Plant Communications
Research Article

CellPress
Partner Journal

Chromosome-scale assembly reveals asymmetric paleo-subgenome evolution and targets for the acceleration of fungal resistance breeding in the nut crop, pecan

Lihong Xiao[1,4,*], Mengjun Yu[2,4], Ying Zhang[1,4], Jie Hu[2], Rui Zhang[2], Jianhua Wang[1], Haobing Guo[2], He Zhang[2], Xinyu Guo[2], Tianquan Deng[3], Saibin Lv[1], Xuan Li[1], Jianqin Huang[1] and Guangyi Fan[2,*]

[1]State Key Laboratory of Subtropical Silviculture, Zhejiang A&F University, No. 666 Wusu St. Lin'an District, Hangzhou 311300, China

[2]BGI-Qingdao, BGI-Shenzhen, No. 2 Hengyunshan Rd. Huangdao District, Qingdao 266555, China

[3]BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China

[4]These authors contributed equally to this article.

***Correspondence: Lihong Xiao (xiaolh@zafu.edu.cn), Guangyi Fan (fanguangyi@genomics.cn)**

https://doi.org/10.1016/j.xplc.2021.100247

ABSTRACT

Pecan (*Carya illinoinensis*) is a tree nut crop of worldwide economic importance that is rich in health-promoting factors. However, pecan production and nut quality are greatly challenged by environmental stresses such as the outbreak of severe fungal diseases. Here, we report a high-quality, chromosome-scale genome assembly of the controlled-cross pecan cultivar 'Pawnee' constructed by integrating Nanopore sequencing and Hi-C technologies. Phylogenetic and evolutionary analyses reveal two whole-genome duplication (WGD) events and two paleo-subgenomes in pecan and walnut. Time estimates suggest that the recent WGD event and considerable genome rearrangements in pecan and walnut account for expansions in genome size and chromosome number after the divergence from bayberry. The two paleo-subgenomes differ in size and protein-coding gene sets. They exhibit uneven ancient gene loss, asymmetrical distribution of transposable elements (especially LTR/*Copia* and LTR/*Gypsy*), and expansions in transcription factor families (such as the extreme pecan-specific expansion in the far-red impaired response 1 family), which are likely to reflect the long evolutionary history of species in the Juglandaceae. A whole-genome scan of resequencing data from 86 pecan scab-associated core accessions identified 47 chromosome regions containing 185 putative candidate genes. Significant changes were detected in the expression of candidate genes associated with the chitin response pathway under chitin treatment in the scab-resistant and scab-susceptible cultivars 'Excell' and 'Pawnee'. These findings enable us to identify key genes that may be important susceptibility factors for fungal diseases in pecan. The high-quality sequences are valuable resources for pecan breeders and will provide a foundation for the production and quality improvement of tree nut crops.

Key words: pecan, genome assembly, paleo-subgenome, pecan scab, fungal disease, population genetics

Xiao L., Yu M., Zhang Y., Hu J., Zhang R., Wang J., Guo H., Zhang H., Guo X., Deng T., Lv S., Li X., Huang J., and Fan G. (2021). Chromosome-scale assembly reveals asymmetric paleo-subgenome evolution and targets for the acceleration of fungal resistance breeding in the nut crop, pecan. Plant Comm. **2**, 100247.

INTRODUCTION

The East Asia–North Eastern American disjunctive genus *Carya* (hickories) are widely grown for their wood, edible nuts, and ornamental value. The most economically significant *Carya* species is pecan (*Carya illinoinensis*), whose nuts are known to be rich in health-promoting factors, such as unsaturated fatty acids, antioxidant polyphenols, and vitamins. Pecan is consumed worldwide both directly and as a primary ingredient in many foods and confectionary products or as cooking oil after pressing

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

(Huang et al., 2019). Pecan is native to North America; its range spans tropical to temperate regions, and it is currently cultivated across six continents (Grauke et al., 2016).

Increasing consumer demands have promoted efforts toward the genetic improvement of pecan as a nut crop. Nonetheless, this work has been largely limited to the domestication and identification of varieties with good performance in yield-related traits, and the majority of cultivars have poor resistance to abiotic and biotic stresses (Goff et al., 1996; Wood et al., 2003; Thompson and Conner, 2012; Bock et al., 2020a). Among the economically damaging fungal diseases, pecan scab, caused by the phytopathogenic fungus *Venturia effusa*, is the most significant disease-related constraint to pecan production in the southeastern regions of the US, where severe fruit infection often results in major or complete crop loss (Bock et al., 2016, 2017, 2018, 2020a, 2020b). Although several natural scab-resistant genotypes have been selected and bred to limit yield losses, scab-susceptible cultivars still predominate in much of the existing and expanding pecan acreage (Thompson and Conner, 2012; Wells, 2014). In the last several years, control of this disease in pecan orchards in major cropping regions has relied largely on repeated and costly fungicide spraying, which also leads to fungicide resistance in the scab pathogen (Bock et al., 2020a). However, the genetic basis of scab resistance is poorly understood, and sources of genetic resistance to the pathogen are urgently needed.

As with many tree species, pecan requires a long generation time to reach full productivity and also displays sporophytic self-incompatibility (Thompson and Conner, 2012). This makes selection for many agronomically valuable traits by classical breeding approaches extremely slow, and it may take over 20 years to release a new cultivar (Thompson and Grauke, 1994; Conner, 2012). Therefore, genome-wide database resources that enable the identification and selection of many genetic loci simultaneously have huge potential to accelerate pecan research and breeding.

A draft genome assembly of the pecan cultivar 'Pawnee' was recently published (Huang et al., 2019). Some molecular and SSR markers have also been developed for pecan (Conner and Wood 2001; Grauke et al., 2003; Beedanagari et al., 2005; Chaney et al., 2015), and others are in progress (Jenkins et al., 2015). Although these studies provide essential resources for the identification of scab-resistant cultivars, there is still a need for a high-quality reference genome sequence of pecan to identify key candidate genes and to facilitate the development of more scab resistance-specific markers to aid in scab resistance breeding. Toward this goal, we used a whole-genome shotgun sequencing strategy that combined Oxford Nanopore long-read sequencing and Hi-C (high-throughput chromosome conformation capture) technology to construct a *de novo* chromosome-scale Pawnee genome assembly consisting of 16 pseudomolecules. Comparison of the pseudomolecules revealed two recent whole-genome duplication (WGD)- and genome rearrangement-related paleo-subgenomes with asymmetry in genome size, gene content, and transposable element (TE) distribution, as well as significant expansion of transcription factor families. We also used whole-genome resequencing data from 86 accessions of 36 genotypes with susceptibility or resistance to pecan scab to identify 47 resistance-related chromosome regions containing 185 putative candidate genes. The candidate gene set highlights genetic selection on putative genes involved in chitin responses, such as chitinase, MAP3K3, GLRs, and so forth, and it provides potential seedling screening markers for the development of fungal disease-resistant varieties.

RESULTS

Chromosome-scale assembly and annotation of pecan

A chromosome-scale assembly of a grafted plant derived from the controlled-cross pecan cultivar Pawnee (Figure 1A) was produced by integration of data generated from Oxford Nanopore sequencing and Hi-C technologies. A total of 71.7 Gb high-quality, cleaned Nanopore sequencing data, representing about 104-fold coverage of the estimated 691.28 Mb genome size with a heterozygosity of 1.52%, were used for *de novo* assembly (supplemental Figures 1 and 2; Table 1 and supplemental Tables 1 and 2). A 636.26-Mb initial assembly with a contig N50 length of 4.20 Mb and a longest contig of 23.88 Mb was obtained by combining *de novo* assembly of Nanopore sequences, error correction with Illumina sequences (generated previously), and removal of redundant and bacterial contamination sequences (supplemental Tables 1 and 3). The resulting 636.41-Mb, high-quality final assembly (Cil_v. 2.0) was generated using 75.93 Gb of Hi-C paired-end sequences, and 90.51% of the contigs in the initial assembly were anchored onto 16 pseudochromosomes with lengths ranging from 20.8 to 50.7 Mb (Figure 1C and supplemental Figure 3; Table 1 and supplemental Tables 1 and 4). Completeness assessment of the assembly revealed complete coverage of 95.1% of the core eukaryotic genes in the BUSCO database (Waterhouse et al., 2018) (supplemental Table 5). At least 93.72% of the Illumina short reads could be mapped to the assembly with coverages of 4-, 10-, and 20-fold (supplemental Table 6). The scatterplots of GC distribution showed a good concentration of nearly 36% and were close to the Poisson distribution (supplemental Figure 4). These metrics indicate the high accuracy and overall completeness of the assembly.

A combination of homology searches and *de novo* prediction resulted in the identification of 304.41 Mb of repetitive sequences in the Cil_v. 2.0 assembly, representing 47.83% of the pecan genome (Table 1 and supplemental Tables 7–9). The TE content of the assembly is 45.37%. LTR is the most abundant type, accounting for 35.15% of the pecan assembly (supplemental Tables 8 and 9), and the *Gypsy* and *Copia* subfamilies are the dominant subtypes (34.69% and 34.23% of TE length). *Gypsy* is usually enriched in the centromere regions of angiosperms, and this is also the case for Pawnee (Figure 1C; supplemental Table 9).

A total of 33472 protein-coding genes were predicted by integrating the *ab initio* prediction, homology search, and transcriptome assembly approaches, and 95.9% of these genes were anchored to the 16 pseudochromosomes (Table 1 and supplemental Tables 9 and 10). In total, 1349 of the predicted protein-coding genes can be completely matched with the BUSCO database (1440 genes), indicating the high completeness (93.7%) of the gene set (supplemental Table 5). Of the protein-coding genes, 31 247 (93.35%) have known functions in the SwissProt,

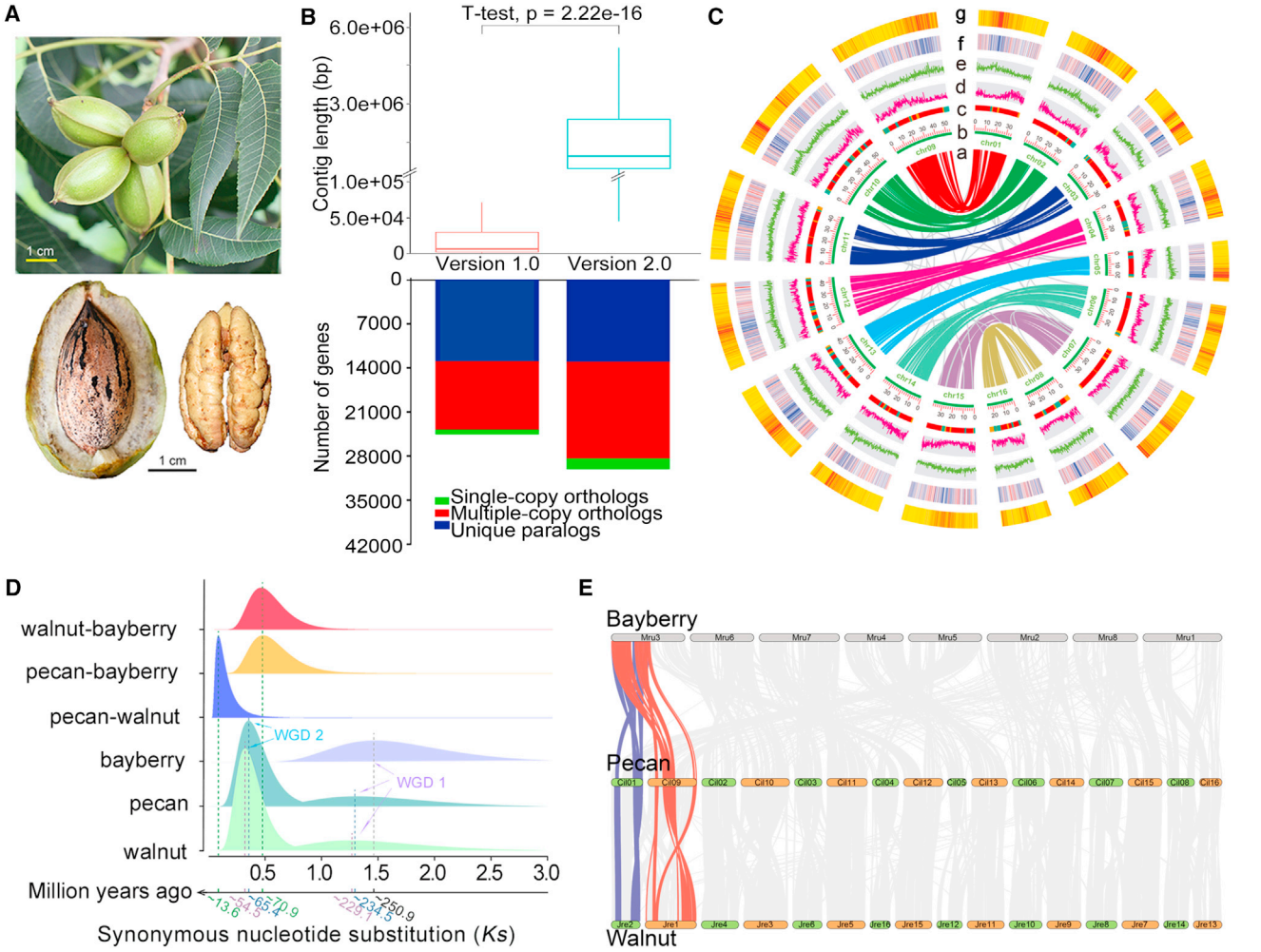


Figure 1. Genome features of the chromosome-scale assembly and evolution of the pecan genome.

(A) Morphology of fruit, nut, and fresh kernel of Pawnee. Scale bar corresponds to 1 cm.

(B) Comparison of contig lengths and clustered gene sets between versions 1.0 and 2.0.

(C) Landscape of the chromosome-scale pecan genome assembly. (a) Synteny of gene pairs from the recent WGD (WGD 2); (b) chromosomes; (c) contigs of version 1.0 on chromosomes of version 2.0; red, green, and yellow indicate contigs >2, 1–2, and <1 Mb in length in the version 1.0 assembly; (d) protein-coding genes; (e) LTR density distribution; (f–g) distribution of *Copia* and *Gypsy* elements.

(D) WGD and divergence within and among species of pecan, walnut, and bayberry.

(E) Syntenic distribution of pecan compared with walnut and bayberry.

Kyoto Encyclopedia of Genes and Genomes (KEGG), TrEMBL, InterPro, and Gene Ontology (GO) public databases (supplemental Tables 10–12). Compared with our previously reported scaffold-scale draft assembly (version 1.0), the chromosome-level assembly filled in 98.06% of the gaps with highly improved contig length, and this led to the identification of 2397 additional protein-coding genes (Figure 1B, 1C, and supplemental Figure S5; supplemental Table 13). Moreover, some agronomic trait-related genes that have been studied previously have been improved in the new genome version. For example, the copy numbers of genes encoding key components of oil accumulation and polyphenol metabolism decreased significantly in the Cil_v. 2.0 assembly (supplemental Table 14), probably because the short-read-based v. 1.0 assembly was inaccurate in the assembly of multicopy genes (Huang et al., 2019). In addition, the Pawnee assembly encodes 121 miRNAs, 565 tRNAs, 414 rRNAs, and 1318 snRNAs (supplemental Table 15).

When compared to the most recent published genomes of pecan (Lovell et al., 2021), the improved Pawnee Cil_v. 2.0 assembly is similar to the assemblies of four pecan genotypes in assembly completeness (Lovell et al., 2021). The Cil_v. 2.0 assembly is also similar to the genotypes 'Oaxaca', 'Lakota', and 'Elliott' in terms of genomic features, except that it has fewer scaffolds/contigs (supplemental Table 13), indicating that there are fewer gaps and missing sequences in the improved assembly. The gap-free Pawnee assembly reported by Lovell et al. (2021) shows slightly higher scores in genomic features than Cil_v. 2.0 (supplemental Table 13). Synteny analysis between the gap-free Pawnee assembly and Cil_v. 2.0 reveals high collinearity and one-to-one chromosome correspondence between the 2 assemblies, with 21 504 gene pairs (~67%) in syntenic blocks (supplemental Figures 6–7 and supplemental Table 16). Differences outside the syntenic blocks of the two Pawnee assemblies probably result from the

Estimated genome size (Mb)	691.28
Total length of scaffolds (Mb)	636.41
Number of scaffolds and contigs	124 & 564
Longest scaffold (Mb)	55.75
N50 of scaffold and contig length (Mb)	38.78 & 2.89
Number of predicted protein-coding genes	33 472
Pseudochromosomes	16
Anchored sequence to pseudochromosome (Mb)	608.60
Protein-coding genes in pseudochromosomes	32 104
Average gene length (CDS + intron) (bp)	5482.63
Masked repeat sequence length (Mb)	304.41
Percentage of repeat sequences (%)	47.83

Table 1. Global statistics of the chromosome-scale pecan genome assembly (Cil_v. 2.0).

syntenic block cutoff (at least five genes) and haplotype selection when assembling the chromosomes, as the outbred Pawnee has a highly heterozygous genome.

Genome evolution and identification of two paleo-subgenomes

We identified orthologous gene pairs in pecan, walnut, and bayberry, estimated species divergence times based on the synonymous nucleotide substitution (*Ks*) sites of orthologous genes, and corrected the times using the earliest fossil records of Myricaceae and Juglandaceae (64–84 mya) (Sauquet et al., 2012; Ho and Phillips, 2009). Our analysis revealed a shallow peak that occurred about 234.5 mya and very close to the ancient WGD event (WGD 1) in walnut and bayberry (~229.1 and ~250.9 mya) (Figure 1D), probably reflecting the paleopolyploidy WGD (γ) event in the angiosperm lineage (Landis et al., 2017). Pecan and walnut also experienced a recent WGD event (WGD 2) at about 65.4 and 54.5 mya (Figure 1D). Estimates of divergence times between species of walnut–bayberry and pecan–bayberry indicated that the speciation event in Juglandaceae and Myricaceae occurred before the tetraploidization in the genera of Juglandaceae, whereas the divergence between pecan and walnut (~13.6 mya) occurred after their tetraploidization (Figure 1D), suggesting a common ancestor between pecan and walnut.

The two WGD events involved a total of 3,683 orthologous gene pairs, 2,829 of which were in WGD 2, reflecting its important contribution to the protein-coding gene set. The gene pairs from WGD 2 were mapped to the 16 pseudochromosomes to visualize the detailed syntenic relationships among chromosomes in pecan. We observed eight chromosome pairs with one-to-one corresponding collinear relationships in the pecan assembly (Figure 1C). Further mapping of all identified orthologous genes in syntenic blocks between chromosomes of any two genomes (among pecan, walnut, and bayberry) revealed a considerable number of pair-to-pair relationships between the chromosomes of pecan and walnut but not between the chromosomes of pecan and bayberry (Figure 1E). Based on the results of syntenic analysis, the chromosomes of pecan, walnut, and bayberry were divided into eight groups, and one-to-one orthologous gene pairs identified in the groups (supplemental Table 17) were used to construct eight phylogenetic trees by the neighbor-joining (NJ) method (supplemental Figure 8). Two paleo-subgenomes—subgenome A (chromosomes Chr09 to Chr16) and subgenome B (chromosomes Chr01 to Chr08)—were identified based on the branch lengths of the NJ trees for both pecan and walnut (Figure 1C, 1D, and supplemental Figure 8). The walnut paleo-subgenomes were the same as those reported in the recently published walnut assembly (Zhang et al., 2020). These results reflected frequent large-scale chromosome rearrangements in the pecan and walnut comparing with bayberry genomes after the divergence of Myricaceae and Juglandaceae, as well as rare rearrangement events between pecan and walnut because of their relatively short divergence time.

Features and evolution of the two paleo-subgenomes

Comparison of the sequence similarity of the two paleo-subgenomes of pecan against the bayberry genome showed higher average identity for each chromosome in subgenome B than in subgenome A (supplemental Table 17). The overall identity distribution between the two subgenomes displayed a trend similar to that of the average identity for each chromosome (supplemental Figure 9). Selection analysis (*Ka*/*Ks*, i.e., ω) revealed that the chromosomes in subgenome B had experienced stronger positive selection than those in subgenome A, in addition to the PAIR-5 chromosomes (supplemental Figure 10; supplemental Table 17). We also detected asymmetry in lengths and protein-coding genes between the paired chromosomes of the two subgenomes. Except for the PAIR-8 chromosomes, all chromosomes in subgenome A were longer and contained more protein-coding genes (345.88 Mb and 18 498 genes) than those in subgenome B (328.41 Mb and 13 606 genes) (supplemental Table 17).

To compare the paleo-subgenome features in pecan, the values of *Ka*, *Ks*, and *Ka*/*Ks* were estimated based on 6316 orthologous gene pairs from the subgenomes. According to the values of *Ks* and *Ka*, 6253 of the analyzed orthologous gene pairs had been subjected to negative selection (*Ka*/*Ks* < 1), whereas 63 had been subjected to positive selection (*Ka*/*Ks* > 1). No genes under neutral selection were detected (*Ka*/*Ks* = 1). These results indicated that these genes may have undergone lower selection pressure, evolving at a slower evolutionary rate. In addition, we also established the relationships among *Ka*/*Ks*, *Ka*, and *Ks* in the pecan genome. We found that *Ka* increased gradually with increasing *Ks* (supplemental Figure 11), with $R = 0.71$, $P < 2.2 \times 10^{-16}$ (Spearman's rank correlation). These data were basically consistent with those in pear ($R = 0.75$) (Cao et al., 2019), suggesting that mechanisms that affect both *Ka* and *Ks* sites may be shared in different genomes. In addition, the *Ka*/*Ks* ratio was negatively correlated with both *Ka* ($R = 0.34$, $P < 2.2 \times 10^{-16}$) and *Ks* ($R = -0.28$, $P < 2.2 \times 10^{-16}$) (supplemental Figure 11). The correlation between *Ka* and *Ka*/*Ks* was greater than that between *Ks* and *Ka*/*Ks*, indicating that *Ka* may be a determining factor for the *Ka*/*Ks* ratio between the subgenomes.

For evolutionary analyses of the two paleo-subgenomes in pecan, a simplified phylogenetic tree was constructed using 1080

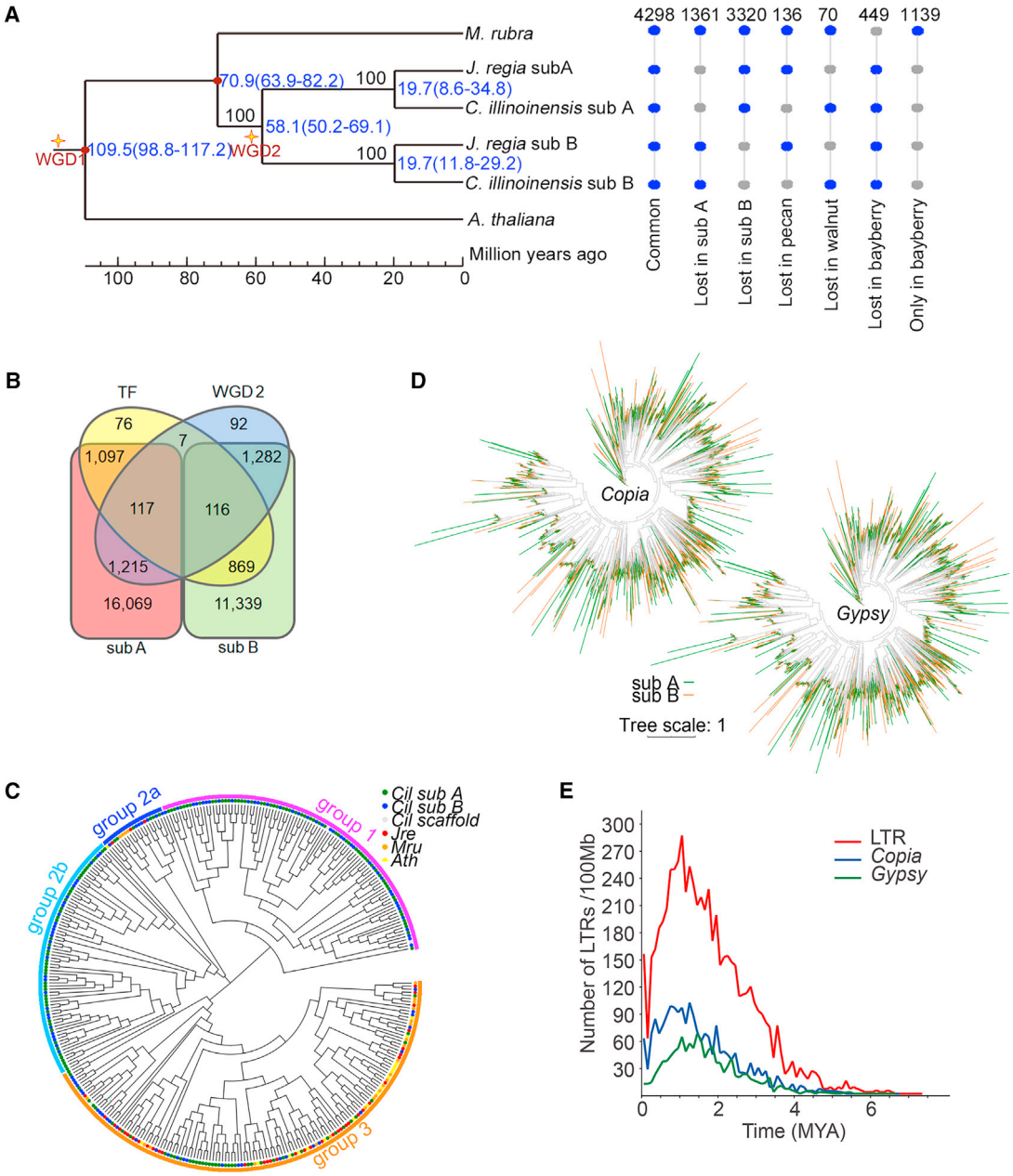


Figure 2. The features, evolution, and divergence of the pecan subgenomes.

(A) Evolutionary analysis of genomes, subgenomes, and gene sets in pecan, walnut, and bayberry, using *Arabidopsis* as an outgroup. Numbers indicate the number of gene families.

(B) Venn diagram of genes in the recent WGD (WGD 2), subgenomes (sub A and sub B), and transcription factors. Numbers indicate the number of genes.

(C) Significantly expanded TF families.

(D) Distribution of *Copia* and *Gypsy* LTRs on the subgenomes.

(E) Insertion time estimates of LTRs.

single-copy orthologous genes from pecan, with walnut, bayberry, and *Arabidopsis* as references (Figure 2A). The topology of the phylogenetic tree confirmed the close relationship between pecan and walnut, as did our previous report (Huang et al., 2019). Interestingly, the group A or B subgenomes of pecan and walnut were clustered into a terminal clade (Figure 2A). The corrected estimate for the split between subgenomes A and B in the 2 species was about 58 mya, earlier than 11.8–29.2 mya, a time representing the splits of subgenomes A or B between the species (Figure 2A) and during which time a speciation event occurred between pecan and walnut (~13.6 mya) (Figure 1D). This analysis indicated that differentiation between the two paleo-subgenomes occurred in the common ancestor of pecan and walnut after the Juglandaceae had diverged from the Myricaceae.

Gene family statistics showed that 4298 families were common to pecan, walnut, and bayberry; 1361 or 3320 families were lost in subgenome A or subgenome B of both pecan and walnut; 136

families were lost only in pecan, and 70 were lost only in walnut; and 449 families were retained in both subgenomes of pecan and walnut (Figure 2A). The considerable alteration of gene families among species and subgenomes largely reflects genome-wide rearrangements before and after the recent WGD during the species' evolution.

The pecan genome encoded a total of 2282 transcription factors (TFs) from 61 TF families, ~53% (1204) of them in subgenome A and ~43% (985) in subgenome B. Only 117 in subgenome A and 116 in subgenome B were related to the recent WGD event (WGD 2) (Figure 2B). Of the 2829 WGD 2 genes, 1332 were encoded by subgenome A, and slightly fewer were encoded by subgenome B (1398). Among the TF families, 18 were significantly over-represented in pecan and walnut, and the *far-red impaired response 1* (*FAR1*) family showed extreme expansion, especially in pecan (supplemental Table 18). A maximum likelihood (ML) tree of 264 FAR1-encoding genes revealed three major clades representing the three groups of this family, and group 1 and group 2b *FAR1s* existed specifically in the pecan genome (Figure 2C). Based on the diverse biological functions of this gene family, expansion of the *FAR1s* may account for the enhanced light signaling in pecan life processes, including plant development, stress response, and immunity (Ma et al., 2016; Wang et al., 2016). Further chromosome mapping revealed an asymmetrical distribution of *FAR1* loci between subgenomes, with more members in subgenome A (Figure 2C and supplemental Figure 12).

The asymmetry of the subgenomes was also shown in the distribution of TEs, especially in the numbers of major LTR types, the *Copia* (132 369 in subgenome A and 90 474 in subgenome B) and *Gypsy* (100 257 in subgenome A and 89 931 in subgenome B) subfamilies (Figure 2D; supplemental Table 9). Insertion time estimates show that the insertion times of total LTRs and *Gypsy* and *Copia* elements are significantly earlier in subgenome A than in subgenome B (Figure 2E).

Population phylogenetic and genetic analysis of pecan scab-associated accessions

A total of 86 pecan accessions, representing scab-associated core varieties in our collection, were selected for genome-wide analysis (supplemental Table 19). This set of germplasm includes 36 core pecan scab-associated varieties: 29 single individual varieties and 7 cloned populations. Genome resequencing of the accessions using the BGI-Seq 500 sequencer generated a total of 1624.41 Gb of sequences after trimming of low-quality reads. On average, ~19 Gb of clean data (27× coverage of the estimated pecan genome size) were obtained for each sample (supplemental Table 20). The filtered reads from each accession were mapped to the pecan Cil_v. 2.0 assembly with an average mapping rate of 96.57%. The mapped reads covered most regions of the reference genome with a coverage ratio from 90.99% to 95.16% among the accessions. A total of 24 972 828 high-quality SNPs were detected. A further filtering step revealed 5 901 970 SNPs that were suitable for population analysis, more than half of which (3 293 001) were unique to subgenome A and 2 608 969 of which were unique to subgenome B.

NJ phylogenetic trees were built to display the phylogenetic relationships among the 36 cultivars based on the variations in each subgenome (Figure 3A). Topologies of both NJ trees clearly formed two major clades for each subgenome, but the cultivars of different major clades varied between subgenome A and subgenome B. The phylogenetic relationships of cultivars in each subclade within major clades of subgenome A or subgenome B were closely related to their genetic relationships (Figure 3A; supplemental Table 21) but showed no obvious correlation with disease resistance (Figure 3A; supplemental Table 20). Internal structure comparison between the two phylogenetic trees revealed the best correspondences of all the leaf (outer) nodes and parts of the inner nodes, with a score of 1, and relatively lower correspondences for the root nodes between the NJ trees of subgenome A and subgenome B, with a score of 0.5.

Further population structure investigation of the 36 cultivars at the subgenome level revealed that $K = 4$ was the best cluster number for the datasets (Figure 3A). To facilitate comparison between subgenomes, we denoted the four K numbers *K1* to *K4* based on the color of the visualized structure and recorded the ancestral types for each cultivar in subgenomes A and B (Figure 3A; supplemental Table 21). We found that 10 of the cultivars originated from a single ancestor and that 6 cultivars derived from 2 to 4 ancestors, and all these 16 cultivars had the same ancestral types in subgenomes A and B (Figure 3A; supplemental Table 21). The remaining 20 cultivars differed in both ancestors and ancestral types between subgenomes (Figure 3A; supplemental Table 21). These results may reflect the complex domestication history and frequent gene flow caused by natural and human selection and inter- and intra-species hybridization and admixture among the cultivars.

To evaluate the genetic diversity among the accessions, we first divided the accessions into two populations based on their pecan scab-resistance grades: the resistant population (denoted R) had grades ≤ 2, and the susceptible population (denoted S) had grades >2 (supplemental Table 20). We then quantified the variations in nucleotide diversity (π value) for each population and the pairwise differentiation level (*Fst*) between the two populations.

Identification of selected regions and candidate genes associated with pecan scab resistance

Given that disease resistance-associated regions in the R population were subjected to stronger selection pressure and therefore had lower polymorphism than corresponding regions in the S population, chromosome regions (100 kb per window) with both π ratios of π_S/π_R and *Fst* values in the top 5% were identified as selected regions associated with pecan scab resistance (Figure 3B; supplemental Table 22). The analyses revealed a total of 47 candidate regions that contained 185 putative protein-coding genes, 141 of which were located in subgenome A and 45 in subgenome B (Figure 3B; supplemental Table 23). The candidate regions were unevenly distributed on 12 chromosomes and were most abundant on chromosomes 6, 10, 11, and 15 (Figure 3C).

Of the 47 selected regions, a region of approximately 83.6 kb on chromosome 9 (between 3.1 and 3.2 Mb) displayed the highest *Fst* value and a relatively high π ratio (*Fst* = 0.208, π = 2.879), and it contained 7 putative protein-coding genes (supplemental Table 23). Two of the seven genes, *Cil_09G_00199V2* and

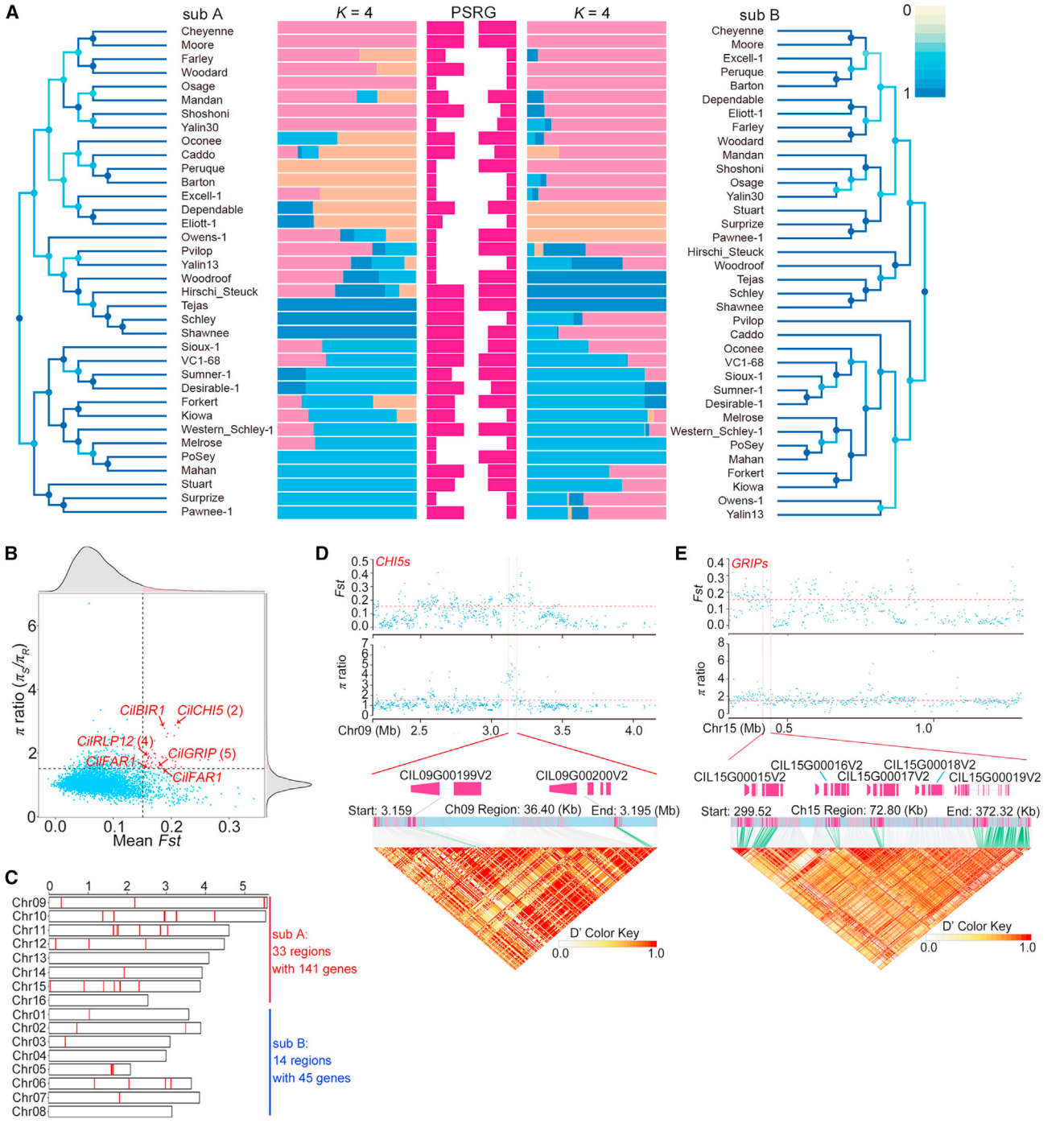


Figure 3. Population genomics and identification of pecan scab-associated candidate genomic regions and genes under selection.

(A) NJ trees and population structures of subgenomes and scab-resistance grades of each accession. A score of 1 on the color scale bar indicates that the tree structure of the node is identical to the tree structure of its best corresponding node. *K* = 4, the best substructure. PSRG, pecan scab-resistance grade. The length of the red bars in the middle indicates the disease-resistance grade.

(B) Plots of the highest 5% π and *Fst* values in pecan scab-related accessions. Arrowheads indicate the loci of key candidate genes associated with pecan scab resistance.

(C) Locations of the selected regions of pecan scab resistance on chromosomes.

(D) *Fst* and π ratio (upper, coordinate diagrams), gene details (middle, color bars), and LD heatmap of the candidate region containing two putative chitinase-encoding genes.

(E) *Fst* and π ratio (upper, coordinate diagrams), gene details (middle, color bars), and LD heatmap of the candidate region containing five putative GRIP-encoding genes. The pairwise LD between the SNPs is indicated as D′ values, where red indicates a value of 1 and yellow indicates 0. Rose-red indicates the CDS regions of genes, and green indicates polymorphic SNP sites in the promoter and CDS regions.

Cil_09G_00200V2, were annotated as chitinases (denoted *CilCHI5_1* and *CilCHI5_2*) (Figure 3B–3D). They were the closest known homologs of an EP3 endochitinase in plants that has been observed to participate in the innate immune response through inhibition of fungal growth (de A. Gerhardt et al., 1997). Detailed analysis revealed two nonsynonymous nucleotide substitutions (missense variants) in *CHI5_1* and four in *CilCHI5_2*, and *CilCHI5_2* also harbored an intron variant (Figure 3D; supplemental Table 24).

An approximately 72.4-kb region on chromosome 15 with an *Fst* value of 0.174 and a π of 1.623 also attracted our attention because it contains 5 tandem repeat genes encoding ionotropic glutamate receptors, *Cil_15G_00015V2* to *Cil_15G_00019V2* (denoted *CilGLR3.6/GRIP1–CilGLR3.6/GRIP5*) (Figure 3B, 3C, and 3E; supplemental Table 20). Plant glutamate receptor-like (*GLR*) homologs have been reported to participate in many plant-specific physiological functions, such as sperm signaling, pollen tube growth, root meristem proliferation, abiotic responses, and innate immunity (Zhu, 2016; Li et al., 2019; Wudick et al., 2018). We detected a total of 183 variants in this tandem repeat region, 28 of which were synonymous substitutions and 21 of which were located in the downstream (14) or upstream (7) regions of the *CilGLR* genes (Figure 3E; supplemental Table 24). Most of the variants (up to 92) were located in introns, and 42 missense variants were found within the *GLRs*. Four variants were detected in the splice-and-intron regions of *GRIP3* and *GRIP5*, and one was identified as a stop-gain variant in *CilGRIP5* (supplemental Table 24).

In addition to two *CilCHI5s* and five *CilGLR3.6s/GRIPs*, we also identified a mitogen-activated protein kinase kinase kinase (MAP3K3/MPKKK3)-encoding gene (*Cil_03G_00295V2*) that has been well studied in model plant species as a key gene in the chitin-signaling cascade (Figures 3B and 4A; supplemental Table 20 [Gong et al., 2020]). We also detected eight transcription factor genes in the selected regions, including two FAR1 family members in the pecan-specific expansion (supplemental Table 23) that may contribute to pecan scab resistance in this species.

Expression patterns of candidate key genes in response to chitin treatment

To investigate the functions of candidate key genes in response to fungal disease, two cultivars with historical records of strong pecan scab resistant (Excell) and susceptible (Pawnee) phenotypes were subjected to chitin treatment for 30, 60, and 180 min (Figure 4B). Expression levels of 10 candidate key genes were examined by real-time qPCR: *CilCHI5_1*, *CilCHI5_2*, five *CilGRIPs*, *CilMAP3K3*, *CilFAR1_1*, and *CilFAR1_2* (supplemental Table 25). Six of the genes responded to the chitin treatment, five of which were induced as early as 30 min after chitin treatment (Figure 4C–4H). *CHI5_2* and *MAP3K3* were induced by chitin with similar expression in both susceptible and resistant cultivars at the early response stage (Figure 4C and 4D), indicating that they may have important roles in defense against fungal pathogens at early infection stages. Two *GLR3.6/GRIP* homologs were significantly induced in "Excell" (Figure 4E and 4F), suggesting their close correlation with fungal pathogen resistance. By contrast, two *FAR* members were quickly upregulated in only the susceptible cultivar Pawnee (Figure 4G and 4H), probably reflecting their involvement in fungal disease resistance.

DISCUSSION

Toward a reference genome for pecan

Pecan is typical of a number of important tree nut crop species with high heterozygosity and high genetic diversity due to self-incompatibility and for which limited genome data are currently available. The publicly available draft genome sequence of the pecan cultivar Pawnee makes it possible to identify most gene sequences of interest but not their chromosomal locations or the exact family members for multicopy genes. A highly continuous and complete reference genome is an essential basis for a wide range of studies on gene functions, molecular and metabolic mechanisms, population genetics, breeding, and so forth. By combining current state-of-the art technologies—Oxford Nanopore long-read (>2 kb) sequencing, Hi-C technology, and high-quality genome assemblers—we constructed a chromosome-scale genome assembly for the pecan cultivar Pawnee. The *de-novo*-assembled Cil_v. 2.0 Pawnee genome displays high continuity, integrity, and quality, with a contig N50 of 3.04 Mb and BUSCO assessments of 95.1% for assembly and 93.7% for protein-coding sequences. The newly assembled genome sequence is about 92% of the estimated genome size, with a total of 608.6 Mb on 16 chromosomes (95.6% of the new assembly). All of the missing sequence lengths are likely to be telomeric and centromeric repeats. The Cil_v. 2.0 assembly contains 33 472 predicted protein-coding models, of which 9516 are unique to Cil_v. 2.0 and, in total, 2397 more than the previous version (v. 1.0) (Huang et al., 2019), reflecting the higher continuity, integrity, and accuracy of the new version.

Asymmetry in the evolution and features of the pecan paleo-subgenomes

An ancient WGD event, i.e., the γ triplication event, has been widely reported in many angiosperms, including Fagales (Tuskan et al., 2006; Jaillon et al., 2007; Huang et al., 2009, 2019; Luo et al., 2015). Our analysis confirmed these ancient WGD events (WGD 1) and a recent WGD (WGD 2) in the pecan genome (Figure 1D). Based on the syntenic relationships of homologous gene pairs in WGD 2, we divided the 16 pecan chromosomes into 8 pairs of homoeologs and further divided them into 2 paleo-subgenomes based on the phylogenetic relationships among pecan, walnut, and bayberry (Figures 1C, 1E, and 2), which were similar to those reported in walnut (Luo et al., 2015; Zhang et al., 2020). Previous studies suggested that the γ triplication generated a genome with n = 21, indicating that the 8 homoeologous chromosome pairs in haploid genomes of Juglandaceae evolved from n = 8 by a recent WGD rather than n = 21 by dysploid reduction (Salse, 2012; Luo et al., 2015). Our analysis of the pecan genome was consistent with this conclusion (Figures 3D and 4B). One hypothesis of x = 8 as the ancestral state was supported by the presence of chromosome number n = 8 in the genera *Roipterlea* and *Myrica*, which are closely related to Juglandaceae (Luo et al., 2015). Timing inference in this study revealed that the recent WGD events in pecan and walnut (65.4 and 54.5 mya) happened in the "juglandoid" WGD (56–66 mya) near the Cretaceous–Tertiary (K–T) boundary about 66 mya (Manchester, 1989; Luo et al., 2015), after the divergence from

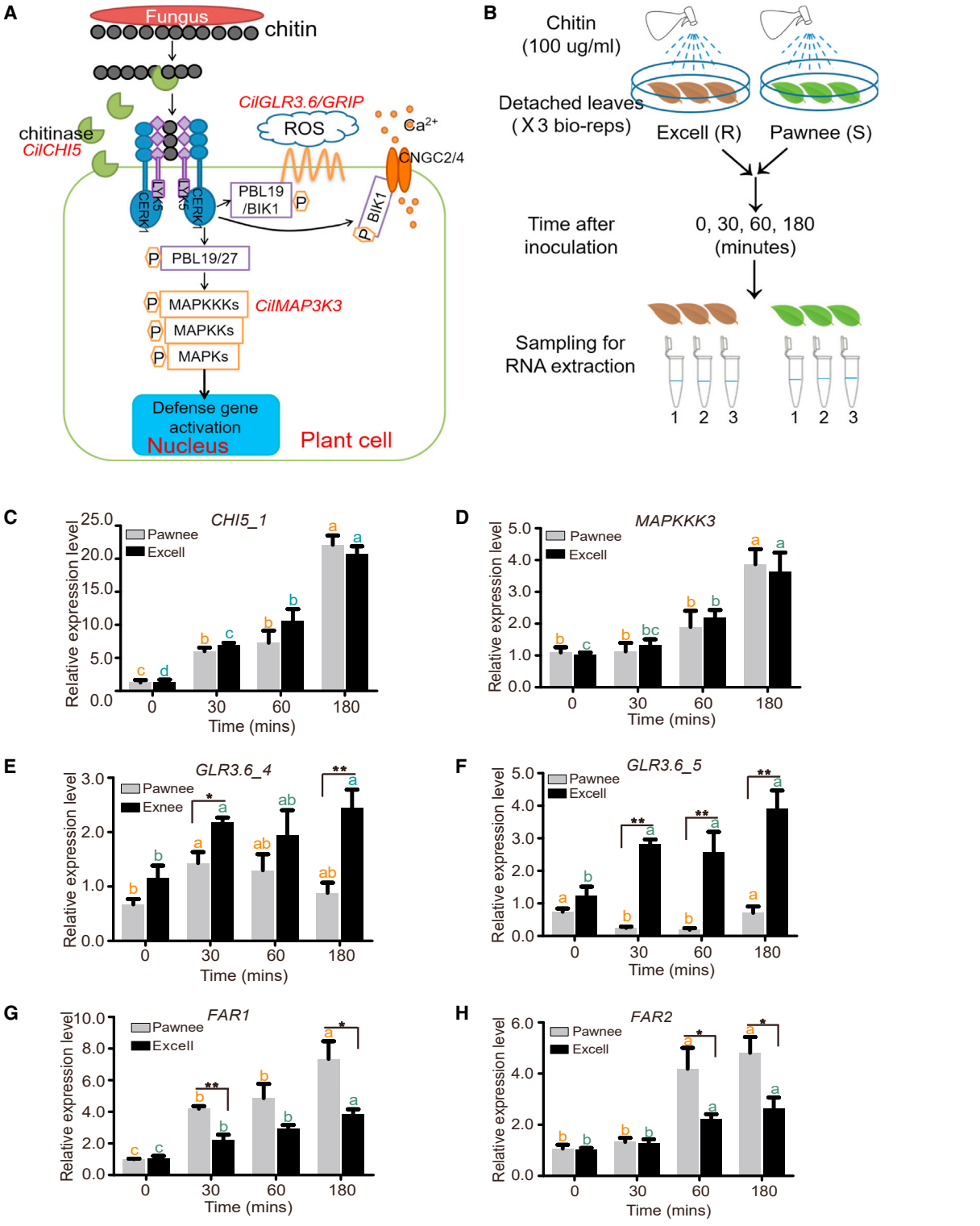


Figure 4. Identification and expression analyses of key candidate genes involved in the chitin signaling pathway.

(A) A simplified chitin signaling pathway in plants and the key genes under selection. Genes in red are putative candidate genes selected for further analysis.

(B) Chitin treatment and sampling strategy for expression analysis by qPCR. R, pecan scab resistance; S, pecan scab sensitivity.

(C–H) Expression of selected candidate key genes involved in chitin signaling pathways in chitin-treated leaves of the pecan cultivars Excell and Pawnee. Letters show the significance of differences between time points within cultivars ($P < 0.05$, one-way ANOVA), and stars show the significance of differences between cultivars (*$P < 0.05$, **$P < 0.01$). qPCR was performed using five replicate leaves from each time point and cultivar.

bayberry (Figures 3D and 4B). The divergence between the paleo-subgenomes preceded the split between pecan and walnut, and the divergence time between paleo-subgenomes followed and/or coincided with the "juglandoid" WGD events and was accompanied by extensive genome rearrangements, probably reflecting rapid genome evolution and adaptive evolution to survive the adverse environmental conditions associated with the K–T boundary (Fawcett et al., 2009; Soltis and Burleigh, 2009; Van de Peer et al., 2009; Luo et al., 2015). The relatively lower syntenic relationship of bayberry with pecan or walnut in this study provided solid evidence for this conclusion (Figure 1E). Moreover, the asymmetry between paleo-subgenome features, such as genome size, number of gene models, TE distribution, and pecan-specific gene family expansion, all strongly supported the inference of a "juglandoid" WGD and a large-scale genome rearrangement-associated evolutionary trajectory during the K–T boundary in the Juglandaceae. Nonetheless, further evidence is still needed to uncover the sources of the "juglandoid" WGD, i.e., data derived from parental ancestral hybridization or from duplication of one ancestral species.

Genome-based insights into breeding targets for fungal disease resistance

Fungal pathogens constitute major threats to land plants and pose growing challenges to global crop production; they have led to losses of approximately 30% in annual global crop production before and after harvest (Gong et al., 2020). In plants, the first layer of innate immunity relies on the perception of conserved pathogen-associated molecular patterns (PAMPs). This perception is mediated by pattern recognition receptors located at the cell surface, including membrane-localized receptor-like kinases and receptor-like proteins, which elicit PAMP-triggered immunity (Wang et al., 2017; Dodds and Rathjen, 2010; Tena et al., 2011). Chitin, an insoluble polymer of β-1,4-linked *N*-acetyl-glucosamine, is a highly conserved building block of fungal cell walls and a broadly effective elicitor of plant immunity. Invasion by fungal pathogens can induce the secretion of plant chitinases into the apoplast to hydrolyze fungal cell walls and release chitin oligomers. PAMP recognition then rapidly initiates a series of early immune responses, including the activation of mitogen-activated protein kinase (MAPK) cascades and the production of reactive oxygen species (ROS) to combat pathogen infection.

To date, many key genes involved in chitin perception and signaling pathways have been identified in model plants such as *Arabidopsis* and rice (Liu et al., 2016; Wang et al., 2017; Gong et al., 2020). In this study, genome-based population genetic diversity enabled us to identify 47 selected regions containing 185 putative candidate genes associated with scab resistance in pecan (Figure 3). Fourteen of them were annotated as receptor(-like) proteins, including one bacterial flagellin receptor-like protein (FLS2) and five proteins with ionotropic GLR activity (supplemental Table 23). FLS2 has been extensively reported to function in bacterial-derived PAMPs but not in fungal-derived PAMPs (Wang et al., 2017), and its identification here may reflect its important role in fungal disease resistance of pecan. However, the chitin receptor kinase *CERK* homolog did not appear to be under selection in this study, possibly because of the small population sample. GLRs in plants participate in diverse and important biological processes, such as photosynthesis, cellular C/N balance, plant organ development, abiotic stress response, plant-pathogen interactions, calcium-mediated signal transduction, and so forth (Kang and Turano, 2003; Kang et al., 2004, 2006; Singh et al., 2006; Li et al., 2013; Manzoor et al., 2013; Cheng et al., 2016). The relatively high *Fst* values of the five *GLR* genes detected in pecan scab-resistant cultivars probably suggest enhanced chitin-induced fungal resistance. Recognition of chitin is known to trigger the intracellular activation of MAPK cascades and the rapid production of ROS (Tsutomu et al., 2017). The activation of MAPK cascades is the core step in chitin-induced immune responses (Yamada et al., 2017), and a homolog of *MAP3K3* (*Cil_03G_00295V2*) was identified and shown to be induced by chitin treatment in scab-resistant pecan cultivars, implying that it may have an important role in the fungal defense response of pecan. Our results also highlighted two putative EP3 endochitinase-like genes, *CHI5s*, which have been reported to function in innate immune responses by degrading the fungal cell wall to inhibit fungal growth in plants (de A. Gerhardt et al., 1997). One of the two detected *CilCHI5* genes was strongly induced in both resistant and susceptible cultivars at the early stages of chitin treatment, indicating its important role in early defense against fungal pathogens. The expression levels of *CHI5_1* showed no obvious differences between the cultivars, indicating that the defense mechanism of resistant varieties may involve downstream signal transduction and corresponding processes. Also, the specific induction of *GLR3.6_4* and *GLR3.6_5* expression in resistant cultivars (Figure 4E and 4F) may serve as a potential marker for the screening of fungal pathogen-resistant varieties at the seedling stage after further experimental validation. Our findings provide important clues and potential targets for uncovering the intrinsic mechanisms of fungal disease resistance and breeding in the future.

MATERIALS AND METHODS

Plant materials

The widely planted cultivar Pawnee, which was released in 1985 as the progeny of a controlled cross between 'Mohawk' × 'Starking Hardy Giant' performed in 1963 (Thompson and Hunter, 1985), was selected for whole-genome sequencing. Fresh young leaves were collected from a grafted plant growing in the plantation of Zhejiang A&F University, Lin'an District, Hangzhou, China in April 2018. To investigate genome-wide associations with scab resistance, fresh young leaves from 86 accessions representing 36 genotypes, including 7 cloned populations, were collected from April to May 2019 (supplemental Table 19). All the collected samples were frozen and transported in liquid nitrogen and stored at −80°C in a freezer before use.

Preparation of genomic DNA and Nanopore sequencing

High-molecular-weight genomic DNA was extracted from young Pawnee leaves using the DNeasy Plant Mini Kit (QIAGEN, Germany) for use in v. 2.0 genome assembly. DNA quality and quantity were determined using a NanoDrop spectrophotometer (Thermo Fisher Scientific, USA), Qubit dsDNA HS Assay Kits, and a Qubit 2.0 fluorometer (Invitrogen, USA). Genomic DNA of over 2 kb in length was purified using a BluePippin automatic nucleic acid electrophoresis and fragment recovery system (Sage Science, USA). The recovered DNA was used to construct libraries for whole-genome sequencing using the Nanopore PromethION platform (Oxford Nanopore Technologies, UK) at Biomarker Technologies, Beijing, China.

Hi-C library preparation and sequencing

To enable a high-quality, chromosome-level assembly of the pecan reference genome, fresh young leaves were collected from the tree and used

for whole-genome sequencing. Leaf samples frozen in liquid nitrogen were fixed with 2% formaldehyde solution in PBS buffer for 30 min, and the reaction was terminated using 2.5 M glycine for 5 min. The fixed samples were sent to BGI-Qingdao (Qingdao, China) for Hi-C library construction using the DNA restriction endonuclease *Dpn*II, according to the standard library preparation protocol (Burton et al., 2013). The BGISEQ-500 platform (BGI-Shenzhen, China) was used for library preparation and sequencing.

Genome survey

Approximately 168 Gb of Illumina data that had been used for scaffold-level assembly of the version 1.0 pecan genome (Huang et al., 2019) were used for a genome survey in the present study. We first used SOAPnuke software (Chen et al., 2018) to remove low-quality paired-end raw reads and then used GenomeScope (Vurture et al., 2017) to estimate the genome size, heterozygosity, and repeat rate based on the 17-mer depth frequency distribution.

Genome assembly

The quality-filtered Nanopore data were assembled using CANU v. 1.6 software (Koren et al., 2017) with optimized parameters (genomeSize=700m minReadLength=500 -correctedErrorRate=0.20 -fast). The accuracy of the initial assembly was then improved three times using Pilon v. 1.22 (Walker et al., 2014), and redundant contigs were removed using Purge_Haplotigs in the CANU package. To remove possible contamination by bacterial sequences identified by GC depth analysis, NT Blast was launched to scan all assembled contigs and eliminate those contigs with best hits to bacterial sequences. Next, all contigs were mapped to the pecan chloroplast genomes deposited in GenBank (accessions MW410238, MH909600, and MH909599) to remove chloroplast sequences. To evaluate the consistency and integrity of the initial polished assembly, Illumina short reads were blast-searched against the genome assembly using BWA v. 0.7.12 (Li and Durbin, 2009), and BUSCO v. 4.1.2 (Simão et al., 2015) analysis was performed to further evaluate the assembly.

To generate a chromosome-level assembly, Hi-C paired-end reads were subjected to quality control using HiC-Pro v. 2.8.0 (Servant et al., 2015). Low-quality bases and adapter sequences were then removed using Bowtie 2 v. 2.2.5 (Langmead et al., 2009), and Juicer v. 1.5 (Durand et al., 2016) was used to analyze the Hi-C datasets. Finally, a 3D *de novo* assembly (3D-DNA, v. 170123) pipeline (Dudchenko et al., 2017) was used to scaffold the assembly onto pseudochromosomes.

Genome annotation

Repetitive sequences were predicted in the pecan genome using homology-based searches combined with *ab initio* approaches. TRF (v. 4.07b; Benson, 1999) was used to identify tandem repeats. RepeatMasker and RepeatProteinMask (v. 3.3.0; http://www.repeatmasker.org/) were used to search for known TEs against the Repbase library and the TE protein database (Jurka et al., 2005). Then, a *de novo* repeat library was built using RepeatModeler software (v. 2.0; http://www.repeatmasker.org) with default parameters, and all TEs were classified using RepeatMasker (http://www.repeatmasker.org/).

The *de novo* prediction of protein-coding genes was performed using AUGUSTUS (v. 3.1; Stanke et al., 2004) and Genscan (v. 1.0; Aggarwal and Ramaswamy, 2002). GeneWise (v. 2.4.1; Birney et al., 2004) was used for homologous annotation against protein datasets from nine species (supplemental Table 8) downloaded from the National Center for Biotechnology Information (NCBI) database. To assist with gene model prediction, paired-end RNA sequencing reads from leaf, epicarp, embryo, and stem tissues in our previous study (Huang et al., 2019) were assembled *de novo* using Trinity (v. 2.8.5; Grabherr et al., 2013), followed by gene model prediction of transcripts using PASA (v. 2.3.3; Campbell et al., 2006). Gene models from these different approaches were integrated into a non-redundant set of gene structures using GLEAN (v. 1.0; Elsik et al., 2007) with default parameters. The final pecan gene set was assessed for completeness of the annotated protein-coding genes.

Functional annotation of protein-coding genes was achieved by homolog searches against the TrEMBL (UniProtKB), SwissProt (Bairoch and Apweiler, 2000), KEGG (Kanehisa and Goto, 2000), GO (Consortium, 2004), and InterProScan (Jones et al., 2014) databases with an E value cutoff of 1×10^{-7}. Non-coding RNAs, including rRNAs, tRNAs, snRNAs, and miRNAs, were identified by searching against various RNA libraries. tRNAscan-SE v. 1.3.1 software (Lowe and Eddy, 1996) was run with eukaryote parameters to identify tRNA genes. The rRNA sequences were annotated based on homology to previously published rRNA sequences in plants. The snRNAs and miRNAs were predicted using the "cmsearch" program in Infernal v. 1.1 (Nawrocki and Eddy, 2013) to search against Rfam v. 13.0 (Kalvari et al., 2018) with an E value cutoff of 0.01.

WGD and synteny analyses

To estimate the timing of WGD events, wgd software (Zwaenepoel and Peer, 2019) was used to calculate the *Ks* distribution of orthologs from pecan, walnut, and bayberry, and then the Gaussian mixture model (GMM) was used to fit a curve for the *Ks* distribution of each species. In the same way, wgd software was also used to estimate the divergence between any two species among pecan, walnut, and bayberry by calculating the *Ks* values of one-versus-one ortholog pairs between two species and fitting curves using the GMM model. The divergence times of WGD events within species and the divergences between species were estimated by the formula *Ks* 1/time 1 = *Ks* 2/time 2 (*Ks* 1, divergence value of ortholog pairs between species; *Ks* 2, WGD peak; time 1, divergence time between species; time 2, WGD time), and corrected based on the earliest fossil records of Myricaceae and Juglandaceae (64–84 mya) (Ho and Phillips, 2009; Sauquet et al., 2012). Gene pairs associated with the recent WGD event in pecan were used for circos mapping among chromosomes (Krzywinski et al., 2009).

To obtain the syntenic relationships between pecan and walnut or bayberry, Blast v. 2.2.6 (Boratyn et al., 2012) was used to identify the syntenic gene pairs between species with "-e 1e-6" and other default parameters. The results were then used for syntenic mapping with MCScanX (Wang et al., 2012) using default parameters.

Insertion time estimates of all LTRs and *Gypsy* and *Copia* elements were obtained as described by Huang et al. (2019) with the model T = K/2r ($r = 1.3 \times 10^{-8}$ per site and per year).

Defining the two paleo-subgenomes

To define the subgenomes of pecan, protein sequences from the v. 2.0 pecan assembly and from the bayberry and walnut genomes were used to generate clusters for gene families. All protein sequences of bayberry and walnut were downloaded from the NCBI database. Orthologous genes with ratios of 2:2:1 in pecan, walnut, and bayberry were selected, and orthologous gene pairs located on two different chromosomes with syntenic relationships in pecan and walnut were filtered out and connected into super sequences according to their chromosomes. The super sequences were aligned with MUSCLE v. 3.8.31 (Edgar, 2004). Regions with gaps were removed using Gblocks v. 0.91b (Talavera and Castresana, 2007) to generate eight chromosome groups (including two chromosomes from pecan, two from walnut, and one from bayberry), and an ML tree was constructed for each group using FastTree (Price et al., 2010) and displayed using MEGA7 (Kumar et al., 2016). Subgenomes A and B and the chromosome numbers of pecan were defined based on evolutionary distance.

Subgenome features

Orthologous gene pairs of sub A and sub B were determined by bilateral Blast searches against the bayberry gene set, and orthologs with the best

identity to bayberry genes in the sub A or sub B genome were selected. A set of 6316 orthologous gene pairs with the best identities was obtained, and MUSCLE v. 3.8.31 (Edgar, 2004) was used for alignment of the gene pairs with codons. KaKs_calculator (v. 2.0) software was used to estimate the *Ka*, *Ks*, and *Ka*/*Ks* values using the NG method (Wang et al., 2010). Frequency distribution histograms and scatterplots of the *Ka*, *Ks*, and *Ka*/*Ks* values were displayed using the ggplot2 package in the R language, and curve fitting of the scatterplots was performed using the "lm" method (Wickham, 2016).

Comparative genome analyses and phylogenetics

To investigate the evolutionary status of subgenomes in the Juglandaceae, protein sequences from the Cil_v. 2.0 pecan assembly and from three reference species (*Arabidopsis*, bayberry, and walnut) were used to generate clusters for gene families. All protein sequences of the three species were downloaded from the NCBI database. Genes with frame shifts that encoded fewer than 30 amino acids and redundant copies in each species were removed, and only the longest transcripts for each gene were selected for further analysis to ensure the analysis quality. To compare orthologous genes from the references with the protein-coding genes from the current pecan assembly (v. 2.0), all one-to-one orthologous gene sets were identified by BLASTP (Altschul et al., 1990) with an E value cutoff of 1×10^{-5}, and similar genes were clustered into families using hcluster, a hierarchical clustering algorithm in the TreeFam v. 0.50 pipeline (Li et al., 2006). All the gene families were aligned with the multi-sequence alignment software MUSCLE v. 3.8.31 (Edgar, 2004). To consider sequence conservation, the aligned single-copy genes from different gene families were further concatenated into super long sequences for the subgenomes of each species using a perl script. An ML phylogenetic tree was constructed using RAxML v. 8.2.4 (Stamatakis, 2006) with the PROTGAMMAAUTO option to automatically determine the optimal amino acid substitution site model; *Arabidopsis* was used as an outgroup, and branch confidence settings were based on 100 bootstrap replicates. The ML tree was used as a starting tree to infer the divergence times between species or subgenomes using the MCMCTree program in the PAML package (Yang, 1997). The calibration times for the divergence between *Arabidopsis* and walnut (98–117 mya) and between walnut and bayberry (43–74 mya) were obtained from the TimeTree database (http://timetree.org/). The divergence time between bayberry and walnut was calibrated based on the earliest fossil records of Myricaceae and Juglandaceae (64–84 mya) (Ho and Phillips, 2009; Sauquet et al., 2012). The common and lost genes among and/or within species were determined based on the results of homologous alignment.

Transcription factors in the A and B subgenomes were predicted by combining homologous searches against *Arabidopsis*, walnut, and bayberry transcription factors in the PlantTFDB v. 3.0 database (http://planttfdb.gao-lab.org/) based on the core domain structure using a hidden Markov model with further manual correction. The visualized heatmap of TF quantity distribution was generated by R script homogenization processing.

Genome resequencing, SNP calling, quality control, and validation

Genomic DNA was extracted from 86 individuals using CTAB methods. One microgram of high-quality DNA from each sample was used for genome resequencing library construction with the MGIEasy DNA Rapid Library Prep Kit (BGI, catalog no. 1000006985), and libraries were sequenced on the BGISEQ-500 platform following the manufacturer's protocol. The raw data (PE100) were filtered using SOAPnuke v. 1.5.6 (Chen et al., 2018) to remove reads with adapters or poly Ns and low-quality reads (reads in which >30% bases had Phred quality ≤25). The quality-controlled reads were then aligned to the Cil_v. 2.0 pecan assembly for SNP and indel calling by Sentieon, a pipeline that integrates BWA (Burrows–Wheeler Aligner) and GATK (Genome Analysis Tool Kit) (Kendig et al., 2019). The Haplotype Caller module of GATK was used for variant calling in the two subgenomes, and the concordance variants were filtered with the parameters "QD < 2.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || FS>60.0 || SOR>3.0". The indels were further filtered with "QD < 2.0 || ReadPosRankSum < -20.0 || FS > 200.0 || SOR>10.0".

Genetics and diversity analysis of the pecan scab-resistant and susceptible accessions

SNPs in subgenome A, subgenome B, and the whole genome of each sample were used to build NJ phylogenetic trees with 1000 bootstrap replicates using TreeBeST (http://treesoft.sourceforge.net/treebest.shtml), and the trees were visualized using the iTOL online tool (Letunic and Bork, 2019; http://itol.embl.de). To obtain a better alignment result, we did not split the subgenomes into two parts for processing. The internal structures of the two phylogenetic trees were compared using the phylo.io online tool (http://phylo.io/index.html). Structures of the accessions on the subgenome scale were analyzed with ADMIXTURE (Alexander and Lange, 2011).

The sampled accessions were divided into two groups based on their pecan scab-resistance grades: the scab-resistant group (R, grade number ≤ 2) and the scab-susceptible group (S, grade number > 2). To determine the pairwise genetic diversity *Pi* (π) and the fixation index *Fst* of the R and S groups, vcftools software (Danecek et al., 2011; http://vcftools.sourceforge.net/) was used with a 100-kb sliding window. Chromosome regions whose values of *Pi* ratio (π_S/π_R) and *Fst* were both in the highest 5% were selected for further analysis.

Hot-block linkage disequilibrium (LD) mapping of the chromosome regions of interest above was visualized using LDBlockShow (Dong et al., 2020) with default parameters.

Chitin treatment and real-time qPCR validation

To validate our results, fully expanded leaves from the scab-resistant cultivar Excell and the scab-susceptible cultivar Pawnee were collected and subjected to chitin (100 μg/ml) treatment for 0, 30, 60, and 180 min (Figure 4B). At least three biological replicates of each sample type were collected for RNA extraction using the RNAprep Pure Plant Kit (TIANGEN, Beijing, China), and cDNA was obtained using the SuperScript III First-Strand Synthesis System (Takara, Dalian, China). Eleven putative genes that were involved in the chitin signaling pathway and ROS elimination were selected for expression analysis, and the 18S rRNA gene was used as the internal control (Mattison et al., 2017). Real-time qPCR was performed on an ABI 7500 Real-Time PCR System (Foster City, CA, USA) with three technical replicates for each gene in each sample. Gene expression levels were calculated using the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001). The genes and primers are listed in supplemental Table 25. Significant differences in relative gene expression levels between samples were determined using the SPSS program (Kretzschmar, 2000).

Data availability

Genome sequences, assembly, and annotation data have been deposited at NCBI GenBank under BioProject/BioSample numbers PRJNA727440/SAMN19020793. Resequencing reads for the 86 individuals have been deposited in the Sequence Read Archive (SRA) under BioProject/BioSample numbers PRJNA735040/SAMN19554720–19554805.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at *Plant Communications Online*.

FUNDING

This work was supported by grants from the Natural Science Foundation of Zhejiang Province, China (grant no. Z20C160001), the State Key Laboratory of Subtropical Silviculture at Zhejiang A&F University (grant no.

ZY20180202), and the Research and Development Fund of Zhejiang A&F University (grant no. 2018FR002).

AUTHOR CONTRIBUTIONS

L.X. and G.F. designed and supervised the research. L.X. wrote the paper. L.X., M.Y., R.Z., H.G., X.G., H.Z., T.D., and G.F. performed the genome assembly, annotation, and evolution analysis. J.H. and G.F. performed the population structure and genetic diversity analysis. L.X., Y.Z., J.W., S.L., and X.L. collected and prepared all samples and performed the experiments. J.H. provided valuable suggestions for the project.

ACKNOWLEDGMENTS

The authors declare no competing interests.

Received: June 1, 2021
Revised: August 18, 2021
Accepted: September 22, 2021
Published: September 24, 2021

REFERENCES

de A. Gerhardt, L.B., Sachetto-Martins, G., Contarini, M.G., Sandroni, M., de P. Ferreira, R., de Lima, V.M., Cordeiro, M.C., de Oliveira, D.E., and Margis-Pinheiro, M. (1997). *Arabidopsis thaliana* class IV chitinase is early induced during the interaction with *Xanthomonas campestris*. FEFS Lett. **419**:69–75.

Aggarwal, G., and Ramaswamy, R. (2002). Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER. J. Biosci. **27**:7–14.

Alexander, D.H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics **12**:246.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. **28**:45–48.

Beedanagari, S.R., Dove, S.K., Wood, B.W., and Conner, P.J. (2005). A first linkage map of pecan cultivars based on RAPD and AFLP markers. Theor. Appl. Genet. **110**:1127–1137.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. **27**:573–580.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and genomewise. Genome Res. **14**:988–995.

Bock, C.H., Grauke, L.J., Conner, P., Burrell, S.L., Hotchkiss, M.W., Boykin, D., and Wood, B.W. (2016). Scab susceptibility of a provenance collection of pecan in three different seasons in the southeastern United States. Plant Dis. **100**:1937–1945.

Bock, C.H., Hotchkiss, M.W., Young, C.A., Charlton, N.D., Chakradhar, M., Stevenson, K.L., and Wood, B.W. (2017). Population genetic structure of *Venturia effusa*, cause of pecan scab, in the southeastern United States. Phytopathology **107**:607–619.

Bock, C.H., Young, C.A., Stevenson, K.L., and Charlton, N.D. (2018). Fine-scale population genetic structure and within-tree distribution of mating types of *Venturia effusa*, cause of pecan scab in the United States. Phytopathology **108**:1326–1336.

Bock, C.H., Alarcon, Y., Conner, P.J., Young, C.A., Randall, J.J., Pisani, C., Grauke, L.J., Wang, X., and Monteros, M.J. (2020a). Foliage and fruit susceptibility of pecan provenance collection to scab, caused by *Vernturia effusa*. CABI Agric. Biosci. **1**:19.

Bock, C.H., Barbedo, J.G.A., Del Ponte, E.M., Bohnenkamp, D., and Mahlein, A.-K. (2020b). From visual estimates to fully automated sensor-based measurements of plant disease severity: status and challenges for improving accuracy. Phytopathol Res. **2**:9.

Boratyn, G.M., Schäffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J., and Madden, T.L. (2012). Domain enhanced lookup time accelerated BLAST. Biol. Direct **7**:12.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. **31**:1119–1125.

Campbell, M.A., Hass, B.J., Hamilton, J.P., Mount, S.M., and Buell, C.B. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. BMC Genomics **7**:327.

Cao, Y., Jiang, L., Wang, L., and Cai, Y. (2019). Evolutionary rate heterogeneity and functional divergence of orthologous genes in *Pyrus*. Biomolecules **9**:490.

Chaney, W., Han, Y., Rohla, C., Monteros, M.J., and Grauke, L.J. (2015). Developing molecular marker resources for pecan. Acta Hort. **1070**:127–132.

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., et al. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience **7**:1–6.

Cheng, Y., Tian, Q.Y., and Zhang, W.H. (2016). Glutamate receptors are involved in mitigating effects of amino acids on seed germination of *Arabidopsis thaliana* under salt stress. Environ. Exp. Bot. **130**:68–78.

Conner, P.J. (2012). Pecan breeding review. Pecan South **45**:34–44.

Conner, P.J., and Wood, B.W. (2001). Identification of pecan cultivars and their genetic relatedness as determined by randomly amplified polymorphic DNA analysis. J. Am. Soc. Hort. Sci. **126**:474–480.

Consortium, G.O. (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. **32**:D258–D261.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., and Sherry, S.T. (2011). The variant call format and VCFtools. Bioinformatics **27**:2156–2158.

Dodds, P.N., and Rathjen, J.P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. Nat. Rev. Genet. **11**:539–548.

Dong, S., He, W., Ji, J., Zhang, C., Guo, Y., and Yang, T. (2020). LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. Brief. Bioinformatics **22**:bbaa227.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., et al. (2017). *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science **356**:92–95.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. **3**:95–98.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**:1792–1797.

Elsik, C.G., Mackey, A.J., Reese, J.T., Milshina, N.V., Roos, D.S., and Weinstock, G.M. (2007). Creating a honey bee consensus gene set. Genome Biol. **8**:R13.

Fawcett, J.A., Maere, S., and Van de Peer, Y. (2009). Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc. Natl. Acad. Sci. U S A **106**:5737–5742.

Goff, W.D., McVay, J.R., and Gazaway, W.S. (1996). Pecan Production in the Southeast. In Alabama Cooperative Extension System Circular ANR-459 (Auburn: University), p. 222.

Gong, B.-Q., Wang, F.-Z., and Li, J.-F. (2020). Hide-and-Seek: chitin-triggered plant immunity and fungal counterstrategies. Trends Plant Sci. **25**:805–816.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, Z.J., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2013). Trinity: reconstructing a full-length transcriptome without a genome from RNA-seq data. Nat. Biotechnol. **29**:644–652.

Grauke, L.J., Iqbal, M.J., Reddy, A.S., and Thompson, T.E. (2003). Developing microsatellite DNA markers in pecan. J. Am. Soc. Hortic. Sci. **128**:374–380.

Grauke, L.J., Wood, B.W., and Harris, M. (2016). Crop vulnerability: *Carya*. HortScience **51**:653–663.

Ho, S.Y.W., and Phillips, M.J. (2009). Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Syst. Biol. **58**:367–380.

Huang, S.W., Li, R.Q., Zhang, Z.H., Li, L., Gu, X.F., Fan, W., Lucas, W.J., Wang, X.W., Xie, B.Y., Ni, P.X., et al. (2009). The genome of the cucumber, *Cucumis sativus* L. Nat. Genet. **41**:1275–1281.

Huang, Y., Xiao, L., Zhang, Z., Zhang, R., Wang, Z., Huang, C., Huang, R., Luan, Y., Fan, T., Wang, J., et al. (2019). The genomes of pecan and Chinese hickory provide insights into *Carya* evolution and nut nutrition. Gigascience **8**:giz036.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Nathalie, C., Sébastien, A., Nicola, V., Claire, J., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449**:463–467.

Jenkins, J., Wilson, B., Grimwood, J., Schmutz, J., and Grauke, L.J. (2015). Towards a reference pecan genome sequence. Acta Hort. **1070**:101–108.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A.L., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics **30**:1236–1240.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. **110**:462–467.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.E., Bateman, A., Finn, R.D., and Petrov, A.I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nacleic Acids Res. **46**:D335–D342.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. **28**:27–30.

Kang, J., and Turano, F.J. (2003). The putative glutamate receptor 1.1 (AtGLR1.1) functions as a regulator of carbon and nitrogen metabolism in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. U S A **100**:6872–6877.

Kang, J., Sohum, M., and Turano, F.J. (2004). The putative glutamate receptor 1.1 (AtGLR1.1) in *Arabidopsis thaliana* regulates abscisic acid biosynthesis and signaling to control development and water loss. Plant Cell Physiol. **45**:1380–1389.

Kang, S., Kim, H.B., Lee, H., Choi, J.Y., Heu, S., Oh, C.J., Kwon, S.I., and An, C.S. (2006). Overexpression in *Arabidopsis* of a plasma membrane-targeting glutamate receptor from small radish increases glutamate-mediated Ca^{2+} influx and delays fungal infection. Mol. Cell **21**:418–427.

Kendig, K.I., Baheti, S., Bockol, M.A., Drucker, T.M., Hart, S.N., Heldenbrand, J.R., Hernaez, M., Hudson, M.E., Kalmbach, M.T., Klee, E.W., et al. (2019). Sentieon DNASeq variant calling workflow demonstrates strong computational performance and accuracy. Front. Genet. **10**:736.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. **27**:722–736.

Kretzschmar, W.A. (2000). SPSS student version 9.0 for Windows[J]. J. Engl. Linguist. **28**:311–313.

Krzywinski, M., Schein, J.E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. **19**:1639–1645.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. **33**:1870–1874.

Landis, J.B., Soltis, D.E., Li, Z., Marx, H.E., Barker, M.S., Tank, D.C., and Soltis, P.S. (2017). Impact of whole-genome duplication events on diversification rates in angiosperms. Am. J. Bot. **105**:348–363.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10**:R25.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**:1754–1760.

Li, H., Coghlan, A., Ruan, J., Coin, L.J.M., Hériché, J.K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L., et al. (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. **34**:D572.

Li, F., Wang, J., Ma, C., Zhao, Y., Wang, Y., Hasi, A., and Qi, Z. (2013). Glutamate receptor-like channel3.3 is involved in mediating glutathione-triggered cytosolic calcium transients, transcriptional changes, and innate immunity responses in *Arabidopsis*. Plant Physiol. **162**:1497–1509.

Li, H., Jiang, X., Lv, X., Ahammed, G.J., Guo, Z., Yu, J., and Zhou, Y. (2019). Tomato GLR3.3 and GLR3.5 mediate cold acclimation-induced chilling tolerance by regulating apoplastic H_2O_2 production and redox homeostasis. Plant Cell Environ. **42**:3326–3339.

Liu, Y., Huang, X., Li, M., and Zhang, Y. (2016). Loss-of-function of *Arabidopsis* receptor-like kinase BIR1-activates cell death and defense responses mediated by BAK1 and DOBIR1. New Phytol. **212**:637–645.

Lovell, J.T., Bentley, N.B., Bhattarai, G., Jenkins, J.W., Sreedasyam, A., Alarcon, Y., Bock, C., Boston, L.B., Carlson, J., Cervantes, K., et al. (2021). Four chromosome scale genomes and a pan-genome annotation to accelerate pecan tree breeding. Nat. Commun. **12**:4125.

Lowe, T.M., and Eddy, S.R. (1996). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**:955–964.

Luo, M.-C., You, F.M., Li, P., Wang, J.-R., Zhu, T., Dandekar, A.M., Leslie, C.A., Aradhya, M., McGuire, P.E., and Dvorak, J. (2015). Synteny analysis in Rosids with a walnut physical map reveals slow genome evolution in long-lived woody perennials. BMC Genomics **16**:707.

Ma, L., Tian, t., Lin, R., Deng, X., Wang, H., and Li, G. (2016). *Arabidopsis* FHY3 and FAR1 regulate light-induced myo-inositol biosynthesis and oxidative stress responses by transcriptional activation of MIPS1. Mol. Plant **9**:541–557.

Manchester, S.R. (1989). Early history of the Juglandaceae. Plant Syst. Evol. **162**:231–250.

Manzoor, H., Kelloniemi, J., Chiltz, A., Wendehenne, D., Pugin, A., Poinssot, B., and Garcia-Brugger, A. (2013). Involvement of the glutamate receptor AtGLR3.3 in plant defense signaling and resistance to *Hyaloperonospora arabidopsidis*. Plant J. **76**:466–480.

Mattison, C.P., Rai, R., Settlage, R.E., Hinchliffe, D.J., Madison, C., Bland, J.M., Brashear, S., Graham, C.J., Tarver, M.R., Florane, C., et al. (2017). RNA-seq analysis of developing pecan (*Carya illinoinensis*) embryos reveals parallel expression patterns among allergen and lipid metabolism genes. J. Agric. Food Chem. **65**:1443–1455.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics **29**:2933–2935.

Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L., and Vandepoele, K. (2009). The flowering world: a tale of duplications. Trends Plant Sci. **14**:680–688.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One **5**:e9490.

Salse, J. (2012). In silico archeogenomics unveils modern plant genome organisation, regulation and evolution. Curr. Opin. Plant Biol. **15**:122–130.

Sauquet, H., Ho, S.Y.W., Gandolfo, M.A., Jordan, G.J., Wilf, P., Cantrill, D.J., Bayly, M.J., Bromham, L., Brown, G.K., Carpenter, R.J., et al. (2012). Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of *Nothofagus* (Fagales). Syst. Biol. **61**:289–313.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. **16**:259.

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

Singh, S.K., Chien, C.T., and Chang, I.F. (2006). The *Arabidopsis* glutamate receptor-like gene GLR3.6 controls root development by repressing the kip-related protein gene KRP4. J. Exp. Bot. **67**:1853–1869.

Soltis, D.E., and Burleigh, J.G. (2009). Surviving the K-T mass extinction: new perspectives of polyploidization in angiosperms. Proc. Natl. Acad. Sci. U S A **106**:5455–5456.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688–2690.

Stanke, M., Steinkamp, R., Waack, S., and Morgensternet, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. **32**:309–312.

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. **56**:564–577.

Tena, G., Boudsocq, M., and Sheen, J. (2011). Protein kinase signaling networks in plant innate immunity. Curr. Opin. Plant Biol. **14**:519–529.

Thompson, T.E., and Conner, P.J. (2012). Pecan. In Fruit Breeding, Handbook of Plant Breeding, *8* (USA: Springer)), pp. 771–801.

Thompson, T.E., and Grauke, L.J. (1994). Genetic resistance to scab disease in pecan. HortScience **29**:1078–1080.

Thompson, T.E., and Hunter, R.E. (1985). Pawnee pecan. Horts **20**:776.

Tsutomu, K., Yamada, K., Yoshimura, S., and Yamaguchi, K. (2017). Chitin receptor-mediated activation of MAP kinases and ROS production in rice and *Arabidopsis*. Plant Signal. Behav. **12**:e1361076.

Tuskan, G.A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313**:1596–1604.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., and Schatz, M.C. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics **34**:2202–2204.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouellie, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One **9**:e112963.

Wang, C., Wang, G., Zhang, C., Zhu, P., Dai, H., Yu, .N., He, Z., Xu, L., and Wang, E. (2017). OsCERK1-mediated chitin perception and immune signaling requires receptor-like cytoplasmic kinase 185 to activate an MAPK cascade in rice. Mol. Plant **10**:619–633.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). Kaks_calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genom. Proteom. Bioinform. **8**:77–80.

Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. **40**:e49.

Wang, W., Tang, W., Ma, T., Niu, D., Jin, J., Wang, H., and Lin, R. (2016). A pair of light signaling factors FHY3 and FAR1 regulates plant immunity by modulating chlorophyll biosynthesis. J. Integr. Plant Biol. **58**:91–103.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. **35**:543–548.

Wells, L. (2014). Pecan planting trends in Georgia. HortTechnology **24**:475–479.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis, 2nd edn (USA: Springer).

Wood, B.W., Conner, P.J., and Worley, R.E. (2003). Relationship of alternate bearing intensity in pecan to fruit and canopy characteristics. HortScience **38**:361–366.

Wudick, M.M., Michard, E., Nunes, C.O., and Feijó, J.A. (2018). Comparing plant and animal glutamate receptors: common traits but different fates? J. Exp. Bot. **69**:4151–4163.

Yamada, K., Yamaguchi, K., Yashiura, S., Terauchi, A., and Kawasaki, T. (2017). Conservation of chitin-induced MAPK signaling pathways in rice and *Arabidopsis*. Plant Cell Physiol. **58**:993–1002.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics **13**:555–556.

Zhang, J., Zhang, W., Ji, F., Qiu, J., Song, X., Bu, D., Pan, G., Ma, Q., Chen, J., Huang, R., et al. (2020). A high-quality walnut genome assembly reveals extensive gene expression divergences after whole-genome duplication. Plant Biotechnol. J. **18**:1848–1850.

Zhu, J.-K. (2016). Abiotic stress signaling and responses in plants. Cell **167**:313–324.

Zwaenepoel, A., and Peer, Y.V.de. (2019). wgd-simple command line tools for the analysis of ancient whole-genome duplications. Bioinformatics **35**:2153–2155.

Plant Communications, Volume *2*

Supplemental information

Chromosome-scale assembly reveals asymmetric paleo-subgenome evolution and targets for the acceleration of fungal resistance breeding in the nut crop, pecan

Lihong Xiao, Mengjun Yu, Ying Zhang, Jie Hu, Rui Zhang, Jianhua Wang, Haobing Guo, He Zhang, Xinyu Guo, Tianquan Deng, Saibin Lv, Xuan Li, Jianqin Huang, and Guangyi Fan

Title: **Chromosome-scale assembly reveals the asymmetric paleo-subgenomes evolution and targets for accelerating fungal resistance breeding in nut crop, pecan**

Authors: Lihong Xiao[1, 4, *], Mengjun Yu[2, 4], Ying Zhang[1, 4], Jie Hu[2], Rui Zhang[2], Jianhua Wang[1], Haobing Guo[2], He Zhang[2], Xinyu Guo[2], Tianquan Deng[3], Saibin Lv[1], Xuan Li[1], Jianqin Huang[1], Guangyi Fan[2, *]

Affiliations:

[1] State Key Laboratory of Subtropical Silviculture, Zhejiang A&F University, Hangzhou 311300, China

[2] BGI-Qingdao, BGI-Shenzhen, Qingdao 266555, China

[3] BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China

[4] Co-first authors

[*] Corresponding authors

Correspondence:

Lihong Xiao, Ph.D.

Tel. +86 0571-61083202

Email: xiaolh@zafu.edu.cn

Address: No. 666 Wusu St. Lin'an District, Hangzhou 311300, China

Guangyi Fan, Ph.D.

Tel. + 86 18576694373

Email: fanguangyi@genomics.cn

Address: No. 2 Hengyunshan Rd. Huangdao District, Qingdao 266000, China

Running title: The chromosome-level genome of pecan

Short summary:

A high-quality chromosome-scale reference genome of pecan reveals two paleo-subgenomes with asymmetry in their features and evolution. Re-sequencing on pecan scab-associated core accessions identifies several key genes in chitin response pathway that may be important susceptibility factors for fungal diseases and valuable resources for pecan breeders. The study provides an example for production and quality improvement of tree nut crops.

Supplemental Information (SI)

Supplemental Figures and legends

Figure S1. Distribution of Nanopore reads.

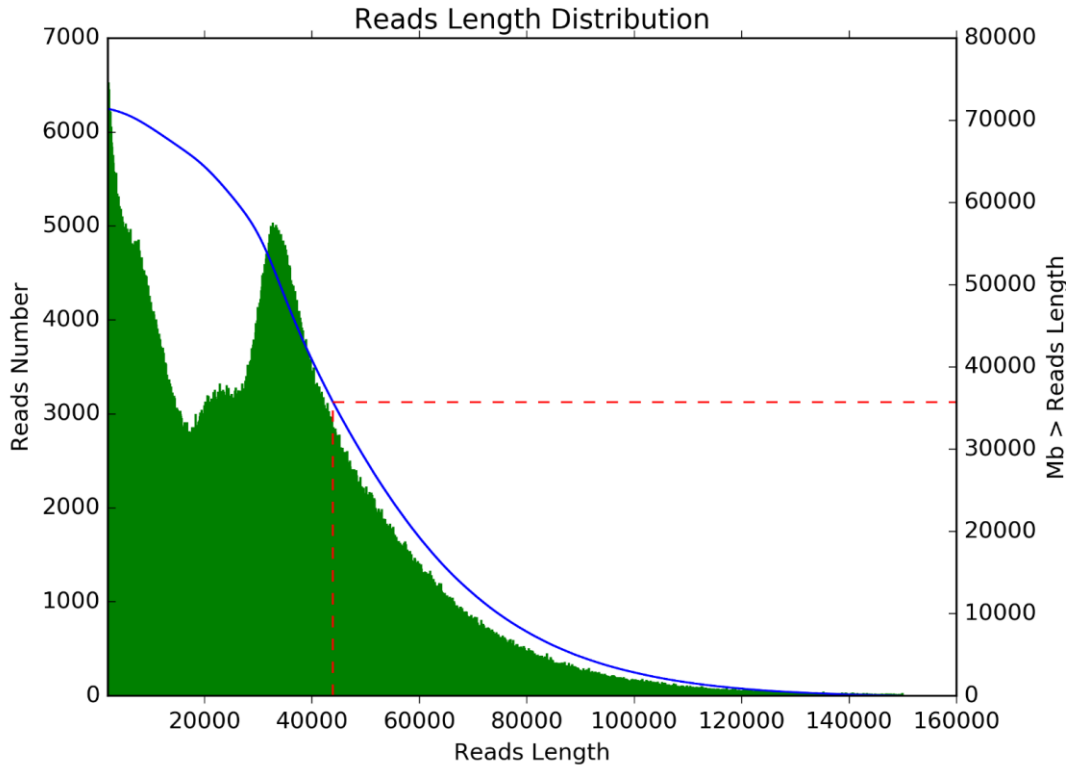


Figure S2. Genome survey by K-mer analysis.

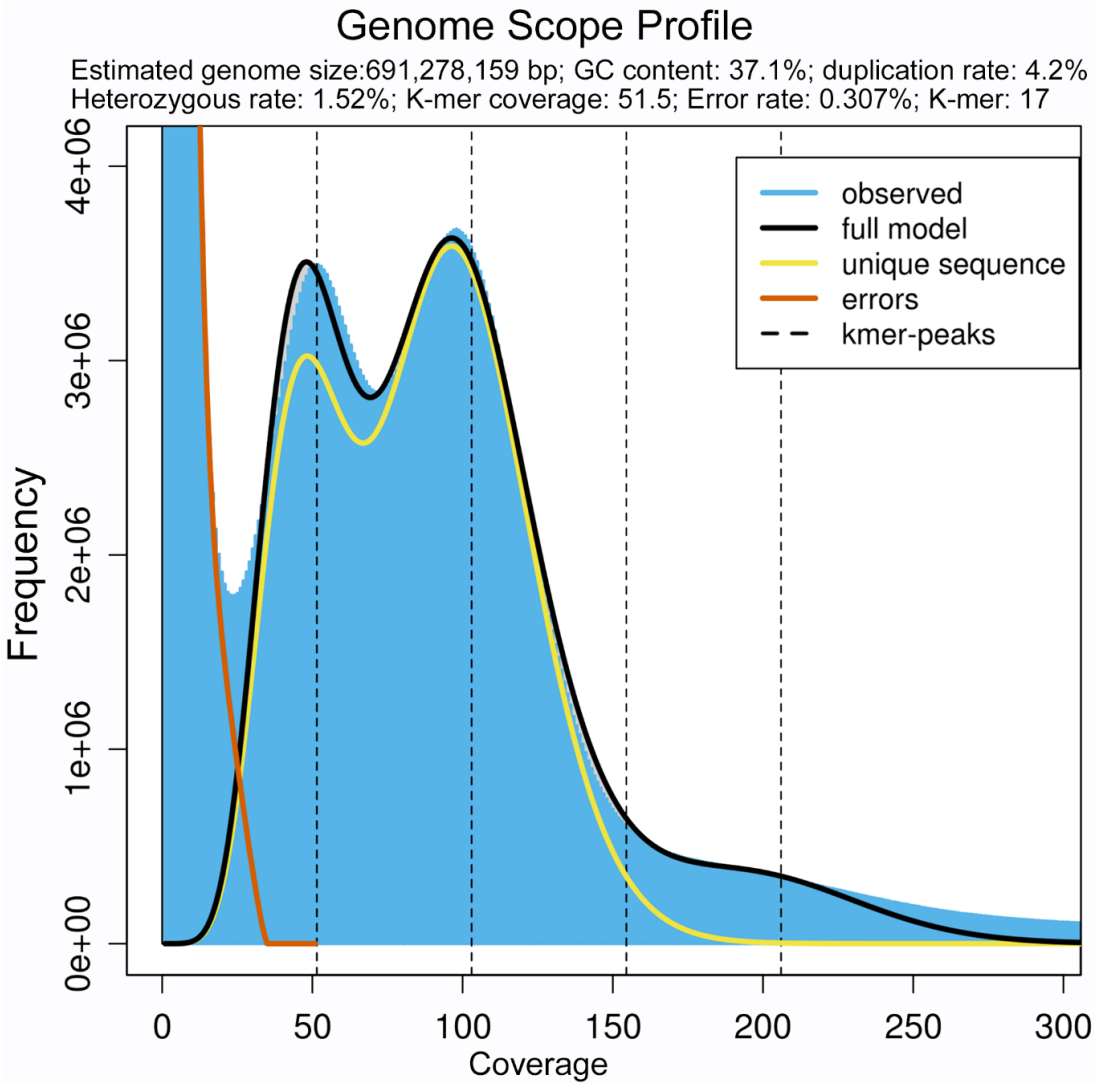


Figure S3. Heatmap of Hi-C assembly.

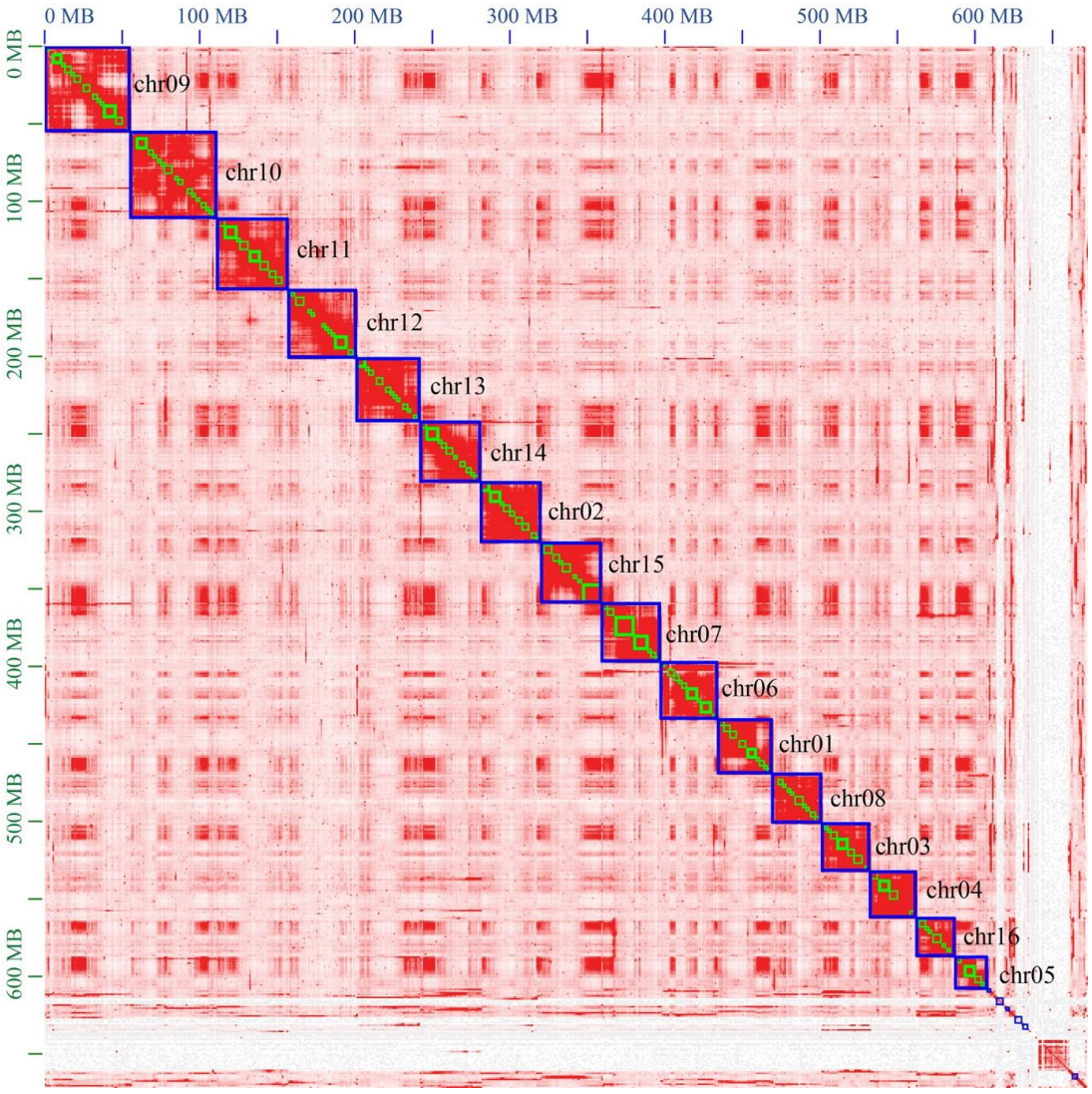


Figure S4. The distribution of GC content.

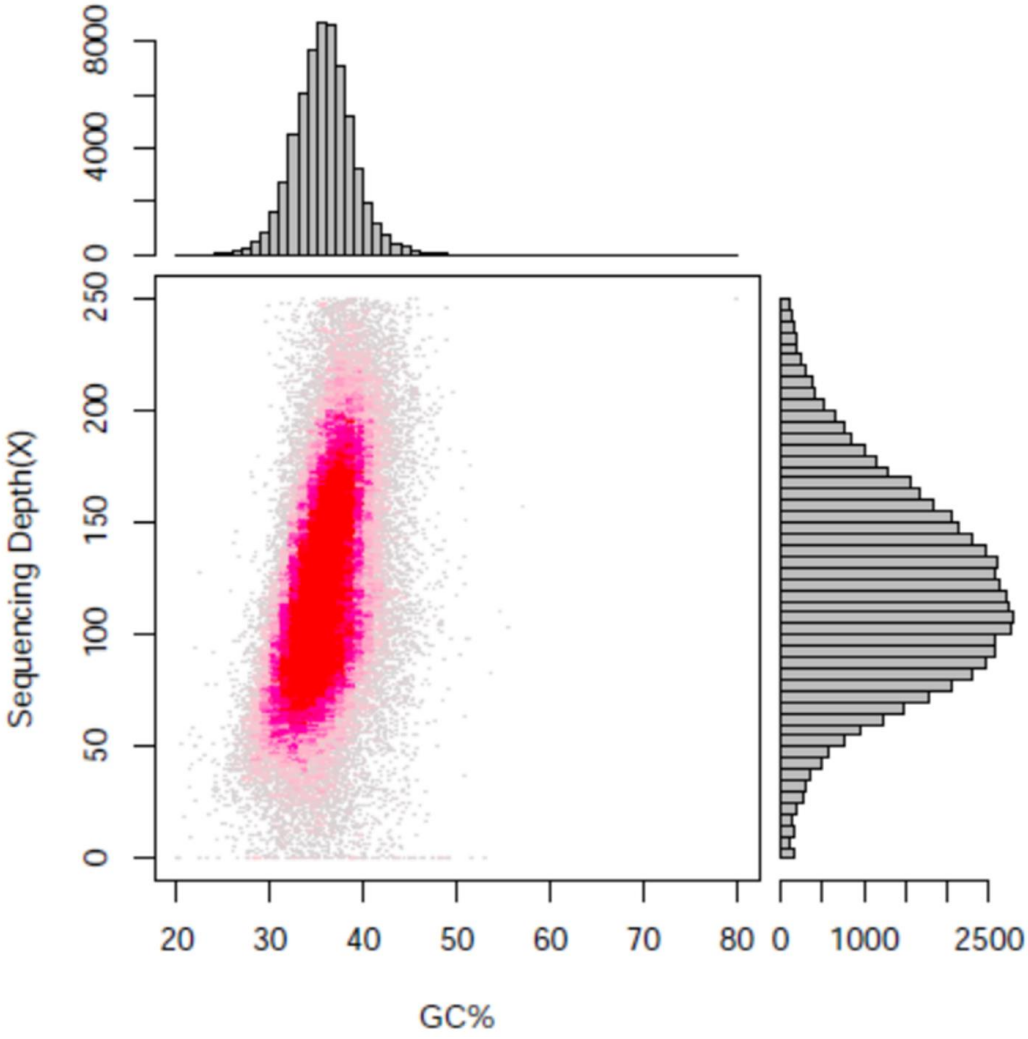


Figure S5. The contigs distribution of version 1.0 assembly on the chromosomes of version 2.0 assembly.

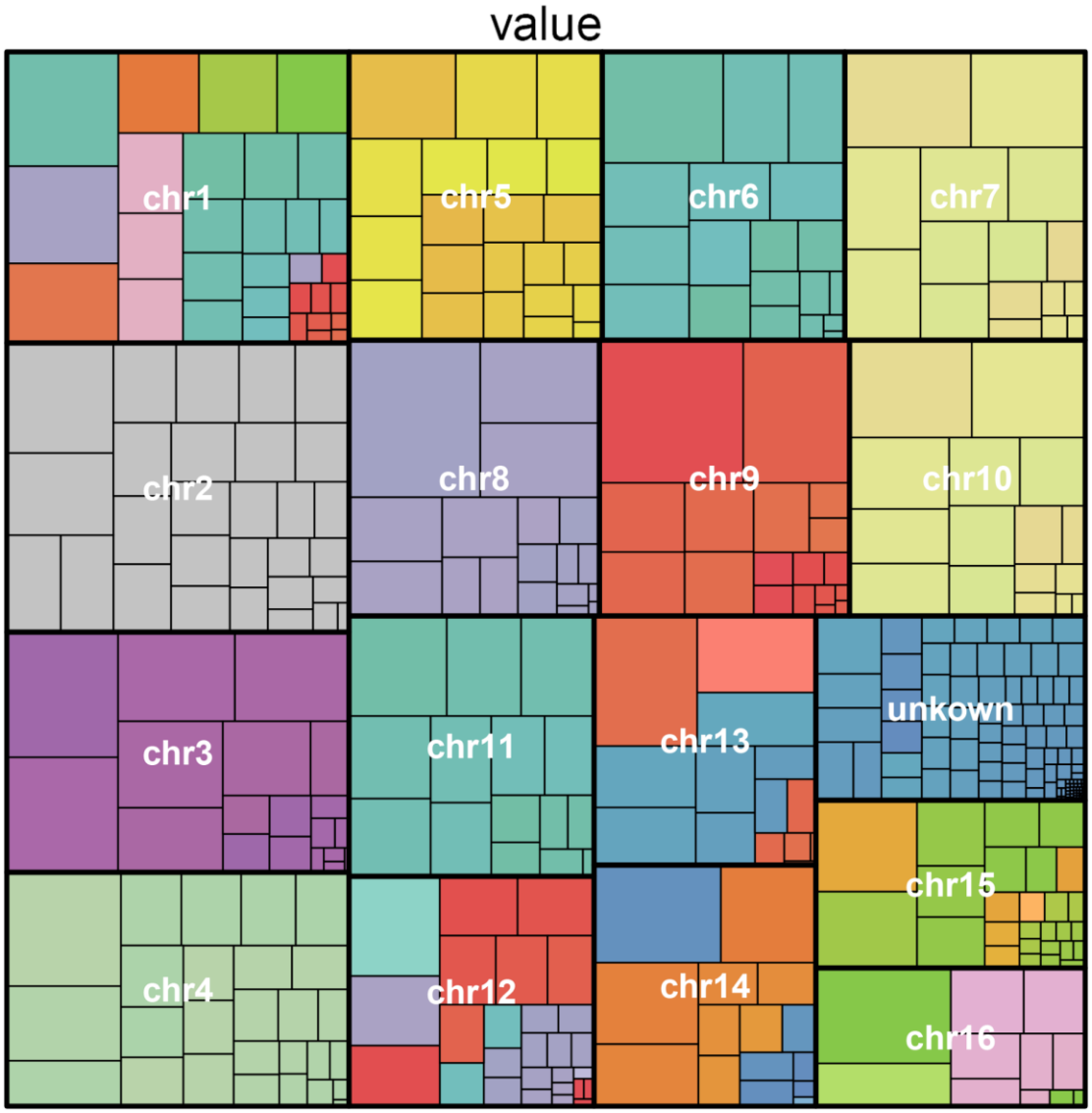


Figure S6. Global synteny between the gap-free ‘Pawnee’ assembly by Lovell et al. (2021) (chr1 – chr16) and Cil_V. 2.0 (CHR01 – CHR16).

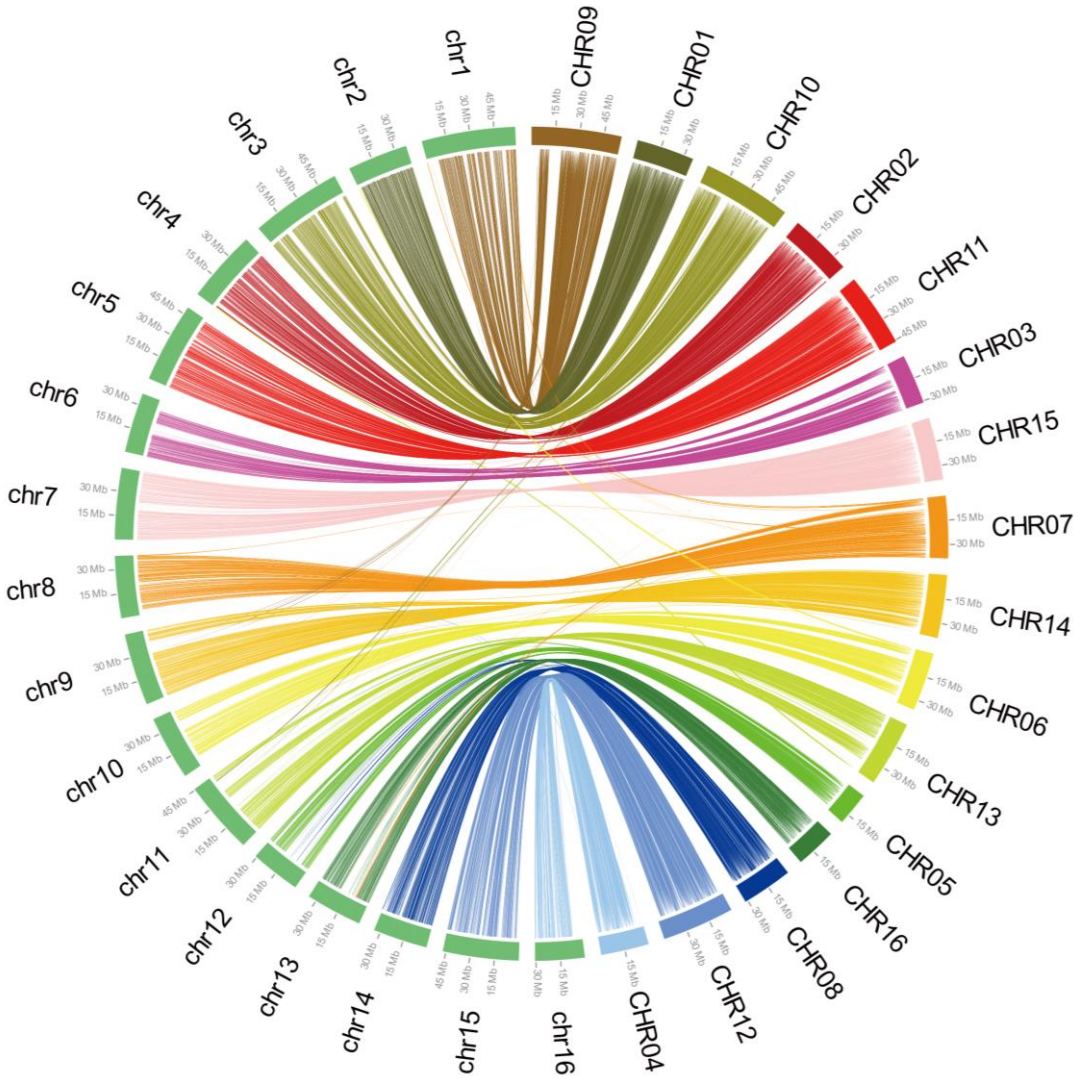


Figure S7. The details of syntenic relationship between the gap-free ‘Pawnee’ assembly by Lovell et al. (2021) (chr1 – chr16) and Cil_V. 2.0 (CHR01 – CHR16).

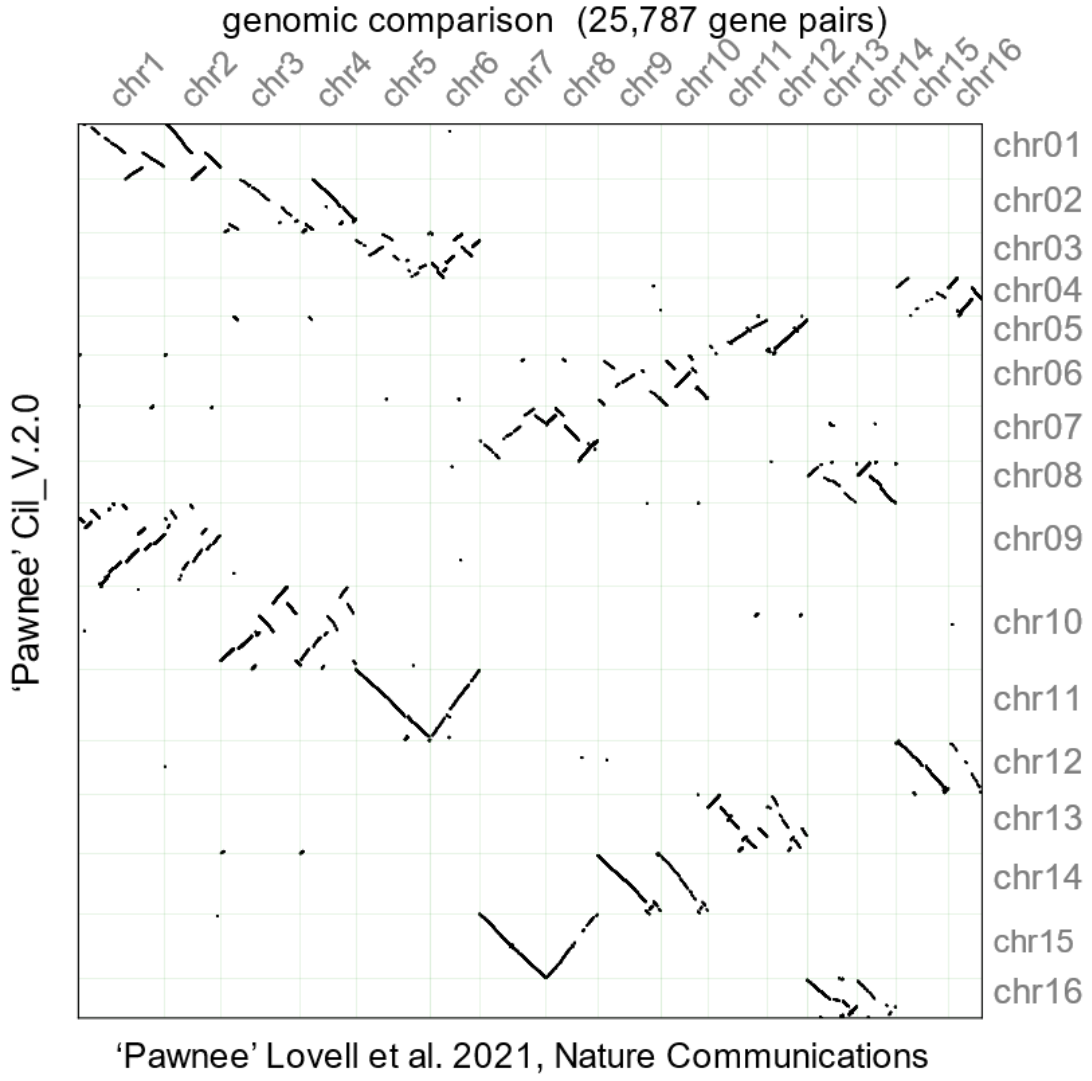


Figure S8. Evolutionary relationships of chromosomes among pecan, walnut and bayberry.

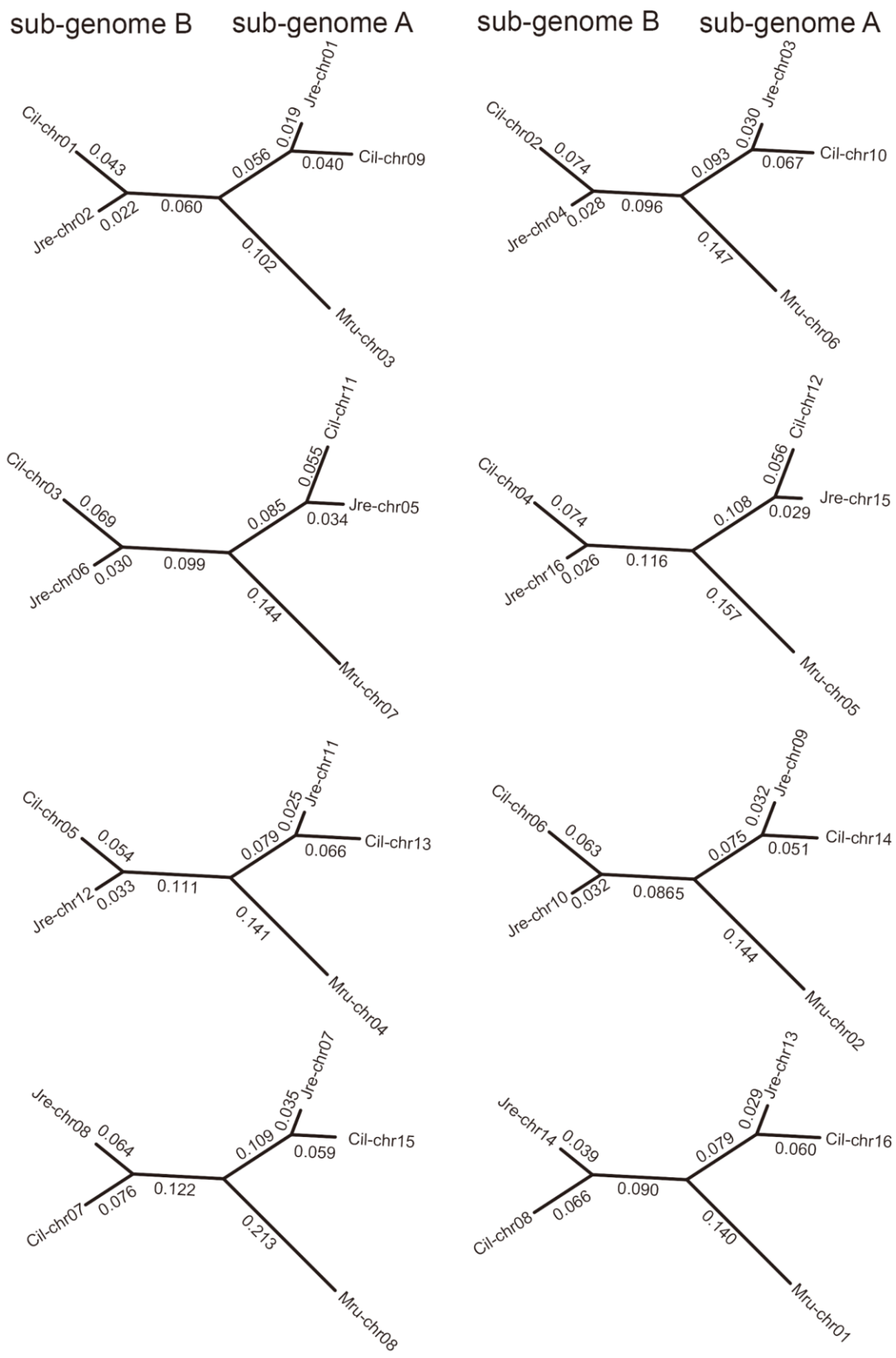


Figure S9. Identity analysis between subgenomes in pecan.

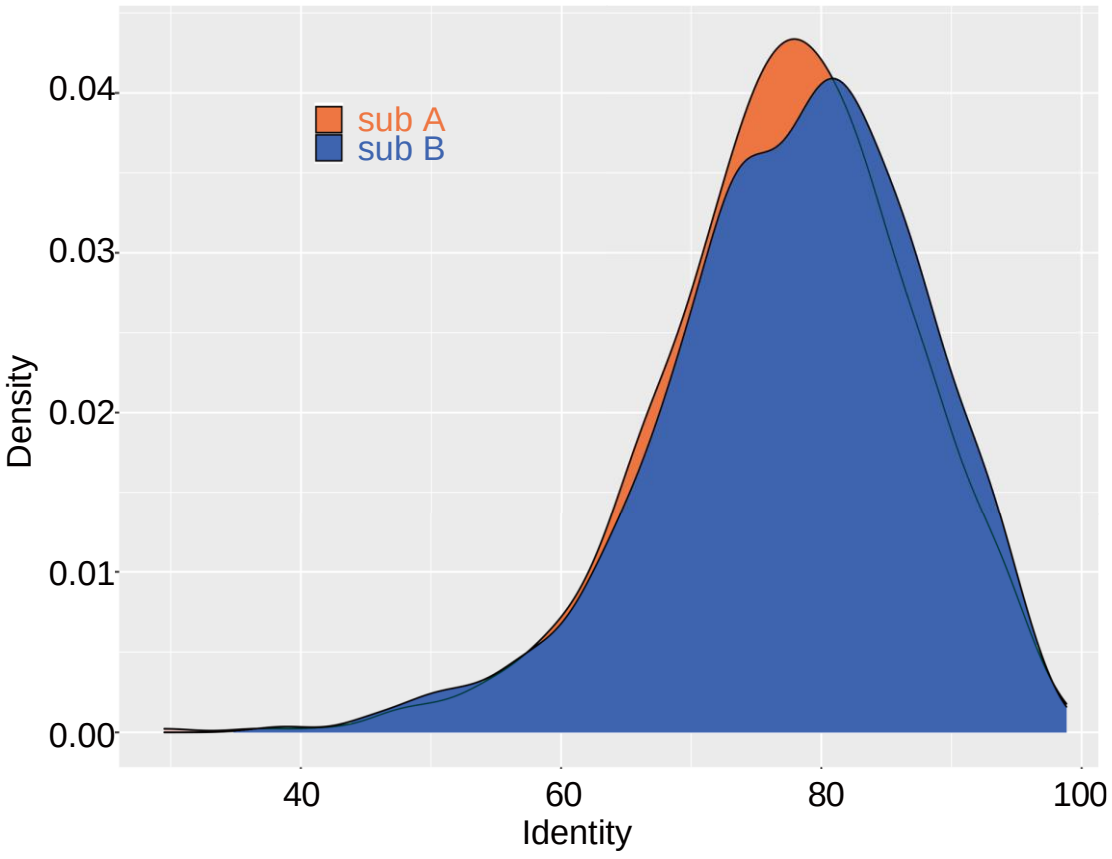


Figure S10. *Ka/Ks* analysis between subgenomes in pecan.

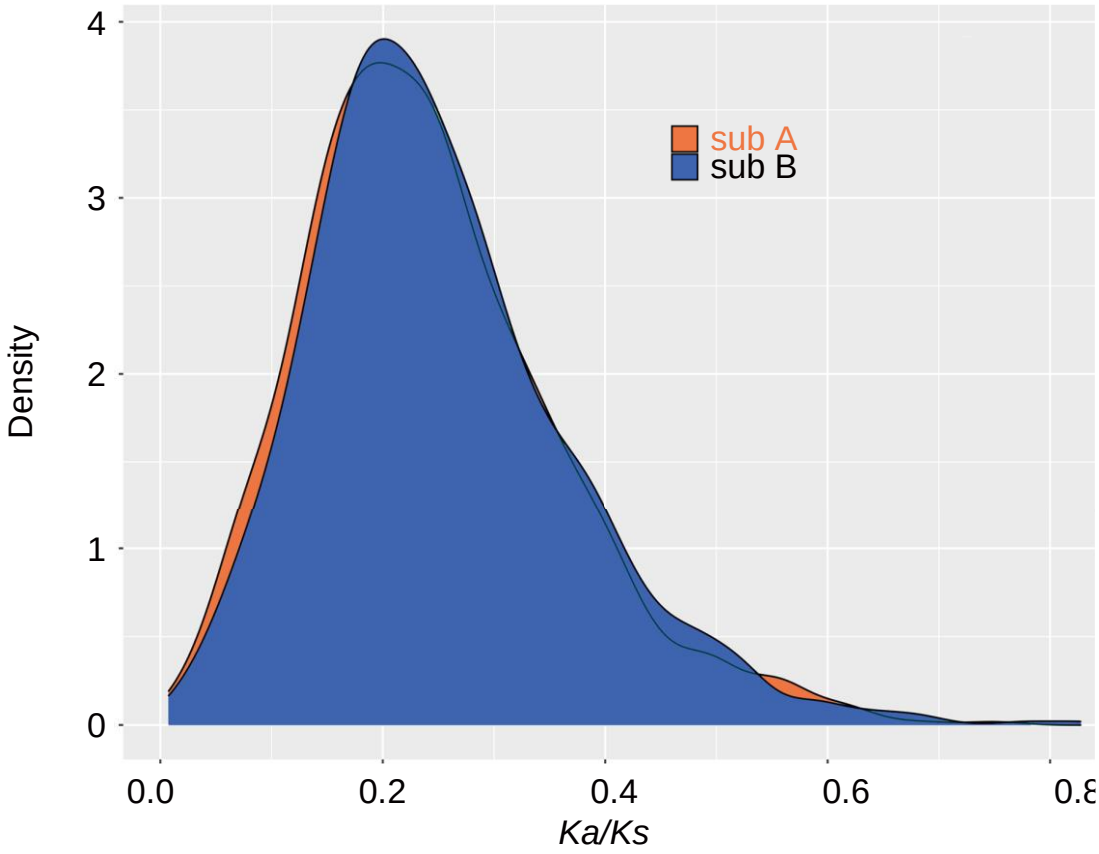


Figure S11. The frequency distributions (upper) and correlation analyses (lower) of *Ka*, *Ks*, and *Ka/Ks*. blue line represented the fitted curve, and the shaded part represented the confidence interval.

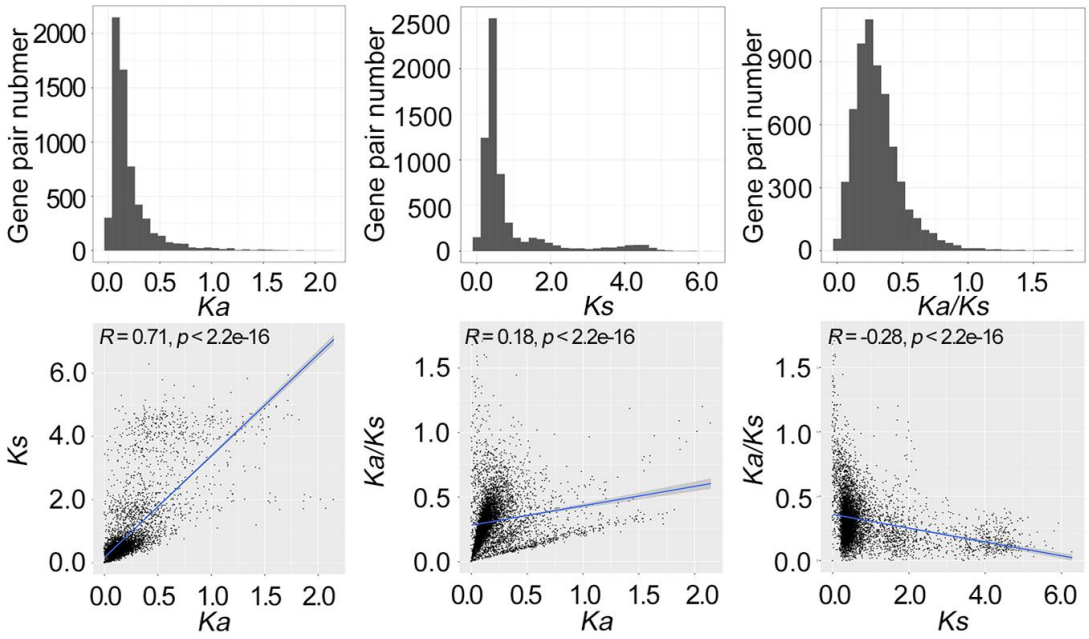
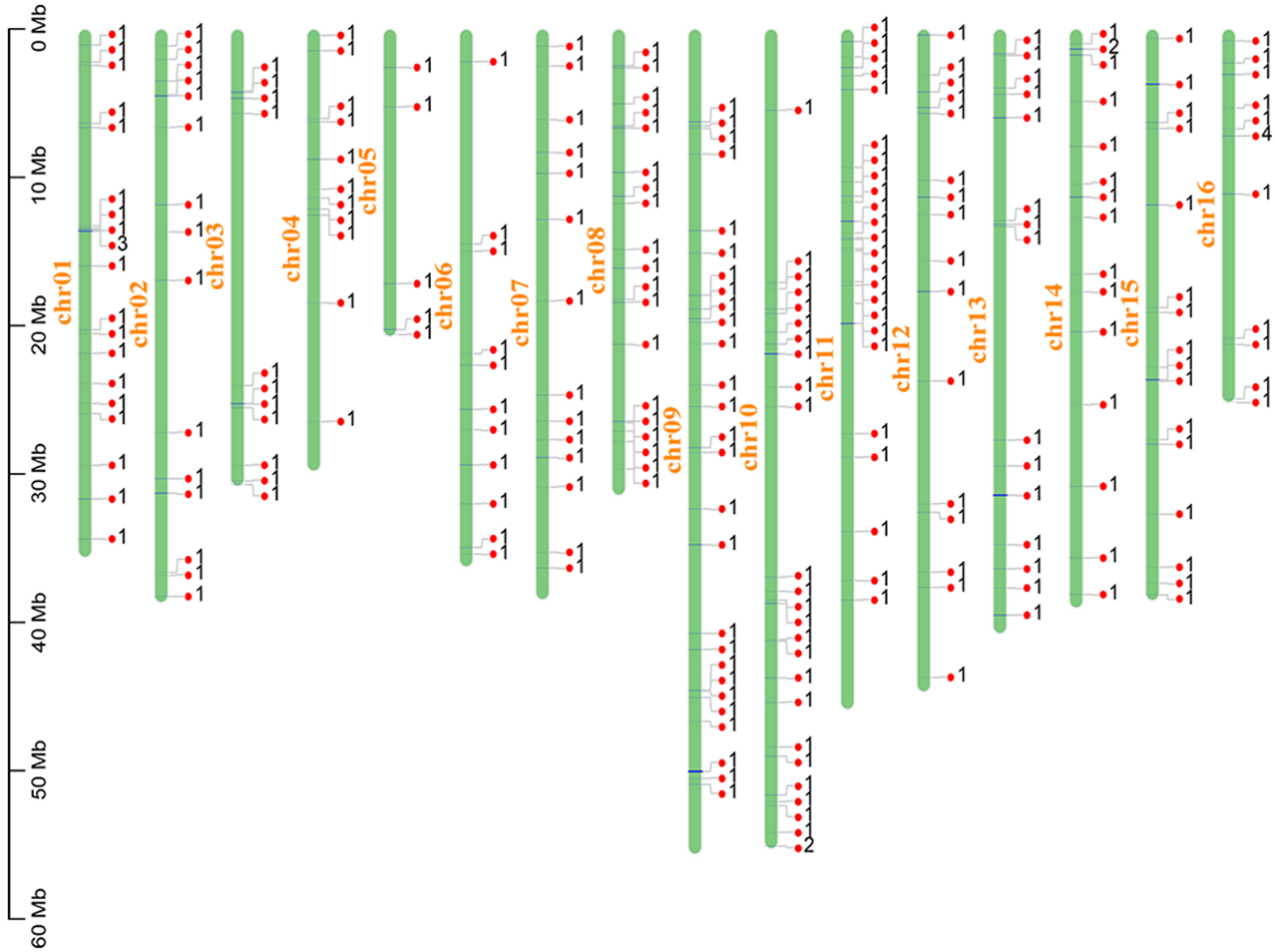


Figure S12. Distribution of *FAR1* transcription factor family members on chromosomes.



Supplemental Tables

Table S1. Data statistics for pecan genome survey and assembly.

Platform	Data type	Number of Reads	Total length (bp)	N50 Length of Reads	N90 Length of Reads	Average Read Length (bp)	Maxmum Read Length (bp)	Mean Qualigy
Nanopore	Raw Data	5,925,714	76,162,104,265	43,224	18,376	12,852	663,065	4.45
	Clean Data	2,184,040	71,697,987,348	44,087	20,238	32,828	663,065	8.84
Hiseq X-Ten (PE150)*	Raw Data	1,122,405,046	168,360,756,900	-	-	-	-	94.17% (Q20) 88.6% (Q30)
	Clean Data	1,121,333,120	168,199,968,000	-	-	-	-	-

	Number of Reads	Total length (bp)	Mapped	Mapping Ratio	Valid Pairs	Percentage
Hi-C (PE100)	1,518,592,046	151,859,204,600	1,139,158,524	75.01%	196,757,892	25.91%

* Our previously published data (Huang et al., 2019).

Table S2. Nanopore clean data length distribution.

Length	Reads Number	Total Length (bp)	Percentage (%)	Average Length (bp)
2,000~5,000	166,917	572,320,083	0.79	3,428.77
5,000~10,000	232,711	1,728,955,943	2.41	7,429.62
10,000~20,000	321,955	4,719,951,481	6.58	14,660.28
20,000~30,000	327,150	8,230,356,547	11.47	25,157.74
30,000~40,000	438,229	15,243,432,557	21.26	34,784.17
40,000~50,000	272,129	12,137,281,885	16.92	44,601.20
50,000~60,000	173,608	9,481,705,172	13.22	54,615.60
60,000~70,000	106,138	6,851,483,605	9.55	64,552.59
70,000~80,000	62,318	4,644,803,948	6.47	74,533.90
>=80,000	82,885	8,087,696,127	11.28	97,577.31

Table S3. Summary of the pecan genome assembly.

	Sequence Type	Total Number	Total Length (bp)	N50 (bp)	N90 (bp)	Longest Read (bp)	Shortest Read (bp)	Gap Length (bp)	GC Content (%)
Nanopore	contig	341	636,255,455	4,195,733	1,070,318	23,879,822	1,663	0	35.89
Hi-C	scaffold	124	636,406,555	38,784,058	25,295,717	55,745,374	1,663	263,863	35.89
	contig	564	636,142,692	2,893,887	789,098	11,551,424	112	-	35.89

Table S4. The statistics of pecan chromosome length.

Chromosome ID	**Length (bp)**	**Percentage of assembly (%)**
chr01	35,734,486	5.62
chr02	38,784,058	6.09
chr03	30,933,461	4.86
chr04	29,925,508	4.70
chr05	20,838,255	3.27
chr06	36,373,273	5.72
chr07	38,592,383	6.06
chr08	31,550,154	4.96
chr09	55,745,374	8.76
chr10	55,374,445	8.70
chr11	45,985,284	7.23
chr12	44,801,196	7.04
chr13	40,882,215	6.42
chr14	39,152,115	6.15
chr15	38,648,230	6.07
chr16	25,295,717	3.97
Total	608,616,154	95.63
un-anchored	63,793,453	4.37

Table S5. Assessment of the gene coverage rate for pecan assembly and predicted protein-coding genes by BUSCO.

	Percentage (%)	
	Assembly	**Protein-coding genes**
Complete BUSCOs	95.1 (409 genes)	93.7 (403 genes)
Complete Single-Copy BUSCOs	85.3 (367 genes)	85.6 (368 genes)
Complete Duplicated BUSCOs	9.8 (42 genes)	8.1 (35 genes)
Fragmented BUSCOs	1.2 (5genes)	4.2 (18 genes)
Missing BUSCOs	3.7 (16 genes)	2.1 (9 genes)

Table S6. Reads coverage statistics of pecan genome assembly.

Reads mapping rate (%)	95.95
Coverage of genome (%)	96.38
Coverage of genome > 4×(%)	95.25
Coverage of genome > 10× (%)	94.45
Coverage of genome > 20× (%)	93.72
Genome average sequencing depth (×)	215.49

Table S7. Repeat sequence prediction.

Type	**Repeat Size(bp)**	**% of genome**
TRF	23,381,072	3.67
RepeatMasker	87,654,647	13.77
RepeatProteinMask	76,449,465	12.01
De novo	279,886,295	43.98
Total	304,405,087	47.83

Table S8. Repeat category statistics.

	Repbase TEs		De novo		TE Proteins		Combined TEs	
	Length (bp)	**% in Genome**	**Length (bp)**	**% in Genome**	**Length (bp)**	**% in Genome**	**Length (bp)**	**% in Genome**
DNA	11,627,231	1.83	543,601	0.09	30,704,387	4.82	37,638,398	5.91
LINE	11,392,813	1.79	7,268,375	1.14	29,105,229	4.57	37,741,823	5.93
SINE	15,026	0	0	0	169,332	0.03	184,288	0.03
LTR	65,295,660	10.26	69,106,873	10.86	213,465,587	33.54	223,597,995	35.13
Other	1,510	0	0	0	0	0	1,510	0
Unknown	0	0	0	0	7,712,931	1.21	7,712,931	1.21
Total	87,654,647	13.77	76,449,465	12.01	272,008,903	42.74	288,757,673	45.37

Table S9. Statistics of repeat sequences and subfamilies of transposable elements (TEs).

Repeat category	subfamily	Size (bp)	% of repeat	Size in sub A (bp)	Size in sub B (bp)
TEs	DNA/Academ	5,469	0.0018	9,341	2,354
	DNA/CMC-Chapaev	62,165	0.0204	32,422	60,462
	DNA/CMC-Chapaev-3	4,573	0.0015	16,279	1,779
	DNA/CMC-EnSpm	13,705,125	4.5023	7,187,368	5,294,574
	DNA/CMC-Transib	37,776	0.0124	25,999	37,221
	DNA/Crypton	11,841	0.0039	8,124	4,542
	DNA/Crypton-H	30	0.0000	30	0
	DNA/Crypton-V	32,639	0.0107	20,358	14,472
	DNA/Dada	27,039	0.0089	16,247	11,406
	DNA/DNA	4,349,717	1.4289	1,878,329	1,913,326
	DNA/Ginger	224,921	0.0739	56,442	39,807
	DNA/Harbinger	157	0.0001	0	42,725
	DNA/hAT	257,672	0.0846	76,133	113,094
	DNA/hAT-Ac	6,752,533	2.2183	3,657,022	2,122,439
	DNA/hAT-Blackjack	5,768	0.0019	3,612	2,440
	DNA/hAT-Charlie	89,895	0.0295	31,249	34,502
	DNA/hAT-hAT5	932	0.0003	516	34
	DNA/hAT-hATm	23,540	0.0077	24,983	568
	DNA/hAT-hATw	44,907	0.0148	1,806	10,499
	DNA/hAT-hobo	1,073	0.0000	131	942
	DNA/hAT-Pegasus	6,548	0.0022	5,483	357
	DNA/hAT-Tag1	212,854	0.0699	103,730	139,728
	DNA/hAT-Tip100	890,171	0.2924	419,440	486,037
	DNA/hAT-Tol2	1,571	0.0005	581	1,268
	DNA/Helitron	3,984,983	1.3091	2,733,752	1,235,857
	DNA/IS3EU	40,653	0.0134	31,592	20,232
	DNA/Kolobok	4,864	0.0016	3,152	2,163
	DNA/Kolobok-Hydra	46,173	0.0152	30,274	18,947
	DNA/Kolobok-T2	21,446	0.0070	14,732	9,034
	DNA/Maverick	208,700	0.0686	149,653	43,345
	DNA/Merlin	24,885	0.0082	14,570	12,415
	DNA/MuLE-MuDR	2,461,172	0.8085	1,003,829	243,805
	DNA/MULE-MuDR	1,555,016	0.5108	776,984	806,123
	DNA/MULE-NOF	586	0.0002	605	180
	DNA/Novosib	406,496	0.1335	39,528	24,091
	DNA/P	60,621	0.0199	81,580	27,582
	DNA/PIF-Harbinger	1,371,592	0.4506	654,087	846,539
	DNA/PIF-HarbS	77	0.0000	77	0
	DNA/PIF-ISL2EU	75,979	0.0250	27,052	47,067
	DNA/PiggyBac	7,110	0.0023	2,477	3,840

DNA/Sola	176,894	0.0581	37,257	33,279
DNA/TcMar	25,474	0.0084	85,690	11,261
DNA/TcMar-Fot1	70,520	0.0232	33,800	54,152
DNA/TcMar-ISRm11	82,655	0.0272	4,336	2,761
DNA/TcMar-Pogo	56	0.0000	72	0
DNA/TcMar-Stowaway	11,676	0.0038	6,794	4,918
DNA/TcMar-Tc1	22,562	0.0074	2,531	25,253
DNA/TcMar-Tc4	42	0.0000	111	62
DNA/TcMar-Tigger	2,420,774	0.7953	1,099,620	442,736
DNA/Zator	42	0.0000	42	0
DNA/Zisupton	33,249	0.0109	3,637	2,548
LINE/Ambal	204	0.0001	61	131
LINE/CR1	1,591	0.0005	252,970	220
LINE/CRE	145	0.0000	10,563	924
LINE/Dong-R4	1,633	0.0005	306	521
LINE/DRE	3,854	0.0013	1,432	11,786
LINE/I	1,253	0.0004	415	784
LINE/Jockey	7,477	0.0025	4,970	3,341
LINE/L1	36,940,270	12.1354	17,420,214	13,392,696
LINE/L1-Tx1	87,445	0.0287	71,425	47,860
LINE/L2	608,135	0.1998	205,047	235,900
LINE/LINE	11,721	0.0039	5,016	5,624
LINE/Penelope	64,425	0.0212	84,909	23,469
LINE/Proto1	1,633	0.0005	726	301
LINE/R1	6,649	0.0022	50,145	1,863
LINE/R2	60,729	0.0200	31,170	27,900
LINE/Rex-Babar	1,415	0.0005	718	625
LINE/RTE-BovB	11,161	0.0037	6,676	2,657
LINE/RTE-RTE	59	0.0000		59
LINE/RTE-X	5,055	0.0017	1,824	1,598
LINE/Tad1	268	0.0001	3,059	35,899
LTR/Cassandra	24,252	0.0080	5425	704
LTR/Caulimoviru	1,189,232	0.3907	661,600	429,225
LTR/Caulimovirus	2,309,802	0.7588	1,308,128	1,095,708
LTR/Copia	**104,193,921**	**34.2293**	**51,853,612**	**35,515,454**
LTR/DIRS	78,269	0.0257	38,803	7,854
LTR/ERV	20,131	0.0066	858	281
LTR/ERV1	221,271	0.0727	89,544	59,614
LTR/ERV4	670	0.0002	381	7554
LTR/ERV-Foamy	61	0.0000	61	0
LTR/ERVK	118,769	0.0390	161,658	130,585
LTR/ERVL	3,191	0.0010	1,618	1,506
LTR/ERVL-MaLR	46	0.0000	46	0
LTR/Gypsy	**105,595,920**	**34.6899**	**46,854,456**	**44,036,255**

	LTR/LTR	23,178,086	7.6144	9,812,895	6,951,532
	LTR/Ngaro	7,059	0.0023	5,619	2,028
	LTR/Pao	166,864	0.0548	25,964	54,808
	SINE/Alu	72	0.0000	3,380	0
	SINE/B2	54	0.0000	54	0
	SINE/B4	4,668	0.0015	3,080	2,061
	SINE/ID	2,600	0.0009	1,719	1,303
	SINE/SINE	169,564	0.0557	16,485	88,089
	SINE/tRNA-7SL	577	0.0002	503	214
	SINE/tRNA-C	4,672	0.0015	2,839	2,508
	SINE/tRNA-Core	655	0.0002	450	0
	SINE/tRNA-Deu-L2	275	0.0001	263	57
	SINE/tRNA-L2	132	0.0000	132	0
	SINE/tRNA-RTE	1,135	0.0004	607	280
	SINE/U	136	0.0000	132	68
Satellite	Satellite	246,671	0.0810	287,088	139,547
Other	Composite	252	0.0001	126	126
Other	DNA_virus	1,258	0.0004	640	637
Simple	Simple_repeat	7,630,721	2.5068	4,184,870	4,107,352
Unknown	Unknown	7,712,931	2.5338	6,531,303	7,424,836

Table S10. Statistics of gene structure prediction in pecan genome assembly.

	Gene set	Number of genes	CDS+intron length (bp)	CDS length (bp)	exon length (bp)	intron (bp)	Exons per gene
Homolog	*Arabidopsis thaliana*	36,944	8,410.66	1,262.22	206.68	1,399.73	6.11
	Cucumis sativus	22,467	6,573.42	1,194.11	231.52	1,293.82	5.16
	Glycine max	21,176	7,865.64	1,300.34	226.89	1,387.66	5.73
	Prunus persica	30,693	12,085.40	1,220.08	240.04	2,661.25	5.08
	Citrullus lanatus	41,721	9,415.54	1,246.60	229.94	1,847.63	5.42
	Eucalyptus grandis	43,556	13,258.85	1,144.50	235.99	3,146.84	4.85
	Malus domestica	18,803	6,188.82	1,003.87	207.23	1,348.76	4.84
	Populus trichocarpa	32,661	10,064.20	1,313.59	237.69	1,933.22	5.53
	Vitis vinifera	24,743	13,946.90	1,328.17	236.56	2,734.52	5.61
De novo	Augustus	49,803	4333.06	1276.33	233.22	683.43	5.47
Transcript	Trinity	154,647			/		
Total		33,472	5482.63	1460.04	220.81	716.76	6.61

Table S11. Comparison of gene structure among the close relatives.

Species	Number	Average transcript length (bp)	Average CDS length (bp)	Average intron length (bp)	Average exon length (bp)	Average exons per gene
Carya illinoinensis	33,472	5,482.63	1460.04	716.76	220.81	6.61
Arabidopsis thaliana	27,465	1,890.03	1223.07	161.1	237.88	5.14
Eucalyptus grandis	36,625	2,568.35	1136.77	409.07	252.64	4.5
Glycine max	46,949	3,616.50	1289.18	537.56	241.89	5.33
Populus trichocarpa	31,612	3,314.73	1392.68	421.85	250.65	5.56
Vitis vinifera	25,726	5,619.01	1348.43	1008.82	257.66	5.23

Table S12. Statistics of function annotation of protein-coding genes in pecan genome assembly (version 2.0).

Database	**Number**	**Percentage (%)**
Swissprot	24,359	72.77
KEGG	25,885	77.33
TrEMBL	30,885	92.27
Interpro	26,047	77.82
GO	18,205	53.39
all annotated	31,247	93.35
Total	33,472	100

Table S13. Comparison on genomic features of all assemblies of pecan.

	GigaScience[a]	Cil_V. 2.0	Nature Communications[b]			
Genomic features	**Pawnee**	**Pawnee**	**Oaxaca**	**Lakota**	**Elliott**	**Pawnee**
Total length of scaffolds (Mb)	651.31	636.41	649.96	668.99	656.69	674.27
Number of scaffolds & contigs	3860/17542	125 & 433	298 & 552	261 & 499	431 & 829	16 & 34
Longest scaffold (Mb)	4.92	55.75	58.44	57	56.14	58.06
N50 of contig length	77.2Kb	3.04Mb	4.4Mb	3.7Mb	4.4Mb	26.5Mb
Number of predicted protein-coding genes	31,075	33,472	31,911	33,280	31,042	32,267
Pseudochromosomes	/	16	16	16	16	16
Anchored sequence to pseudochromosome (Mb)	/	608.6	637.0	642.9	627.1	674.3
Genome in chromosomes (%)	/	96%	98%	96.10%	95.50%	100%
Average number of exons per gene	5.0	6.6	5.4	5.5	5.5	5.5
Percentage of repeat sequences (%)	50.43	47.83	46.5	33.8	32.3	49.7

Note: a, Huang et al., 2019. b, Lovell et al, 2021.

Table S14. Statistics of key genes related to non-structural polyphenol metabolism and oil accumulation in pecan genome versions 2.0 and 1.0.

		Cil_V. 2.0				Vertion
		subA	subB	scaffold	Total	1.0
Polyphenol metabolism	**4CL**	16	12	1	**29**	**34**
	C4H	2	1	0	**3**	**6**
	LAC15	30	11	1	**42**	**49**
	MAY123	8	7	0	**15**	**14**
	PAL	1	2	0	**3**	**8**
	WDR	0	1	0	**1**	**2**
	CHS	2	1	0	**3**	**4**
	CHI	1	1	0	2	2
	F3H	1	0	0	1	1
	F3'H	4	1	0	5	5
	DFR	3	0	0	3	3
	LDOX	1	0	0	1	1
	ANR	1	1	0	**2**	**3**
	GSTF	0	1	0	1	1
	MATE	2	0	0	2	2
	HATPase	0	1	0	**1**	**2**
	Sum	**72**	**40**	**2**	**114**	**137**
Oil accumulation	**ABCAT**	1	1	0	2	2
	ABI3	1	0	0	1	1
	ABI4	1	0	0	1	1
	ACP4	1	1	0	2	2
	alpha-CT	1	2	0	3	3
	alpha-PDH	1	1	0	**2**	**1**
	BC	1	0	0	1	1
	BCCP1	1	1	0	**2**	**4**
	beta-CT	0	0	0	**0**	**1**
	beta-PDH	1	0	0	**1**	**2**
	DHLAT/EMB3003(E2)	0	1	0	1	1
	DHLAT/LTA2(E2)	1	1	0	2	2
	ER/ENR1(MOD1)	0	1	0	**1**	**2**
	FATA	1	0	0	**1**	**2**
	FATB	1	1	0	2	2
	HACPS	1	0	0	1	1
	HAD	1	0	0	**1**	**2**

KAR	1	1	0	**2**	**3**
KASI	1	1	0	**2**	**3**
KASII	1	1	0	**2**	**3**
KASIII	1	1	0	2	2
LACS8	1	0	0	1	1
LACS9	1	0	0	**1**	**2**
LPD2(E3)	1	0	0	1	1
LS	1	1	0	2	2
LT	0	1	0	1	1
MCMT	1	1	0	2	2
PII	1	0	0	**1**	**2**
SAD/DES5	0	0	1	1	1
SAD/DES6	0	1	0	**1**	**2**
SAD/FAB2	1	0	0	**1**	**4**
TGD1	1	0	0	1	1
WRI1	3	1	0	**4**	**5**
WRI3	2	2	0	4	4
Sum	**31**	**21**	**1**	**53**	**69**

Table S15. Predicted non-coding RNAs in pecan genome assembly (version 2.0).

Type		Copy number	Average length (bp)	Total length (bp)	% of genome
miRNA		121	126.8	15,343	0.0024
tRNA		565	74.86	42,295	0.0066
rRNA	rRNA	414	180.32	74,651	0.0117
	18S	82	505.44	41,446	0.0065
	28S	69	107.87	7,443	0.0012
	5.8S	21	122.29	2,568	0.0004
	5S	242	95.84	23,194	0.0036
snRNA	snRNA	1,318	109.55	144,382	0.0227
	CD-box	1,118	105.23	117,642	0.0185
	HACA-box	52	124.08	6,452	0.001
	splicing	147	137.13	20,158	0.0032

Table S16. Statistics of synteny and genes in Cil_V. 2.0 and gap-free 'Pawnee' assembly by Lovell et al. (2021).

Chromosome Pair	Sub-genome	Chromosome[a]	Gene Number[a]	Chromosome[b]	Gene Number[b]	Genes in Block[b]	Block Number	Average gene number/Block
PAIR1	sub A	CHR09	3,011	chr1	3,080	2,001	16	125.06
	sub B	CHR01	1,984	chr2	1,995	1,422	3	474.00
PAIR2	sub A	CHR10	2,967	chr3	2,845	1,967	10	196.70
	sub B	CHR02	1,911	chr4	1,984	1,292	9	143.56
PAIR3	sub A	CHR11	2,561	chr5	2,644	1,790	7	255.71
	sub B	CHR03	1,626	chr6	1,772	1,147	10	114.70
PAIR7	sub A	CHR15	2,305	chr7	2,377	1,781	3	593.67
	sub B	CHR07	2,003	chr8	1,830	1,255	8	156.88
PARI6	sub A	CHR14	2,173	chr9	2,253	1,569	9	174.33
	sub B	CHR06	1,836	chr10	1,677	1,126	8	140.75
PAIR5	sub A	CHR13	2,106	chr11	2,110	1,353	10	135.30
	sub B	CHR05	1,385	chr12	1,430	955	7	136.43
PAIR8	sub A	CHR16	1,431	chr13	1,802	1,016	15	67.73
	sub B	CHR08	1,493	chr14	1,374	939	6	156.50
PAIR4	sub A	CHR12	1,944	chr15	1,864	1,097	9	121.89
	sub B	CHR04	1,368	chr16	1,197	794	6	132.33
Sum		16	32,104	16	32,234	21,504	136	158.12

Note: [a], Chromosome number and genes in Cil_V. 2.0 assembly; [b], Chromosome number and genes in the gap-free 'Pawnee' assembly by Lovell et al. (2021).

Table S17. Statistics of sub-genomes features.

Chromosome Pair	Sub-genome	Chromosome	Chromosome length	Gene number	TE	Identity	Ka/Ks
PAIR1	subA	CHR09	55,745,374	3,011	0.446	78.64	0.2261
	subB	CHR01	35,734,486	1,984	0.449	78.33	0.2298
PAIR2	subA	CHR10	55,374,445	2,967	0.451	77.72	0.2499
	subB	CHR02	38,784,058	1,911	0.474	75.99	0.2507
PAIR3	subA	CHR11	45,985,284	2,561	0.452	77.79	0.2419
	subB	CHR03	30,933,461	1,626	0.468	76.95	0.2524
PAIR4	subA	CHR12	44,801,196	1,944	0.511	78.73	0.2456
	subB	CHR04	29,925,508	1,368	0.502	77.17	0.2531
PAIR5	subA	CHR13	40,882,215	2,106	0.469	77.48	0.2602
	subB	CHR05	20,838,255	1,385	0.413	76.17	0.2600
PAIR6	subA	CHR14	39,152,115	2,173	0.44	79.06	0.2396
	subB	CHR06	36,373,273	1,836	0.47	77.20	0.2551
PAIR7	subA	CHR15	38,648,230	2,305	0.409	76.99	0.2475
	subB	CHR07	38,592,383	2,003	0.467	75.24	0.2769
PAIR8	subA	CHR16	25,295,717	1,431	0.436	78.84	0.2540
	subB	CHR08	31,550,154	1,493	0.489	77.40	0.2627

Table S18. Statistics of transcription factor families among pecan, walnut and bayberry.

TF_family	*Cil*_sub A	*Cil*_sub B	*Cil*_Scaffold	*Cil* sum	*Jre*_sub A	*Jre*_sub B	*Jre* sum	*Mru*	*Ath*	Included_domains	P-value(chi-square test)
ABI3VP1	23	17	2	42	20	17	37	60	89	B3	0.068
AP2-EREBP	105	85	3	193	109	93	202	128	145	AP2	1
ARF	20	10	1	31	17	16	33	18	23	Auxin_resp	1
ARR-B	7	4	1	12	7	5	12	9	13	G2-like,Myb_DNA-binding,Response_reg	0.9827
Alfin-like	11	12	2	25	10	12	22	27	19	Alfin-like	0.872
BBR/BPC	5	6	0	11	6	5	11	6	8	GAGA_bind	1
BES1	5	4	1	10	4	5	9	6	8	DUF822	1
BSD	5	7	1	13	6	6	12	10	12	BSD	1
C2C2-CO-like	9	4	0	13	10	6	16	7	17	CCT,zf-B_box	0.8128
C2C2-Dof	22	19	3	44	26	22	48	27	36	zf-Dof	1
C2C2-GATA	18	13	0	31	18	15	33	30	30	GATA	0.6565
C2C2-YABBY	5	4	0	9	6	5	11	9	7	YABBY	0.8887
C2H2	39	38	4	81	46	46	92	55	65	zf-C2H2	0.8879
C3H	37	28	6	71	46	32	78	49	63	zf-CCCH	0.8617
CAMTA	3	2	0	5	5	4	9	7	5	CG-1,IQ	0.5713
CCAAT	16	19	0	35	12	14	26	22	43	CBFB_NFYA,CBFD_NFYB_HMF,CCAAT-Dr1,NF-YB,NF-YC	1
CPP	4	3	0	7	6	4	10	5	10	TCR	0.7437
CSD	2	1	0	3	3	1	4	4	3	CSD	0.815
DBP	0	0	0	0	1	0	1	1	1	DNC,PP2C	0.9268
E2F-DP	7	3	1	11	7	4	11	7	8	E2F_TDP	1
EIL	3	2	1	6	4	2	6	5	6	EIN3	0.9774

FAR1	154	107	3	264	22	18	40	26	18	FAR1	**1.39E-09**
FHA	10	8	1	19	12	9	21	14	16	FHA	0.9653
G2-like	28	23	2	53	33	26	59	36	43	G2-like	0.8481
GRAS	41	29	3	73	50	27	77	49	34	GRAS	0.8219
GRF	5	8	0	13	8	7	15	10	9	QLQ,WRC	1
GeBP	3	4	0	7	4	4	8	7	22	DUF573	0.2378
HB	9	7	0	16	9	6	15	8	8	Homeobox,KNOX1,KNOX2	0.7427
HRT	1	0	0	1	1	0	1	0	2	HRT	1
HSF	16	13	2	31	19	15	34	18	24	HSF_DNA-bind	1
LFY	0	1	0	1	0	1	1	1	1	FLO_LFY	1
LIM	12	5	0	17	9	6	15	11	12	LIM	1
LOB	26	24	2	52	29	22	51	34	43	DUF260	1
MADS	38	29	1	68	41	25	66	65	108	SRF-TF	0.2382
MYB	192	157	14	363	205	179	384	247	274	Myb_DNA-binding	1
MYB-related	159	130	12	301	167	148	315	205	227	Myb_DNA-binding	1
NAC	72	51	4	127	70	56	126	75	114	NAM	1
NOZZLE	2	2	0	4	1	1	2	2	3	NOZZLE_Angio	0.9087
OFP	15	7	3	25	17	9	26	13	19	Ovate	1
PBF-2-like	2	1	0	3	1	0	1	2	3	Whirly	1
PLATZ	12	9	1	22	7	4	11	16	12	PLATZ	0.5853
RWP-RK	7	4	1	12	6	5	11	9	15	RWP-RK	0.8289
S1Fa-like	2	0	0	2	2	1	3	2	3	S1FA	0.8413
SAP	1	0	0	1	1	0	1	2	1	STER_AP	1
SBP	13	16	2	31	16	12	28	16	17	SBP	0.6644
SRS	6	6	0	12	5	6	11	5	11	DUF702	1
Sigma70-like	1	6	0	7	1	6	7	5	6	Sigma70_r2,Sigma70_r3,Sigma70_r4	1

TAZ	3	3	0	6	3	4	7	4	5	zf-TAZ	1
TCP	12	19	2	33	18	19	37	17	24	TCP	1
TIG	4	5	0	9	4	5	9	6	4	TIG	0.9576
TUB	5	5	2	12	6	6	12	10	11	Tub	0.9827
Tify	8	8	1	17	11	12	23	13	15	tify	0.7606
Trihelix	24	20	1	45	25	23	48	33	30	trihelix	1
ULT	2	1	0	3	4	1	5	2	2	ULT	1
VARL	1	2	0	3	1	2	3	1	3	VARL	1
VOZ	5	2	1	8	3	3	6	3	5	VOZ	0.9053
WRKY	45	39	3	87	51	40	91	57	73	WRKY	0.9704
bHLH	91	74	6	171	97	85	182	111	141	HLH	0.9583
bZIP	19	17	0	36	21	19	40	24	22	bZIP_1,bZIP_2,bZIP_Maf	1
mTERF	19	27	2	48	22	30	52	33	36	mTERF	1
zf-HD	9	6	2	17	11	7	18	11	17	ZF-HD_dimer	1

Note: *Jre*, *Juglans regia* (walnut); *Mru*, *Myrica rubra* (bayberry); *Cil*, *Carya illinoisnensis* (pecan).

Table S19. Information of samples for whole genome re-sequencing.

Sample ID	Plant ID	Scab resistance*	Source**	Sampling location	Population***
Barton	ML11	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Melrose	HL35	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Osage	HL16	1	Foreign	China: Paiyashan Farm, Hunan	R
Peruque	HL38	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
PoSey	HL30	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Pvilop	HL32	1	Foreign	China: Paiyashan Farm, Hunan	R
Surprize	HL25	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Woodroof	ML38	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Owens-1	ML32	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Yalin13	YL13	1	Domestic	China: Jing'an, Jiangxi	R
Yalin30	YL30	1	Domestic	China: Jing'an, Jiangxi	R
Farley	ML20	2	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Mandan	ML25	2.3	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Caddo	ML13	3	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Dependable	ML1	3	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Forkert	ML21	3	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Kiowa	ML23	3	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Oconee	HL17	3	Foreign	China: Paiyashan Farm, Hunan	S
Stuart	ML5	3	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Cheyenne	HL12	4	Foreign	China: Paiyashan Farm, Hunan	S
Hirschi/Steuck	HL29	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Mahan	ZL57	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Moore	ML29	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S

Schley	ML6	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Shawnee	ZL59	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Shoshoni	ML34	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Tejas	ZL60	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Woodard	ML37	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
VC1-68	VC1-68	4	Foreign	USA: NCGR Provenance Orchards, Somerville	S
Excell	ML3	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Excell	ML3-2	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Excell	ML3-3	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Excell	ML3-4	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Excell	ML3-5	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Excell	ML3-6	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Excell	ML3-7	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Excell	ML3-8	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Excell	ML3-9	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Excell	ML3-11	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Excell	ML3-12	1	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Elliott	ML7	1.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Elliott	ML7-4	1.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Elliott	ML7-6	1.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Elliott	ML7-7	1.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Elliott	ML7-8	1.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Elliott	ML7-9	1.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Elliott	ML7-10	1.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	R
Sumner	ML8	2.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Sumner	ML8-2	2.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	S

Sumner	ML8-3	2.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Sumner	ML8-4	2.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Sumner	ML8-6	2.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Sumner	ML8-7	2.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Sumner	ML8-8	2.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Sumner	ML8-9	2.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Sumner	ML8-10	2.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Sumner	ML8-11	2.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Sumner	ML8-12	2.67	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Desirable	ML9	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Desirable	ML9-2	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Desirable	ML9-3	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Desirable	ML9-4	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Desirable	ML9-5	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Pawnee	ZL49-1	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Pawnee	ZL49-2	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Pawnee	ZL49-3	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Pawnee	ZL49-5	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Pawnee	ZL49-6	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Pawnee	ZL49-7	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Pawnee	ZL49-8	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Pawnee	ZL49-9	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Pawnee	ZL49-10	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Pawnee	ZL49-11	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Pawnee	ZL49-13	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Pawnee	ZL49-14	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S

Pawnee	ZL49-16	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Pawnee	ZL49-17	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Sioux	HL8	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Sioux	HL42	4	Foreign	China: Paiyashan Farm, Hunan	S
Western Schley	ZL58-3	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Western Schley	ZL58-4	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Western Schley	ZL58	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Western Schley	ZL58-11	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Western Schley	ZL58-12	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Western Schley	ZL58-13	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S
Western Schley	ZL58-15	4	Foreign	China: Zhejiang A&F University campus, Hangzhou	S

Note: *, the grade of pecan scab resistance were assessed by integrating the record from XXX. **, "Foreign" indicates the original germplasms were introduced from outside China and "Domestic" means the accessions were collected from seed-germinated seedling in China. ***, the populations were defined based on the pecan scab resistance grade: R, resistance population with the grade <=2; S, susceptiable population with the grade >2.

Table S20. Statistics of the resequencing data and the SNPs for each accessions.

Sample ID	Plant ID	Raw_reads (M)	Raw_bases (Gb)	Clean_reads (M)	Clean_bases (Gb)	Clean_ratio(%)	Clean_coverage (X)	Q20(%)	Q30(%)	GC(%)
Barton	ML11	190.80	19.08	190.12	19.01	99.64	27.50	98.47	94.91	37.09
Melrose	HL35	202.54	20.25	201.85	20.19	99.66	29.21	97.09	91.73	37.49
Osage	HL16	237.93	23.79	235.60	23.56	99.02	34.08	98.80	94.91	37.66
Peruque	HL38	169.97	17.00	169.41	16.94	99.67	24.51	97.19	92.05	37.17
PoSey	HL30	185.78	18.58	185.12	18.51	99.64	26.78	97.03	91.58	37.61
Pvilop	HL32	188.53	18.85	186.61	18.66	98.98	26.99	98.56	94.99	36.83
Surprize	HL25	192.61	19.26	192.09	19.21	99.73	27.79	96.97	91.36	37.39
Woodroof	ML38	203.03	20.30	202.10	20.21	99.54	29.24	98.45	94.62	36.95
Owens-1	ML32	173.03	17.30	172.27	17.23	99.56	24.92	98.52	94.90	37.58
Yalin13	YL13	227.62	22.76	226.89	22.69	99.68	32.82	97.88	93.66	37.41
Yalin30	YL30	179.08	17.91	178.60	17.86	99.73	25.84	97.81	93.28	37.14
Farley	ML20	168.36	16.84	167.87	16.79	99.71	24.29	98.44	94.53	37.05
Mandan	ML25	172.26	17.23	171.73	17.17	99.69	24.84	98.49	94.77	37.25
Caddo	ML13	248.04	24.80	246.72	24.67	99.47	35.69	98.73	96.10	38.69
Dependable	ML1	222.27	22.23	221.56	22.16	99.68	32.06	98.70	95.70	37.03
Forkert	ML21	183.42	18.34	182.89	18.29	99.71	26.46	98.53	94.91	37.01
Kiowa	ML23	185.02	18.50	184.47	18.45	99.70	26.69	98.34	94.07	37.03
Oconee	HL17	193.78	19.38	191.95	19.19	99.05	27.76	98.51	94.74	37.01
Stuart	ML5	159.09	15.91	158.64	15.86	99.72	22.94	98.62	95.60	37.26
Cheyenne	HL12	175.20	17.52	173.82	17.38	99.21	25.14	98.42	94.61	37.15
Hirschi/Steuck	HL29	150.11	15.01	149.70	14.97	99.73	21.66	96.98	91.40	37.53
Mahan	ZL57	159.84	15.98	159.42	15.94	99.74	23.06	98.09	94.31	37.96

Moore	ML29	185.77	18.58	184.87	18.49	99.52	26.75	98.52	94.72	37.01
Schley	ML6	198.88	19.89	198.20	19.82	99.66	28.67	98.61	95.88	37.22
Shawnee	ZL59	264.27	26.43	263.20	26.32	99.60	38.07	98.37	94.44	37.64
Shoshoni	ML34	170.75	17.07	169.97	17.00	99.55	24.59	98.54	95.18	37.39
Tejas	ZL60	148.16	14.82	147.24	14.72	99.38	21.29	98.02	93.71	44.20
Woodard	ML37	169.32	16.93	168.55	16.85	99.55	24.38	98.59	95.16	37.23
VC1-68	VC1-68	226.99	22.70	225.88	22.59	99.51	32.68	98.26	94.16	40.06
Excell	ML3	199.76	19.98	199.22	19.92	99.73	28.82	98.65	95.48	36.72
Excell	ML3-2	151.03	15.10	149.99	15.00	99.32	21.70	98.53	94.95	36.02
Excell	ML3-3	179.32	17.93	178.84	17.88	99.73	25.87	98.17	94.75	37.02
Excell	ML3-4	234.32	23.43	233.75	23.37	99.75	33.81	98.62	95.49	36.61
Excell	ML3-5	190.18	19.02	189.92	18.99	99.87	27.47	98.57	94.93	37.38
Excell	ML3-6	152.70	15.27	152.56	15.26	99.90	22.07	98.56	94.72	36.94
Excell	ML3-7	166.64	16.66	166.47	16.65	99.90	24.09	98.50	94.55	37.25
Excell	ML3-8	179.47	17.95	179.29	17.93	99.90	25.94	98.59	94.97	37.21
Excell	ML3-9	192.67	19.27	192.36	19.24	99.84	27.83	98.65	95.33	37.00
Excell	ML3-11	214.80	21.48	214.52	21.45	99.87	31.03	98.60	95.05	37.03
Excell	ML3-12	195.12	19.51	194.80	19.48	99.84	28.18	98.64	95.41	36.91
Elliott	ML7	202.68	20.27	202.05	20.21	99.69	29.24	98.71	95.96	37.32
Elliott	ML7-4	190.73	19.07	190.44	19.04	99.85	27.54	98.46	94.83	37.14
Elliott	ML7-6	151.99	15.20	151.79	15.18	99.87	21.96	98.39	94.53	37.45
Elliott	ML7-7	193.29	19.33	192.97	19.30	99.83	27.92	98.41	94.53	36.95
Elliott	ML7-8	197.55	19.76	197.27	19.73	99.85	28.54	98.47	94.75	36.75
Elliott	ML7-9	211.40	21.14	211.00	21.10	99.81	30.52	98.54	95.20	36.90
Elliott	ML7-10	174.58	17.46	174.26	17.43	99.82	25.21	98.55	95.25	36.92
Sumner	ML8	200.54	20.05	200.03	20.00	99.75	28.93	98.64	95.31	37.62

Sumner	ML8-2	162.36	16.24	161.90	16.19	99.72	23.42	97.62	92.08	37.27
Sumner	ML8-3	142.77	14.28	141.83	14.18	99.34	20.51	97.59	92.69	43.50
Sumner	ML8-4	173.06	17.31	172.38	17.24	99.61	24.94	98.38	94.16	36.47
Sumner	ML8-6	249.19	24.92	247.15	24.71	99.15	35.75	97.52	92.34	37.80
Sumner	ML8-7	165.17	16.52	164.73	16.47	99.74	23.83	97.67	92.30	37.12
Sumner	ML8-8	165.01	16.50	164.56	16.46	99.73	23.81	97.73	92.75	38.47
Sumner	ML8-9	236.38	23.64	235.56	23.56	99.65	34.08	96.86	90.98	37.28
Sumner	ML8-10	234.48	23.45	233.70	23.37	99.67	33.81	96.88	91.03	37.49
Sumner	ML8-11	153.29	15.33	152.79	15.28	99.67	22.10	96.76	90.70	37.79
Sumner	ML8-12	376.17	37.62	374.53	37.45	99.56	54.17	98.50	94.89	38.88
Desirable	ML9	185.56	18.56	185.06	18.51	99.73	26.78	98.61	95.07	37.44
Desirable	ML9-2	155.46	15.55	155.05	15.51	99.74	22.44	97.65	92.21	37.79
Desirable	ML9-3	126.63	12.66	126.31	12.63	99.75	18.27	97.54	91.85	37.36
Desirable	ML9-4	127.31	12.73	127.02	12.70	99.77	18.37	97.66	92.34	38.05
Desirable	ML9-5	202.03	20.20	200.85	20.09	99.42	29.06	98.46	95.08	40.67
Pawnee	ZL49-1	194.92	19.49	193.96	19.40	99.51	28.06	98.33	94.39	37.72
Pawnee	ZL49-2	221.20	22.12	220.25	22.02	99.57	31.85	98.42	94.92	37.59
Pawnee	ZL49-3	291.60	29.16	290.82	29.08	99.73	42.07	98.19	94.59	38.83
Pawnee	ZL49-5	192.43	19.24	191.92	19.19	99.74	27.76	98.20	94.43	37.34
Pawnee	ZL49-6	152.36	15.24	151.95	15.19	99.73	21.97	98.25	94.69	37.43
Pawnee	ZL49-7	159.84	15.98	159.37	15.94	99.71	23.06	98.09	94.09	37.42
Pawnee	ZL49-8	134.35	13.43	134.01	13.40	99.75	19.38	98.15	94.38	37.80
Pawnee	ZL49-9	172.45	17.24	171.95	17.19	99.71	24.87	98.16	94.30	37.80
Pawnee	ZL49-10	191.63	19.16	190.99	19.10	99.67	27.63	98.15	94.25	37.98
Pawnee	ZL49-11	165.03	16.50	164.61	16.46	99.74	23.81	98.22	94.48	37.62
Pawnee	ZL49-13	159.89	15.99	159.49	15.95	99.75	23.07	98.09	94.21	37.75

Pawnee	ZL49-14	194.60	19.46	194.00	19.40	99.69	28.06	98.07	94.12	37.36
Pawnee	ZL49-16	188.31	18.83	187.66	18.77	99.65	27.15	98.13	94.53	37.27
Pawnee	ZL49-17	190.62	19.06	190.03	19.00	99.69	27.49	98.11	94.30	37.55
Sioux	HL8	224.19	22.42	223.53	22.35	99.70	32.33	96.88	91.08	37.39
Sioux	HL42	262.74	26.27	260.80	26.08	99.26	37.73	98.58	95.39	37.23
Western Schley	ZL58	185.75	18.58	185.17	18.52	99.69	26.79	98.34	94.75	38.12
Western Schley	ZL58-3	229.59	22.96	228.81	22.88	99.65	33.10	97.79	92.62	37.24
Western Schley	ZL58-4	177.26	17.73	176.75	17.67	99.70	25.56	97.75	92.34	36.83
Western Schley	ZL58-11	191.00	19.10	190.52	19.05	99.74	27.56	96.31	87.93	39.03
Western Schley	ZL58-12	139.21	13.92	138.60	13.86	99.55	20.05	96.16	87.50	37.10
Western Schley	ZL58-13	171.53	17.15	170.76	17.07	99.53	24.69	96.96	91.19	36.83
Western Schley	ZL58-15	146.53	14.65	146.04	14.60	99.66	21.12	96.03	88.07	36.66
Total		16303.12	1630.31	16244.28	1624.41	8569.52	-	-	-	-
Average		189.57	18.96	188.89	18.89	99.65	27.32	98.08	93.85	37.59

Table S21. Parents information and estimate ancestral sources and types in subgenomes for each cultivar.

Accession	Subgenome A	Subgenome B	Parents (Female X Male)
Barton	K2	K1	Moore X Success
Caddo	K1,K4,K3,K2	K3,K1	Brooks X Alley
Cheyenne	K1	K1	Clark X Odom
Dependable	K4,K2	K4,K1	Jewett X Success
Desirable	K4,K3	K3,K4	Success X Jewett
Elliott	K4,K2	K4,K1	Not known (Seedling)
Excell	K1,K2	K4,K1	Not known (Seedling)
Farley	K1,K2	K3,K4,K1	Not known (Seedling)
Forkert	K1,K3,K2	K3,K1	Success X Schley
Hirschi/Steuck	K1,K4,K3,K2	K3,K2,K4,K1	Not known (Seedling)
Kiowa	K1,K3,K2	K3,K1	Mahan X Desirable
Mahan	K3	K3	Not known (Seedling)
Mandan	K1,K3,K2	K2,K1	BW-1 X Osage
Melrose	K1,K3	K3,K2,K1	Not known (Seedling)
Moore	K1	K1	Not known (Seedling)
Oconee	K3,K2	K3,K1	Schley X Barton
Osage	K1	K3,K4,K1	Major X Evers
Owens-1	K1,K4,K3,K2	K3,K2,K4,K1	Not known (Chance seedling)
Pawnee	K3	K3	Mohawk X Starking Hardy Giant
Peruque	K2	K1	Not known (Seedling)
PoSey	K3	K3	Not known (Seedling)
Pvilop	K1,K4,K3	K3,K4,K1	No record
Schley	K4	K4	Not known (Seedling)
Shawnee	K4	K4	Schley X Barton
Shoshoni	K1	K1	Odom X Evers
Sioux	K1,K3	K3,K1	Schley X Carmichael
Stuart	K3	K3	Not known (Seedling)
Sumner	K4,K3	K3,K4	Not known (Seedling)
Surprize	K3	K2	Not known (Chance seedling)
Tejas	K4	K4	Mahan X Risien 1
VC1-68	K1,K3	K3,K1	Not known (Seedling)
Western Schley	K1,K3	K3,K2,K4,K1	Not known (Seedling)
Woodard	K1,K2	K3,K4,K1	Not known (Seedling)
Woodroof	K1,K4,K3	K3,K4,K1	Not known (Seedling)
Yalin13	K1,K4,K3,K2	K3,K2,K4,K1	Not known (Seedling)
Yalin30	K1	K3,K4,K1	Not known (Seedling)

Note: Pink=K1, Oringe=K2, lake blue=K3, sky blue=K.

Table S22. Nuecleotide diversity (π) and pairwise population differentiation level (*Fst*) between the pecan scab resistance and susceptible populations (100 Kb per window).

Table S23. Information of candidate genes in the selected regions (Top 5% of *Fst* and π ratio).

Table S24. Phenotypes of two selected regions including the protein-encoding genes of chitinases (*CHIs*) and ionotropic glutamate receptors (*GRIPs*).

(Tables S22-S24 are shown by separate files)

Table S25. Primers information for real-time qPCR.

Gene ID	**Gene name**	**Forward primer**	**Reverse primer**	**Tm**	**Product length (bp)**
Cil_09G_00199V2	*CHI-1*	CTAATAATGTCTCGGTGTCTG	CATCTATCTCCTCTATGTAGCA		243
Cil_09G_00200V2	*CHI-2*	CAACGATGTCTCAGTGTCT	CAGCAATCTCACGCATAGA		182
Cil_15G_00015V2	*GLR-1*	CCACAGTTACAGTTCCAAGA	GCCTGACTTACAACACTACT		264
Cil_15G_00016V2	*GLR-2*	TGATGCGAAGTGAATATGTG	TCAAGTGAATGAGCAAGAAG		184
Cil_15G_00017V2	*GLR-3*	TTCCAAGAGATTCGCCAAT	AAGCAGAGCAAGCAAGAA		209
Cil_15G_00018V2	*GLR-4*	GAGTTACACTGCAAGTCTGA	ATTCTTCTGCCGAGTTGAG		191
Cil_15G_00019V2	*GLR-5*	AGTCTGACCTCAATCCTTAC	CATATCCTTCTTCCGAGTTG		181
Cil_03G_00295V2	*MAPKKK3*	TCAACGAGGACACATACAAG	TCTCCGATGAAGCCGATT		208
Cil_12G_00571V2	*FAR-1*	CCTGGCGGAGATTGATAC	GTGGACACTTAGACAGAGAA		233
Cil_03G_00293V2	*FAR-2*	AGAAGAGGAGAGCATAGACT	AAGCAAGGCATACCGTAAT		249
Internal control*	*18S rRNA*	ACATCTTACCACGATACATAAC	AACTTGCGTTCAAAGACTC		134

Note: *, 18S rRNA (NCBI accession no. AF174619.1) were used as the interal control according to Mattison et al. (2017).