P. & Edmunds S. C. (2012). GigaDB: announcing the GigaScience database. *GigaScience,1,1(2012-07-12), 1* (1), 11. doi: 10.1186/2047-217X-1-11

Sork V. L., Fitz-Gibbon S. T., Puiu D., Crepeau M., Gugger P. F., Sherman R., ... Salzberg S. L. (2016). First Draft Assembly and Annotation of the Genome of a California Endemic Oak Quercus lobata Née (Fagaceae). *G3 (Bethesda), 6* (11), 3485-3495. doi: 10.1534/g3.116.030411

Subramanian B., Gao S., Lercher M. J., Hu S. & Chen W. H. (2019). Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Research, 47* (W1), W270-W275. doi: 10.1093/nar/gkz357

Sudhir K., Glen S., Michael S. & Blair H. S. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology & Evolution, 34* (7), 1812-1819. doi: 10.1093/molbev/msx116

SYAU (2020a). Genome sequencing of Quercus mongolica. GenBank, GenBank accession number: SRR11093900.

SYAU (2020b). Qm_Genome _PacBio. GenBank, GenBank accession number: SRR11119092.

SYAU (2020c). Hi-C assisted assembly sequencing of Quercus mongolica. GenBank, GenBank accession number: SRR11096897.

SYAU (2020d). Qm_transcriptome_PacBio. GenBank, GenBank accession number: SRR11119237.

SYAU (2020e). ASM1169623v1. GenBank, GenBank accession number: GCA_011696235.1.

SYAU (2020f). Annotation information of Quercus mongolica genome. FigShare: 11888118.v4.

Tuskan G. A., Difazio S., Jansson S., Bohlmann J., Grigoriev I., Hellsten U., ... Rokhsar D. (2006). The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). *Science, 313* (5793), 1596-1604. doi: 10.1126/science.1128691

Ulker B. & Somssich I. E. (2004). WRKY transcription factors: from DNA binding towards biological function. *Current Opinion in Plant Biology, 7* (5), 491-498. doi: 10.1016/j.pbi.2004.07.012

Velasco R., Zharkikh A., Affourtit J., Dhingra A., Cestaro A., Kalyanaraman A., ... Viola R. (2010). The genome of the domesticated apple (Malus × domestica Borkh.). *Nature Genetics, 42* (10), 833-839. doi: 10.1038/ng.654

Walker B. J., Abeel T., Shea T., Priest M., Abouelliel A., Sakthikumar S., ... Earl A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One, 9* (11), e112963. doi: 10.1371/journal.pone.0112963

Watanabe M., Hoshika Y., Inada N. & Koike T. (2018). Photosynthetic activity in relation to a gradient of leaf nitrogen content within a canopy of Siebold's beech and Japanese oak saplings under elevated ozone. *Ence of the Total Environment, 636* 1455-1462. doi: 10.1016/j.scitotenv.2018.04.423

Watanabe Y., Satomura T., Sasa K., Funada R. & Koike T. (2010). Differential anatomical responses to elevated CO2 in saplings of four hardwood species. *Plant Cell & Environment, 33* (7), 1101-1111. doi: 10.1111/j.1365-3040.2010.02132.x

Xie T., Zheng J. F., Liu S., Peng C., Zhou Y. M., Yang Q. Y., ... Zhang H. Y. (2015). De Novo Plant Genome Assembly Based on Chromatin Interactions: A Case Study of Arabidopsis thaliana. *Molecular Plant, 8* (3), 489-492. doi: 10.1016/j.molp.2014.12.015

Xing Y., Liu Y., Zhang Q., Nie X., Sun Y., Zhang Z., ... Qin L. (2019). Hybrid de novo genome assembly of Chinese chestnut (Castanea mollissima). *GigaScience, 8* (9), z112. doi: 10.1093/gigascience/giz112

Yang Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology & Evolution, 24* (8), 1586-1591. doi: 10.1093/molbev/msm088

Yin J., Kim H. H., Hwang I. H., Kim D. H. & Lee M. W. (2019). Anti-Inflammatory Effects of Phenolic Compounds Isolated from Quercus Mongolica Fisch. ex Ledeb. on UVB-Irradiated Human Skin Cells. *Molecules, 24* (17), 3094. doi: 10.3390/molecules24173094

Zeidler D., Zahringer U., Gerber I., Dubery I., Hartung T., Bors W., ... Durner J. (2004). Innate immunity in Arabidopsis thaliana: Lipopolysaccharides activate nitric oxide synthase (NOS) and induce defense genes. *PNAS, 101* (44), 15811-15816. doi: 10.1073/pnas.0404536101

Zeng Y. F., Wang W. T., Liao W. J., Wang H. F. & Zhang D. Y. (2016). Multiple glacial refugia for cool-temperate deciduous trees in northern East Asia: the Mongolian oak as a case study. *Molecular Ecology, 24* (22), 5676-5691. doi: 10.1111/mec.13408

Zhang H., Mcdowell N. G., Adams H. D., Wang A., Wu J., Jin C., ... Guan D. (2020). Divergences in hydraulic conductance and anatomical traits of stems and leaves in three temperate tree species coping with drought, N addition and their interactions. *Tree Physiology, 40* (2), 230-244. doi: 10.1093/treephys/tpz135

Zhao X. & Hao W. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research, 35* W265-W268. doi: 10.1093/nar/gkm286

Zhou J., Wang B., Wang G., Jiang Y., Yang R., Shi S., ... Li Q. (2017). Identification of Volatile Chemical Components in Leaves and Barks of Two Types of Oak Trees. *Science of Sericulture,* (3), 459-466.