The Tung Tree (*Vernicia Fordii*) Genome Provides A Resource for Understanding Genome Evolution and Oil Improvement

Lin Zhang[1,2,3,#,*,a], Meilan Liu[1,2,#,b], Hongxu Long[1,2,#,c], Wei Dong[4,#,d], Asher Pasha[3,e], Eddi Esteban[3,f], Wenying Li[1,2,g], Xiaoming Yang[5,h], Ze Li[1,i], Aixia Song[4,j], Duo Ran[1,2,k], Guang Zhao[1,2,l], Yanling Zeng[1,2,m], Hao Chen[1,2,n], Ming Zou[6,o], Jingjing Li[6,p], Fan Liang[6,q], Meili Xie[6,7,r], Jiang Hu[6,s], Depeng Wang[6,t], Heping Cao[8,*,u], Nicholas J. Provart[3,*,v], Liangsheng Zhang[4,*,w], Xiaofeng Tan[1,2,*,x]

[1]*Key Laboratory of Cultivation and Protection for Non-Wood Forest Trees, Ministry of Education, Central South University of Forestry and Technology, Changsha 410004, China*

[2]*Key Lab of Non-wood Forest Products of State Forestry Administration, College of Forestry, Central South University of Forestry and Technology, Changsha 410004, China*

[3]*Department of Cell and Systems Biology / Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario M5S 3B2, Canada*

[4]*State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou 350002, China*

[5]*College of Forestry, Nanjing Forestry University, Nanjing 210037, China*

[6]*Nextomics Biosciences Co., Wuhan 430073, China*

[7]*Oil Crop Research Institute, Chinese Academy of Agricultural Sciences, Wuhan 430062, China*

[8]*U.S. Department of Agriculture, Agricultural Research Service, Southern Regional Research Center, New Orleans, LA 70122, USA*

# Equal contribution.

* Corresponding authors.

E-mail: t20071128@csuft.edu.cn (Zhang L), heping.Cao@ars.usda.gov (Cao H), nicholas.provart@utoronto.ca (Provart NJ), zls@fafu.edu.cn (Zhang L), t19781103@csuft.edu.cn (Tan X).

Running title: *Zhang L et al / Tung Tree (Vernicia fordii) Genome*

[a]ORCID: 0000-0002-7412-5784.

[b]ORCID: 0000-0003-3837-2398.

[c]ORCID: 0000-0002-6145-0385.

[d]ORCID: 0000-0003-2201-7164.

[e]ORCID: 0000-0002-9315-0520.

[f]ORCID: 0000-0001-9016-9202.

[g]ORCID: 0000-0001-8488-1578.

[h]ORCID: 0000-0002-6751-1732.

[i]ORCID: 0000-0002-2759-123X.

[j]ORCID: 0000-0003-2300-9238.

[k]ORCID: 0000-0003-3660-8838.

[l]ORCID: 0000-0002-0344-0760.

[m]ORCID: 0000-0003-2140-9510.

[n]ORCID: 0000-0001-5739-2330.

[o]ORCID: 0000-0003-0352-2851.

[p]ORCID: 0000-0002-0142-5495.

[q]ORCID: 0000-0003-1556-1436.

[r]ORCID: 0000-0003-1679-4012.

[s]ORCID: 0000-0002-8521-9161.

[t]ORCID: 0000-0001-9014-710X.

[u]ORCID: 0000-0002-0958-1468.

[v]ORCID: 0000-0001-5551-7232.

[w]ORCID: 0000-0003-1919-3677.

[x]ORCID: 0000-0001-5508-2224.

The manuscript has 6,547 words, 1 table and 8 figures, 13 supplementary files, 33 supplementary figures, and 59 supplementary tables.

Abstract

Tung tree (*Vernicia fordii*) is an economically important woody oil plant that produces tung oil containing a high proportion of eleostearic acid (~80%). Here we report a high-quality, chromosome-scale tung tree genome sequence of 1.12 Gb with 28,422 predicted genes and over 73% repeat sequences. Tung tree genome was assembled by combining Illumina short reads, PacBio single-molecule real-time long reads and Hi-C sequencing data. Insertion time analysis revealed that the repeat-driven tung tree genome expansion might be due to long standing long terminal repeat (LTR) retrotransposon bursts and lack of efficient DNA deletion mechanisms. An electronic fluorescent pictographic (eFP) browser was generated based on genomic and RNA-seq data from 17 various tissues and developmental stages. We identified 88 nucleotide-binding site (NBS)-encoding resistance genes, of which 17 genes may help the tung tree resist the *Fusarium* wilt shortly after infection. A total of 651 oil-related genes were identified and 88 of them were predicted to be directly involved in tung oil biosynthesis. The fewer phosphoenolpyruvate carboxykinase (PEPC) genes, and synergistic effects between transcription factors and oil biosynthesis-related genes may contribute to high oil content in tung seeds. The tung tree genome should provide valuable resources for molecular breeding and genetic improvement.

KEYWORDS: Tung tree genome; Tung oil; Genome evolution; Electronic fluorescent pictographic browser; Oil biosynthesis

Introduction

Tung tree (*Vernicia fordii*), a woody oil plant native to China, is widely distributed in the subtropical area. Tung tree taxonomically belongs to the Euphorbiaceae family, along with several other economically important species including cassava (*Manihot esculenta*), castor oil plant (*Ricinus communis*), rubber tree (*Hevea brasiliensis*) and physic nut (*Jatropha curcas*). Species commonly referred to as tung trees include three major subspecies (*V. fordii*, *V. montana*, and *V. cordata*), of which *V. fordii* is the most widely cultivated species due to wide geographic distribution, medium stature

for easy plantation management, and high-quality oil production. Tung trees have been planted for tung oil production or ornamental purpose for more than 1000 years in China [1]. Tung trees have been widely distributed in 16 Chinese provinces and many countries after they were introduced into America, Argentina, Paraguay and other countries for plantation and tung oil production at the beginning of the 20th century [1] (Figure S1).

Tung seeds contain 50%−60% tung oil, which is composed of approximately 80% α-eleostearic acid (α-ESA), a type of unusual fatty acid. As the major component in tung oil, α-ESA has three conjugated double bonds (9 cis, 11 trans, 13 trans), and thus is easily oxidized. Due to its excellent characteristics, tung oil has been widely used as a drying ingredient in paints, varnishes, coating and finishes since ancient times [2]. Tung oil also can be used for synthesizing thermosetting polymers and resins with superior properties [3,4], and has been proposed as a potential source of biodiesel [5−7]. Tung oil was one of the chief exports until 1980s, and then declined due to the development of chemical coatings. Interestingly, tung oil has been attracted world-wide attention in recent years due to production security, environmental concerns, and negative effect of synthetic chemical coatings on human health [8−10]. New technologies have been developed to improve the performance of tung oil-based coatings [3,11,12].

As an oil crop, economic traits involved in fatty acid biosynthesis and oil accumulation are the targets of improved breeding efficiency for tung tree. However, identification of important genes, gene families and marker loci associated with oil content, fatty acid composition, and fruit yield has been hampered due to a lack of genetic and genomic information. Only a few functional genes, mainly involved in the formation and regulation of fatty acids such as *fatty acid desaturase* (*FAD2*, *FAD3*, *FADx*) and *diacylglycerol acyltransferase* (*DGAT*), have been investigated to date [13−17].

In the present study, we report the sequencing and assembly of *V. fordii* genome, which was achieved by combining whole-genome shotgun sequencing of Illumina short reads and real-time (SMRT) long reads on a Pacific Biosciences (PacBio) platform. We also used a Hi-C map to cluster the majority of the assembled contigs onto 11 pseudochromosomes. We conducted evolutionary comparisons and comprehensive transcriptome analysis of genes involved in oil biosynthesis to elucidate the genetic characteristics of oil synthesis and genetic difference as

compared to other plant species.

Results

Genome sequencing, assembly and validation

The self-bred progeny 'VF1-12' of *V. fordii* cv. Putaotong was used for genome sequencing (File S1). The genome of *V. fordii* was estimated to be 1.31 Gb in size with a low heterozygous rate of 0.0976% (Tables S2 and S3; File S2; Figure S4). After removing low-quality reads, we obtained a total of 177.68 Gb of high quality data, including 160.21 Gb of Illumina sequencing data and 187.47 Gb of SMART data, corresponding 135.73 × coverage of the tung tree genome (Table S4; Figure S5). The assembled tung tree genome was 1.12 Gb covering 85% of the estimated genome size, and contained 34,773 contigs with a maximum length of 544.11 Kb and 4,577 scaffolds with a maximum length of 5.09 Mb (**Table 1**; Table S5). Among them, 3,333 contigs and 29,721 scaffolds were more than 2 Kb in length (Table S5). After Hi-C data assessment and assembly, 1.06 Gb (95.15%) of the genome sequences were anchored onto 11 pseudochromosomes, with maximum clustered sequence lengths, minimum clustered sequence lengths and scaffold N50 of 120.57 Mb, 63.43 Mb and 87.15 Mb, respectively (Table 1; Tables S6−S11; **Figure 1**; Figures S6 and S7).

The CEGMA prediction indicated that 87.9% complete elements and 95.97% partial elements in tung tree genome could be hit for the 248 most conserved genes (Table S12). The BUSCO analysis showed that 1,379 (95.7%) of BUSCO genes were complete, of which 1338 (92.9%) and 41 (2.8%) were single-copy and duplicated, respectively (Table S13). RNA-seq data showed that 90.36%, 96.83% of flower samples 1 and 2 unigenes, 95.35%, 95.50%, and 96.48% of seed samples 1–3 unigenes showed good alignments to the assembled tung tree genome with mapping rate > 90%, respectively (Tables S14−S19). Furthermore, 88.3% to 95.6% of the reads from the five samples could be mapped to our genome assembly (Table S20). The validation results suggested that our tung tree genome assembly was of high quality in this study.

Genome annotations

In total, 28,422 genes were predicted with an average transcript length of 3,785 bp, average CDS length of 1,034 bp, average exon number of 4.85 per gene, average exon

length of 213 bp, and average intron length of 714 bp (Table 1; Table S21; Figure S8). The GC content was 31.93% across the genome, 41.91% in coding sequences and 31.16% in intron regions (Table 1; Tables S22−S24). BUSCO analysis showed that 1290 complete BUSCOs (89.6%) could be searched of all BUSCO groups, indicating that most of the gene models were complete (Table S25).

Among the total 28,422 genes, 23,143 genes (81.4%) were functionally annotated. Tremble, Swissprot and NR allowed the annotation of 79.6%, 63.8%, and 81.1% of all genes, respectively (Table S26). Gene ontology (GO) annotation revealed that 12,581 genes could be grouped into three categories with 65.97% in molecular function (GO:0003674), 20.1% in cellular component (GO:0005575), and 58.52% in biological process (GO:0008150) (Figure S9). We were able to use kyoto encyclopedia of genes and genomes (KEGG) to annotate 6835 genes to 235 pathways, of which oil biosynthesis and metabolism-related glycerolipid metabolism (ko00561), fatty acid biosynthesis (ko00061), fatty acid elongation (ko00062), and fatty acid degradation (ko00071) were of particular interest in this paper (Table S27).

In addition, we identified several types of non-coding RNAs in tung tree genome, including 465 microRNA (miRNA) genes, 740 transfer RNA (tRNA) genes, 116 ribosomal RNA (rRNA) genes, and 1414 small nuclear RNA (snRNA) genes (Table S28).

Gene family evolution and phylogeny

A total of 22,991 tung tree genes clustered into 15,038 gene families including 8,865 gene families shared by all eight species, and 635 tung tree-unique families and 5,431 tung tree-specific unclustered genes (Table S29). GO annotation of the tung tree-specific families showed that the genes involved in macromolecule metabolic processes (GO:0043170), cellular macromolecule metabolic processes (GO:0044260) and protein metabolic processes (GO:0019538) were highly enriched (Table S30; Figure S10). A total of 933 genes could be annotated using KEGG database, of which 586 genes were mapped to KEGG pathways. We observed KEGG enrichment in translation (110), carbohydrate metabolism (61), biosynthesis of other secondary metabolites (42), amino acid metabolism (44), folding, sorting and degradation (44), signal transduction (43), biosynthesis of other secondary metabolites (42), and environmental adaptation (36) (Table S31). We identified 11,985 gene families that were shared among the five Euphorbiaceae species (Figure S11A). The tung tree

193 shared 13,408, 13,387, 13,519, and 13,216 gene families with *J. curcas*, *H.*
194 *brasiliensis*, *M. esculenta*, and *R. communis*, respectively, of which 9,778, 6,643,
195 7,980, and 10,675 gene families had a one-to-one orthologous relationship (Figure
196 S11A). Additionally, compared with *A. thaliana*, *P. trichocarpa*, and *V. vinifera*, 3,421
197 gene families were found to be specific to Euphorbiaceae (Figure S11B).

198 A phylogenetic tree was generated based on a total of 2,085 single gene families
199 among the eight species (**Figure 2**A; Figure S12). We estimated that *V. fordii* and *J.*
200 *curcas* diverged around 34.55 million years ago (Mya) (Figure 2A). These data
201 indicate that *V. fordii* is more closely related to *J. curcas* than *M. esculenta*, *R.*
202 *communis*, and *H. brasiliensis* in Euphorbiaceae family, which is consistent with their
203 phylogenetic classification based on morphological characteristics.

204 The expansion and contraction of gene families occur since plants are subjected to
205 selection pressure during their evolution, thereby playing significant roles in plant
206 phenotypic diversification [18]. Expansion and contraction analysis on 15,662 shared
207 gene families based on the phylogenetic tree produced 475 expanded gene families
208 encompassing 1,612 genes, and 1,815 contracted families in tung tree as compared to
209 other plant species (Figure 2A). Of the 1,612 expanded genes, 839 could be annotated
210 using the GO database. GO annotation revealed highly enriched genes related to
211 macromolecule metabolic processes (GO:0043170), cellular macromolecule
212 metabolic processes (GO:0044260), and nucleotide binding (GO:0000166) (Table S32;
213 Figure S13).

214 The Ka/Ks ratio, also called ω or dN/dS, represents the number of
215 non-synonymous substitutions per non-synonymous site (Ka) to the number of
216 synonymous substitutions per synonymous site (Ks), indicating selective pressure
217 acting on a protein-coding gene in genetics. The values of Ks and Ka substitution
218 rates and the Ka/Ks ratio were estimated in each homologous cluster. A total of 586
219 positively selected genes (PSGs) in tung tree genome were identified, of which 475
220 were annotated using Swissprot functions (Table S33). GO annotation revealed that
221 the PSGs related to pigment metabolic processes (GO:0042440), mitochondrial
222 membrane (GO:0031966) and nuclear part (GO:0044428) are highly enriched (Table
223 S34; Figure S14).

224

225 **Whole genome duplication and collinearity**

226 All of the seven species showed peak 2 with the values ranging from 1.08 to 1.48 for

4DTV analysis, and 0.42 to 0.59 for Ks analysis (**Figure 3**). However, peak 1 was only observed in *V. fordii*, *J. curcas* and *R. communis* (Figure 2B). The results suggest that only an ancient genome triplication event (*i.e.*, γ event shared by core eudicots) and no recent independent whole-genome duplication (WGD) events occurred in the subsequent ~ 34.55 Mya evolutionary history in the tung tree lineage.

Plotting collinear regions of tung tree with itself showed that only 122 syntenic blocks containing 2,010 collinear gene pairs were identified in the tung tree genome (Figure 1; Table S35). A total of 3,559 genes comprised the collinear gene pairs, accounted for only 12.52% of tung tree genes, which is similar with *V. vinifera* (13.91%) and much lower than *M. esculenta* (33.86%) (Tables S36 and S37). The low collinear rate of tung tree genome suggests that a minority of the tung tree genome was duplicated during its evolution, which is consistent with the finding that the tung tree did not undergo a recent WGD event.

The tung tree genome generally showed a one-to-one and one-to-two syntenic relationships with *V. vinifera* (one duplication) and *M. esculenta* (two duplications), respectively (Figure 2C). Tung tree genome shared a total of 694 syntenic blocks, containing 22,133 collinear gene pairs with *M. esculenta*, and 589 syntenic blocks containing 14,570 collinear gene pairs with *V. vinifera* (Figure 2C; Figures S15 and S16). Most collinear regions between tung tree and *M. esculenta* revealed that one chromosome in tung tree corresponded to two chromosomes in *M. esculenta* (Figure 2C; Figure S17). For instance, VfChr1 in tung tree corresponded to MeChr12 and MeChr13 in cassava, and, similarly, VfChr2 to MeChr4 and MeChr11, VfChr3 to MeChr7 and MeChr10, VfChr5 to MeChr1 and MeChr2, as well as VfChr6 to MeChr1 and MeChr5, respectively. These results indicate that that VfChr1, VfChr2, VfChr3, and VfChr5 of tung tree might be formed by fragmentation and recombination of ancestral chromosomes. The collinear regions between tung tree and *V. vinifera* did not exhibit the remarkable corresponding chromosome relationships, in contrast to those between tung tree and *M. esculenta*.

Repeat-driven genome expansion

Tung tree had larger genome size than physic nut and castor bean, which was mainly attributed to repeat expansion in tung tree genome. Repetitive element analysis showed that tung tree genome harbored the greatest repeat content (73.34%) among the five sequenced Euphorbiaceae species (Table S40), which was slightly higher than

the rubber tree (71%) [19], and much higher than the castor oil plant (50.33%) [20], Physic nut (49.8%) [21], and cassava (less than 40%) [22]. The repeat sequences were distributed at both ends of each tung tree chromosome (Figure 1). We identified 66,3931 simple sequence repeats (SSRs) in the tung tree genome. The annotated SSRs were mostly mononucleotide (39.62%) and dinucleotide (13.38%) (File S3). Retroelements comprised the majority (51.89%) of the tung tree genome, of which 50.77% belonged to long terminal repeat (LTR) retrotransposons (Table S41). Of the repeat sequences, two types of LTR retrotransposons, *Ty1/Copia* (84,180 in number) and *Ty3/Gypsy* (284,597 in number) were most abundant, accounting for 15.13% and 53.46%, respectively (Figure 3A and B; File S3; Table S41). The *Ty1/Copia* and *Ty3/Gypsy* were ~ 0.53 Gb of total length, occupying 50.31% of the assembled tung tree genome.

Kimura analysis showed that two LTR retrotransposons (*Ty1/Copia* and *Ty3/Gypsy*) and DNA transposons were almost simultaneously amplified, with similar peaks for amplification bursts (Figure S18). Insertion time analysis of intact LTR retrotransposons indicated that both of *Ty1/Copia* and *Ty3/Gypsy* experienced multiple bursts over the last 3-4 Mya and they were younger than other unclassified transposable elements (File S3; Figures S19 and S20). In addition, median-copy families and high-copy families were younger than single-copy families (Figure S21). In light of our analysis, the dramatic expansion in tung tree genome size might be due to long standing LTR retrotransposon bursts and lack of efficient DNA deletion mechanisms. VL0001 was the largest *Ty3/Gypsy* family with 130 copies, accounting for 7.54% of the high-copy families and 4.35% of LTR retrotransposons (Figure 3C; Table S42).

Based on our RNA-seq data, 1,738 out of the total 2,991 LTR retrotransposons were expressed across six tissues. *Ty3/gypsy* LTR retrotransposons generally exhibited higher expression levels than *Ty1/Copia* retrotransposons, ranging from 0.71-fold in seed to 4.09-fold in leaf with approximately two-fold higher on average (File S3; Table S44). Among the 1,738 LTR retrotransposons, 701 showed the highest expression level in seeds, of which 60.77% belonged to high-copy families (Figure 3D; File S3). This suggests that abundant high-copy LTR retrotransposons may be more active than other LTR retrotransposons families in developing tung seeds. In addition, 184, 204, 244, 148, and 257 LTR retrotransposons exhibited the highest expression levels in root, stem, leaf, female flower, and male flower, respectively (File

S3; Figure S22). Among these LTRs, high-copy LTR families also accounted for the highest proportion in the other five tissues.

The tung tree eFP browser

A total of 28,422 genes were identified from the tung tree genome, of which 23,143 genes were annotated. The genome-wide gene identification allowed us to investigate gene expression on a large-scale in tung tree. To provide easy access and enable visualization of the expression levels of tung tree genes, flowers and seeds at different developmental stages were sampled for RNA-seq analysis (File S4). Based on RNA-seq data from 17 tung samples, a "Tung Tree eFP Browser" (at http://bar.utoronto.ca/efp_tung_tree/cgi-bin/efpWeb.cgi) was implemented to permit visualization of gene expression patterns with "absolute", "relative" and "compare" modes in these tissues using the annotated gene IDs (File S4). The search interface generated an "electronic fluorescent pictograph" colored according to transcript abundance data for individual tung tree gene in various tissues/organs. As exemplified (Figure S23), the *VfFADx-1* (Vf11G0298) using linoleic acid (C18:2Δ9,12) as substrates to produce α-ESA (18:3Δ9,11,13) exhibited expression patterns consistent with its role in oil biosynthesis. In addition, the Tung Tree eFP Browser could be used for functional characterization of tung tree gene copies with different expression patterns. For instance, three feruloyl CoA ortho-hydroxylase homologues (Vf03G0652, Vf00G0634 and Vf03G0623) exhibited conservation of function as revealed by similar expression patterns in various tissues/organs (**Figure 4**; Table S51). Among three purple acid phosphatase homologues, the Vf11G0977 displayed neofunctionalization, *i.e.*, functional diversification due to its expression in roots compared to the other homologues (Figure 4; Table S51).

NBS-encoding resistance genes

Disease resistance is one of the most important traits in tung tree breeding programs. The *V. fordii* is susceptible to wilt (*Fusarium oxysporum*), black spot (*Mycsphaqrella aleuritids*) and twig dieback (*Nectria aleuriidia*). Information on disease resistance-related genes will be helpful for understanding plant resistance mechanisms. Identification and characterization of these genes on a genome-wide scale will provide a basis for improvement of disease resistance in tung tree. Genes encoding nucleotide-binding sites (NBSs) are the largest class of plant disease resistance genes.

Based on whether they contain a Toll/interleukin-1 receptor (TIR) domain, NBS resistance genes can be further categorized into two subclasses (TIR and non-TIR) (File S5).

A total of 88 genes with an NBS domain were identified in tung tree, of which 28 (31.82%) were organized in tandem arrays (Table S52; **Figure 5**A; Figure S25). The number of NBS-encoding genes in *V. fordii* was similar to *Z. mays* (107), but remarkably lower than *R. communis* (232), *M. esculenta* (312), *J. curcas* (275), and *H. brasiliensis* (483) (Table S52). The 88 NBS-encoding genes were classified into four subfamilies, including 23 coiled-coil (CC)-NBS, 16 NBS-leucine-rich repeat (LRR), 7 CC-NBS-LRR, and 42 NBS, however they did not form four independent classes in the phylogenetic tree (Figure 5A). Intriguingly, all of the tung tree NBS-encoding resistance genes do not belong to the TIR type (Table S52).

The NBS genes were distributed nonrandomly across all 11 chromosomes (Figure S24). More than 85% *NBS* genes were clustered in groups, and clusters were most abundant on chromosomes 2, 9, and 3 (Figure S24). Enrichment of *NBS* genes in these corresponding genomic regions indicated that resistance gene evolution might involve tandem duplication and divergence of linked gene families, as described in other plant genomes such as rubber tree [23] and pear [24]. RNA-seq data showed that the 88 tung tree NBS genes displayed differential expression patterns in roots after *F. oxysporum* infection (Figure 5B; File S5). The expression level of 17 genes including 8 *NBS*s, 3 *NBS-LRR*s, 2 *CC-NBS*s, and 4 *CC-NBS-LRR*s increased at early stage after infection (FOE) and decreased at late stage after infection (FOL) (Figure 5B). These results suggest that these genes may help the tree resist the pathogen shortly after infection.

Evolution of genes involved in oil biosynthesis

Tung oil is the most important product from tung tree. Tung oil biosynthesis starting from acetyl-CoA involves 18 enzymatic steps with multiple isozymes in each step (**Figure 6**A). The oil is packed in subcellular structures called oil bodies or lipid droplets (Figure 6B; File S6). Tung seed oil droplets formed following the pattern of α-ESA accumulation in the seeds (Figure 6B and C). No visible oil droplet was observed in 10 weeks after flowering (WAF) seeds and small oil droplets were observed in 15 WAF seeds. The number and sizes of oil droplets were dramatically increased in more mature seeds (20, 25, and 30 WAF).

Tung oil biosynthesis in the seeds started in mid-June (10 WAF), increased rapidly until 25 WAF with the oil content of 55.42% (Figure 6C), and ended by 30 WAF. Oleic acid (C18:1Δ9) accounted for minor percentage, whereas linoleic acid (C18:2Δ9,12) accounted for the major content (43%) in young seeds (10 and 15 WAF). Both gradually decreased in more mature seeds. Accumulation of linoleic acid and α-ESA (α-C18:3Δ9,11,13) showed opposite patterns in the developing tung seeds (Figure 6C) because linoleic acid is the same substrate for synthesizing α-ESA and α-ALA (α-linolenic acid, C18:3Δ9,12,15). The α-ESA synthesis started after 15 WAF and then increased rapidly up to 72.35% of seed oil following seed ripening (Figure 6C). The α-ALA accumulation was observed in 10 WAF seeds and accounted for minor percentage during the whole developmental stages, although it shares the same substrate with α-ESA. These developmental patterns of α-ESA biosynthesis and oil droplet formation were used as the criteria for selecting seed stages for our transcriptomic analysis.

We annotated 22,419 genes in the tung tree genome and identified 651 genes related to oil biosynthesis (Table S53). Among them, 88 genes were predicted more directly involved in oil biosynthesis (Figure 6A; File S7; Table S54). This study provided far more tung oil-related genes than those deposited in the GenBank databases (29 genes). These genes belonged to 18 families whose expression profiles were described in Figure 6A. The number of tung oil-related genes (88) was within the range of other plant species including 91 genes in *J. curcas*, 84 genes in *R. communis*, 87 genes in *A. thaliana*, 105 genes in *S. indicum*, and 210 genes in *G. Max* (Table S54).

Several key genes important in oil biosynthesis have been studied extensively, including acetyl CoA carboxylase (*ACCase*), *FAD*s, *DGAT*s and *oleosins* (*OLE*s) (Figure 6A). The current study indentified one additional *DGAT3* and two additional *FAD*s besides those reported previously. We also reported for the first time that tung tree genome had six *phospholipid:diacylglycerol acyltransferase* genes (*PDAT*) (Figure 6A).

ACCase and phosphoenolpyruvate carboxykinase (PEPC) are probably the key enzymes determining the metabolic pathways towards oil or protein biosynthesis in the seeds (Figure 6A) [25]. We identified nine *ACCase* genes in tung tree genome with high expression levels in the mid-late developing stages of tung seeds (Figure 6A). There are 10 *ACCase* genes in soybean, and 6-7 genes in other species (Table

S54). We also identified three *PEPC* genes in tung tree genome which were expressed in the early developing stages of tung seeds (Figure 6A; Table S54). There are 16 *PEPC* genes in soybean and more *PEPC* genes in other species than tung tree. Comparison of soybean whose seed has high protein content (~ 40%) and low oil content (~ 20%), the fewer *PEPC* genes in tung tree genome might be the reason of high oil (~ 55%) and low protein content (~ 5%) in tung seeds, probably contributing to carbon flow towards fatty acid biosynthesis in tung seeds.

FAD protein family catalyzes the desaturation of fatty acids [5] and therefore is responsible for polyunsaturated lipid synthesis in developing seeds of oil crops. FAD2 and FAD3 are the main enzymes responsible for the Δ12 linoleic acid and Δ15 linolenic acid desaturation, respectively. We identified one *FAD2*, two *FAD3* and two *FADx* genes in tung tree (Table S54). *FAD2* and *FADx-1* were highly expressed in mid-late stages of developing seeds; whereas *FAD3* was expressed higher in early stages of seeds (Figure 6A). FADx, a divergent FAD2, converts linoleic acid to α-ESA, the major component of tung oil [14], but the evolutionary relationship between FADx and FAD2 is still uncertain. According to the newly generated phylogenetic tree in this study (**Figure 7**), we found FAD2/x clade could be divided into two clades (FAD2 and FADx) in eudicot plants, suggesting that the two clades were due to gene duplication in eudicot ancestors. The eudicot ancestors have γ WGD event, and gene duplication is likely to be retained by the WGD event. Further synteny analysis revealed that FAD2s and FADxs were likely to be generated by WGDs event (Table S56), which corresponded to the γ WGD event shared by core eudicots. Notably, the FADx clade lost many genes in species like the members of Brassicaceae.

DGAT protein family catalyzes the last step of triacylglycerol (TAG) biosynthesis and is regarded as the rate-limiting step for TAG accumulation. Three DGATs were reported in tung tree in previous studies. *DGAT2* was proposed to be the most important *DGAT* for TAG biosynthesis in tung tree seeds. Our transcripomics study found four *DGAT*s (*DGAT1*, *DGAT2*, and two *DGAT3*) expressed in tung seeds (Figure 6A; Table S55). *DGAT2* was confirmed to be the most abundantly expressed *DGAT* in tung seeds which corresponded to oil accumulation (20–30 WAF), but *DGAT3-1* was the dominant form of *DGAT* in immature seeds (10–15 WAF) and other tissues including stem, root, leaf and female flower (Figure 6A; Table S55).

Recently, it has become obvious that TAG synthesis also can be catalyzed by PDAT. We reported for the first time that there were five PDATs in tung tree genome.

PDAT1-1 and *1-4* were expressed more in mid-late stages of developing seeds but the other three *PDAT* genes were expressed more in the early stages of developing seeds (Figure 6A).

OLEs are the major proteins in plant oil bodies. Genome-wide phylogenetic analysis and multiple sequence alignment demonstrated that the five tung *OLE* genes represented the five OLE subfamilies and all contained the "proline knot" motif (PX5SPX3P) shared among 65 OLE from 19 tree species [26]. We confirmed the five tung tree *OLE* genes coding for small hydrophobic proteins. These five *OLE*s were highly expressed in mid-late stage of developing tung seeds (Figure 6A; Table S55).

A total of eight *long chain fatty aycl-CoA synthetases* (*LACS*) genes were identified in tung tree genome, of which *LACS1* and *2* were more highly expressed in early stage but *LACS7*, *8* and *9* were highly expressed in mid-late stages of developing seeds (Figure 6A). Additionally, 9 *glycerol-3-phosphateacyltransferases* (*GPAT*s), 7 *lysophosphatidic acid acyltransferases* (*LPAT*s), and 6 *phosphatidate phosphatases* (*PP*s) genes were identified in tung tree genome whose expression levels of some genes were higher in early stage rather than late stages of developing seeds and verse visa (Figure 6A; Table S55).

To explore possible synergistic effects among genes in oil accumulation, we performed a weighted correlation network analysis of transcript expression in developing seeds at five stages (FPKM values $\geq$ 1) (File S8). We identified 10 co-expression modules for each stage sample, among which oil biosynthesis-related genes at 20 WAF were highly enriched in two significant modules (PCC values $\geq$ 0.8 , *P* value $\leq$ 0.1): MEbrown and MEyellow containing 1,156 and 908 genes, respectively (Tables S57 and S58; Figures S31 and S32). We did not find oil biosynthesis-related genes in other significant modules. In MEyellow and MEbrown modules, 18 and 13 genes were respectively identified to play pivotal roles in fatty acid synthesis and oil accumulation, such as fatty acid synthases (FASs), the upstream rate-limiting enzyme *ACCase* subunits (*α-CT*, *BCCP-1*, *BCCP-2*, *BCCP-2*, and *BC-1*), and genes related to TAG assembly like *GPDH*, *LPAT*, etc (**Figure 8**). A number of transcription factors were also identified in the two modules and co-expression networks (Figure 8) including *WRINKLED1* (*WRI1*), *FUSCA3* (*FUS3*), *LEAFY*

462 *COTYLEDON1* (*LEC1*), and *ABSCISIC Acid INSENSITIVE3* (*ABI3*), which has been
463 reported to facilitate oil accumulation by interacting each other or with oil
464 biosynthesis-related genes [27−31]. We selected four tung tree transcription factors
465 (*FUS3*, *ABI3*, *LEC1-1* and *LEC1-2*) to conduct yeast two-hybrid assay (File S9) and
466 observed that *FUS3* and *LEC1-2* were interacted (Figure S33). The gene
467 co-expression networks indicate that transcription factors and oil biosynthesis-related
468 genes have synergistic effects in oil biosynthesis, which may contribute to high oil
469 content in tung seeds.
470

471 ## Discussion

472 The whole genomes of an increasing number of plant species have been sequenced
473 due to rapid development of new sequencing technologies in recent years. The
474 genome information provides researchers a useful resource for better understanding
475 plant evolutionary history and exploring important genes to uncover the mechanisms
476 controlling various traits during long-term evolution process. As an economically
477 important tree species, tung tree has been cultivated and utilized for thousands of
478 years. Presently its oil has a great potential for producing environmentally-friendly
479 coatings with low VOCs. However, producing tung oil on an industrial scale is
480 hampered by low yield. Our genome sequencing effort will facilitate the breeding of
481 elite cultivars with yield-related traits including fruit setting rate and seed oil content.
482 In this study, the large amount of repeat sequences and low GC content made the tung
483 tree genome a challenge for WGS strategies using NGS technology even though the
484 tung tree genome was estimated to be extremely low heterozygosity. To overcome the
485 challenge of high repeat content, we generated long reads from 10 kb and 20 kb
486 libraries via PacBio sequencing. Finally, we used the Hi-C map to generate a
487 chromosome-scale assembly of the tung tree genome. The genome sequence covered
488 ~ 85.50% of the estimated genome size and harbored 28,422 genes. Among the
489 Euphorbiaceae family, rubber tree and cassava instead of tung tree, physic nut and
490 castor bean were found to have undergone a recent WGD event, although they all

shared an ancient WGD event. Interestingly, rubber tree and cassava have more genes than the other three species (Figure 2A). The recent WGD event could cause chromosomal rearrangements, fissions or fusions and is one of the reasons resulting in expansion of gene families [18], which may contribute to more gene expansions in rubber tree and cassava than those in tung tree, physic nut and castor bean. The genome sequence of tung tree opens a window to functional and molecular breeding of economically important woody oil plants within the Euphorbiaceae family.

Tung tree had a larger genome size than physic nut and castor bean. In most cases, genome expansions are caused by repeated sequence insertion, like those occurred in tea tree, rubber tree, and Ginkgo (*G. biloba*) [32]. Similar to the three species, *Ty3/Gypsy* families contributed the most to the tung tree genome expansion. Based on our insertion time analysis, we proposed that lack of efficient deleting mechanisms of repeated DNA sequences might have resulted in long-term and continuous LTR retrotransposon bursts and growth, eventually leading to the whole genome size expansion. This is also consistent with the findings in tea tree and *P. abies* [33]. We also found that different LTR retrotransposon families were differentially expressed in various tissues, confirming the retrotransposon activity in the tung tree genome. The eFP Browser has proved to be a useful tool to display gene expression levels visually in several plant species including *A. thaliana*, *P. trichocarpa*, *G. Max*, *S. tuberosum*, *S. lycopersicum*, *C. sativa*, *F. vesca* and other species [34−37]. Based on tung tree genome sequences generated in this study, we created a Tung Tree eFP Browser to display tung tree RNA-seq data from 17 different tissues and stages. This eFP Browser work should facilitate further research in tung tree and other Euphorbiaceae plants.

Plant disease resistance has always been a research hotspot. NBS genes are the largest class of plant disease resistance genes. They confer the capacity for the plant to resist the intrusion of outside pathogens, including bacteria, fungi and virus [38]. The present studies suggested that the TIR domain-containing NBS genes are widely distributed in dicots but not monocots, whereas they are lost in tung tree genome. To date only tung tree and sesame [22] out of dicots have been reported for TIR domain-containing NBS gene loss. This finding provides a new paradigm to

522 investigate the evolution of disease resistance genes. CC is the functional domain of
523 many proteins and CC structure plays an important role in protein-protein interaction
524 [39]. LRR is the signal region in transmembrane domain and loss of it can result in
525 loss of function [40]. In this study, the highest proportion of *CC-NBS-LRR* genes (4/7,
526 57.14%) responded to *F. oxysporum* infection at early stage, suggesting that CC and
527 LRR domains play more important roles than other domains.

528 Tung tree is a high-efficient photosynthetic tree with strong photosynthesis rate.
529 Sucrose, the major photosynthesis product, is synthesized in the chloroplast and
530 exported to the sink tissues such as seeds for seed development and metabolite
531 accumulation. Sucrose is converted into hexose phosphate, triose phosphate,
532 phosphoenolpyruvate (PEP), and pyruvate. PEP is a key intermediate metabolite for
533 synthesizing both fatty acids and proteins. PEP is converted into pyruate by pyruate
534 kinase (PK), which is subsequently converted into acyl-CoA and enters fatty acid
535 biosynthesis pathway after ACCase action. On the other hand, PEP is catalyzed by
536 PEPC to produce oxaloacetic acid, which is subsequently used for protein synthesis.
537 Therefore, ACCase and PEPC are probably the keys enzymes determining the
538 metabolic pathway towards oil or protein biosynthesis in the seeds [25]. We identified
539 nine *ACCase* genes in tung tree genome with high expression levels in the mid-late
540 developing stages of tung seeds, which are indicative of their importance in tung oil
541 biosynthesis. There are 10 *ACCase* genes in soybean, and 6-7 genes in other species.
542 We also identified three *PEPC* genes in tung tree genome with high expression levels
543 in the early developing stages of tung seeds. By contrast, there are 16 *PEPC* genes in
544 soybean and more *PEPC* genes in other species than tung tree. Because soybean has
545 more *PEPC* genes and higher protein/ lower oil content in the seed, it is possible that
546 the fewer *PEPC* genes in tung tree diverted less carbon flow towards protein
547 biosynthesis and resulted in high oil/low protein content in tung seeds.

548 Tung oil is the major economically important product from tung tree.
549 Identification and characterization of all genes involved in tung oil biosynthesis is
550 essential for improving tung oil production and economic value. Many tung oil
551 biosynthetic genes have been identified in our laboratories, including those coding for

552 *diacylglycerol acyltransferases* (*DGAT*) [13,17], *delta-12 oleic acid desaturase*
553 (*FAD2*) and *delta-12 fatty acid conjugase* (*FADx*) [14], *omega-3 fatty acid desaturase*
554 (*FAD3*) [41], *acyl-CoA binding proteins* [42], *cytochrome b5* [43], *cytochrome b5*
555 *reductase* [15], *glycerol-3-phosphate acyltransferase* (*GPAT*) [44], *plastid-type*
556 *omega-3 fatty acid desaturase* (*TnDES2*) [45], *aquaporin* and *glutaredoxin* [46], and
557 *ß-ketoacyl-ACP synthase* (*KAS*) [47]. Interestingly, we identified an additional *FADx*
558 gene, *FADx-2*, which was probably generated by gene duplication and
559 sub-functionalization based on the different expression patterns of *FADx-1* and
560 *FADx-2* genes. In comparison with *FADx-2*, *FADx-1* was the dominant form
561 responsible for α-ESA synthesis in developing seeds of tung tree. We also identified 9
562 *ACCases*, 4 *DGAT*s, 7 *FAD*s, 6 *PDAT*s, 5 *OLE*s, 8 *LACS*s, 9 *GPAT*s, 7 *LPAT*s, and 6
563 *PP*s genes in the tung tree genome. This study provided a more complete picture for
564 genes involved in tung oil biosynthesis. The numbers of tung oil-synthesizing genes
565 are within the range of other species. These suggest that there is no gene expansion in
566 tung tree and the amount and types of oils in various species may not be directly
567 related to the number of genes in oil biosynthesis.

568 Transcriptomic analysis evaluated the expression profiles of all these genes. Our
569 results indicated that the expression patterns of some of the most important genes
570 were well-coordinated with oil biosynthesis and accumulation in tung tree seeds.
571 Specifically, *DGAT2* was shown to be the most abundantly expressed *DGAT* in tung
572 seeds but *DGAT3-1* was the dominant form of *DGAT* in immature seeds and other
573 tissues including stem, root, leaf and female flower, in agreement with our previous
574 results [13,17]. *FAD2* and *FADx* were highly expressed in mid-late stages of
575 developing seeds; whereas *FAD3* was expressed higher in early stages of seed, also in
576 agreement with published results [14]. All five *OLE*s were highly expressed in
577 mid-late stage of developing tung seeds, similar results to what we reported
578 previously [26]. Our expression analysis provided novel insights into the potential
579 role of *PDAT*s in tung oil biosynthesis by showing that *PDAT1-1*, *1-4*, and *2-2* were
580 expressed more in mid-late stages of developing seeds but the other three *PDAT* genes
581 were expressed more in the early stages of developing seeds, which were not reported

582 previously. Our gene co-expression analysis revealed that oil biosynthesis-related
583 genes were enriched in two significant modules only at 20 WAF when the seed oil
584 started to accumulate rapidly. The enriched oil biosynthesis-related genes included
585 most of FAS genes, part of TAG biosynthesis genes and some transcription factors.
586 The complete gene co-expression networks provide insights into oil biosynthesis by
587 gene-gene synergistic function.

588 In conclusion, this study provides whole-genome sequence information, eFP
589 browser, and extensive RNA-seq data. These critical lines of information should be
590 used as valuable resources for functional genomics studies and tree improvement of
591 economically important traits such as oil content and disease resistance in the tung
592 tree.

593

594 ## Materials and methods

595 ### Plant materials

596 The self-bred progeny 'VF1-12' of the elite *V. fordii* cv. Putaotong was used for whole
597 genome sequencing in this study (File S1). Young leaves were collected from 'VF1-12'
598 in the spring for genome sequencing. Young plantlets were used for Hi-C library
599 construction and sequencing. A total of 17 fresh tissues including stems, roots, male
600 flowers, female flowers, and seeds at different developmental stages were collected
601 for RNA-seq. The developing seeds were also used for oil content measurement and
602 fatty acid analysis.

603

604 ### Whole-genome sequencing, assembly and assessment

605 The tung tree genome size was estimated by a modified Lander-Waterman algorithm
606 *i.e.*, a formula G = Bnum/Bdepth = Knum/Kdepth [48]. Heterozygosity was estimated
607 by the k-mer distribution and GenomeScope [49]. Nuclear DNA was isolated from
608 fresh leaf tissues by using a DNeasy Plant Mini Kit (Qiagen, CA, USA). A series of
609 DNA libraries were constructed and sequenced with an Illumina HiSeq 2000
610 sequencing platform (Illumina, CA, USA) (File S10). In addition,
611 SMRTbell template libraries of 20 kb were constructed and sequenced on the PacBio
612 RSII. After removing low-quality reads, the whole genome assembly of tung tree was
613 performed with a hierarchical assembly strategy due to its homozygous genome with

highly repetitive sequences (File S11). The genome completeness was assessed by Core Eukaryotic Genes Mapping Approach (CEGMA) [50], Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis [51] and RNA-seq reads mapping [52].

Hi-C data preparation and contig clustering

The Hi-C library was prepared with the standard procedure described [53]. Raw Hi-C data were generated using HiSeq2500 sequencing platform and then were processed to filter low-quality reads and trim adapters. Clean reads were mapped to the assembled scaffolds by BWA-aln after truncating the putative Hi-C junctions in sequence reads. HiC-Pro software (version 2.7.1) was used to filter the invalid ligation read pairs, including dangling-end and self-ligation, re-ligation and dumped products. Finally the scaffolds were clustered, ordered and orientated onto chromosomes using the valid read pairs by LACHESIS (http://shendurelab.github.io/LACHESIS/).

Genome annotation

Gene prediction was conducted using *de novo* prediction, homology information and RNA-seq data (File S12). Gene functions were assigned according to the best match derived from the alignments to proteins annotated in SwissProt and TrEMBL databases using Blastp, and the pathway in which the gene might be involved was annotated by KAAS [54]. Motifs and domains were annotated using Inter ProScan (Version 5.2-45.0) [55] by searching against publicly available databases in InterPro [56]. The rRNA, snRNA and miRNA genes were predicted by INFERNAL software using the Rfam database. The rRNA subunits were identified by RNAmmer [57] based on hidden Markkov models (HMMs). The tRNA genes were predicted by tRNAscan-SE [58] with eukaryote parameters. A *de novo* and homology-based approach was used to identify repetitive sequence and transposable elements (TEs) in the tung tree genome.

Evolutionary analysis

Phylogeny of a total of eight species was constructed based on single-copy gene families by the maximum likelihood (ML) method (File S13). The divergence times were estimated based on all single-copy genes and 4-fold degenerate sites with the program MCMCTree of the PAML package [59]. The neutral evolutionary rate was

calculated via Bayes estimation with Markov Chain Monte Carlo algorithm. Gene families which underwent expansions or contractions were identified using the CAFE (Computational Analysis of gene Family Evolution) program [60].The selection pressure of tung tree in the phylogenetic tree was calculated by Codeml. The significance of the identified PSGs was verified using a Chi-square test. WGD events were identified by 4DTv (four-fold synonymous third-codon transversion) and synonymous Ks analysis.

Data access

The project of tung tree genome sequencing, Hi-C and transcriptome sequencing is registered at NCBI under BioProject accession PRJNA503685 (http://www.ncbi.nlm.nih.gov/bioproject/503685), PRJNA445350 (http://www.ncbi.nlm.nih.gov/bioproject/445350) and PRJNA483508 (http://www.ncbi.nlm.nih.gov/bioproject/483508). The data are publicly available at NCBI under accession number SUB4731026, SRP136294 and SRP155790, respectively.

Author contributions

XFT and LZ conceived and initiated the study. XFT, LZ, and HPC designed experiments and coordinated the project. XFT, LZ, HXL, MLL, ZL, YLZ, and HC performed the field controlled pollination and managing and sampling experimental materials for genome and transcriptome sequencing. JH and DPW supervised the data generation and analysis. MZ, JJL, FL, JH, and DPW conducted the genome assembly and annotation. MZ and MLX were involved in the WGD event determination. XMY, MZ, and LSZ performed repeat annotation, phylogenetic analysis and expression analysis. AP, EE, WYL, and NJP performed tung tree eFP browser construction. WD, AXS, and LSZ coordinated collinear analysis and phylogenetic analysis. MLL, DR, and GZ performed NBS gene family analysis. HXL, MLL, LZ, and HPC contributed to identification, expression and phylogenetic analysis of oil-related gene families. WD and MLL conducted gene co-expression analysis. MLL performed Yeast two-hybrid

678 assay. WYL and MLL drew all figures of the manuscript. LZ and HPC wrote the
679 manuscript. LZ, HPC, NJP, LSZ, and XFT revised the manuscript. All authors
680 discussed results and commented on the manuscript.
681

682 ## Competing interests

683 The authors have declared no competing interests.
684

685 ## Acknowledgments

686 This work was supported by the National Key R&D Program of China (Grant No.
687 2017YFD0600703), the National Natural Science Foundation of China (Grant No.
688 31770720), the outstanding youth of the Education Department of Hunan Province
689 (Grant No. 17B279), and the USDA-ARS Quality and Utilization of Agricultural
690 Products National Program 306 through CRIS 6054-41000-103-00-D. Mention of trade
691 names or commercial products in this publication is solely for the purpose of providing
692 specific information and does not imply recommendation or endorsement by the U.S.
693 Department of Agriculture. USDA is an equal opportunity provider and employer.
694

695 ## References

696 [1] Tan XF, Jiang GX, Tan FY, Zhou WG, Lv PH, Luo KM, et al. Research report on
697 industrialization development strategy of *Vernicia fordii* in China. Nonwood
698 Forest Res 2011;29:1−7.
699 [2] Zhang L, Jia B, Tan X, Thammina CS, Long H, Liu M, et al. Fatty acid profile and
700 unigene-derived simple sequence repeat markers in tung tree (*Vernicia fordii*).
701 PLoS One 2014;9:e105298.
702 [3] Huang Y, Pang L, Wang H, Zhong R, Zeng Z, Yang J. Synthesis and properties of
703 UV-curable tung oil based resins via modification of Diels–Alder reaction,
704 nonisocyanate polyurethane and acrylates. Prog Org Coat 2013;76:654−61.
705 [4] Liu CG, Shang QQ, Jia PY, Dai Y, Zhou YH, Liu ZS. Tung oil-based unsaturated
706 co-ester macromonomer for thermosetting polymers: Synergetic synthesis and

copolymerization with styrene. ACS Sustain Chem Eng 2016;4:3437−49.

[5] Park JY, Kim DK, Wang ZM, Lu P, Park SC, Lee JS. Production and characterization of biodiesel from tung oil. Appl Biochem Biotechnol 2008;148:109−17.

[6] Shang Q, Lei J, Jiang W, Lu H, Liang B. Production of tung oil biodiesel and variation of fuel properties during storage. Appl Biochem Biotechnol 2012;168:106−15.

[7] Chen YH, Chen JH, Luo YM. Complementary biodiesel combination from tung and medium-chain fatty acid oils. Renew Energy 2012;44:305−10.

[8] Meininghaus R, Gunnarsen L, Knudsen HN. Diffusion and sorption of volatile organic compounds in building materials-impact on indoor air quality. Environ Sci Technol 2000;34:3101−8.

[9] Tsakas MP, Siskos AP, Siskos PA. Indoor air pollutants and the impact on human health, chemistry, emission control, radioactive pollution and indoor air quality. InTech, 2011: 447-484.

[10] Wei WJ, Zhang YP, Xiong JY, Li M. A standard reference for chamber testing of material VOC emissions: Design principle and performance. Atmos Environ 2012;47:381−8.

[11] Yang XH, Zhang SX, Li WY. The performance of biodegradable tung oil coatings. Prog Org Coat 2015;85:216−20.

[12] Yoo Y, Youngblood JP. Tung oil wood finishes with improved weathering, durability, and scratch performance by addition of cellulose nanocrystals. ACS Appl Mater Inter 2017;9:24936−46.

[13] Shockey JM, Gidda SK, Chapital DC, Kuan J-C, Dhanoa PK, Bland JM, et al. Tung tree DGAT1 and DGAT2 have nonredundant functions in triacylglycerol biosynthesis and are localized to different subdomains of the endoplasmic reticulum. Plant Cell 2006;18:2294−313.

[14] Dyer JM, Chapital DC, Kuan JC, Mullen RT, Turner C, McKeon TA, et al. Molecular analysis of a bifunctional fatty acid conjugase/desaturase from tung. Implications for the evolution of plant fatty acid diversity. Plant Physiol

737 2002;130:2027−38.

738 [15] Shockey JM, Dhanoa PK, Dupuy T, Chapital DC, Mullen RT, Dyer JM. Cloning,
739 functional analysis, and subcellular localization of two isoforms of
740 NADH:cytochrome b5 reductase from developing seeds of tung (*Vernicia fordii*).
741 Plant Sci 2005;169:375−85.

742 [16] Cao HP, Chapital DC, Howard OD, Deterding LJ, Mason CB, Shockey JM, et al.
743 Expression and purification of recombinant tung tree diacylglycerol
744 acyltransferase 2. Appl Microbiol Biot 2012;96:711−27.

745 [17] Cao HP, Shockey JM, Klasson KT, Chapital DC, Mason CB, Scheffler BE.
746 Developmental regulation of diacylglycerol acyltransferase family gene
747 expression in tung tree tissues. PloS One 2013;8:e76946.

748 [18] Zhang L, Li X, Ma B, Gao Q, Du H, Han Y, et al. The tartary buckwheat genome
749 provides insights into rutin biosynthesis and abiotic stress tolerance. Mol Plant
750 2017;10:1224−37.

751 [19] C T, M Y, Y F, Y L, S G, X X, et al. The rubber tree genome reveals new insights
752 into rubber production and species adaptation. Nat Plants 2016;2:16073.

753 [20] Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, et al. Draft genome
754 sequence of the oilseed species *Ricinus communis*. Nat Biotechnol
755 2010;28:951−U3.

756 [21] Wu PZ, Zhou CP, Cheng SF, Wu ZY, Lu WJ, Han JL, et al. Integrated genome
757 sequence and linkage map of physic nut (*Jatropha curcas* L.), a biodiesel plant.
758 Plant J 2015;81:810−21.

759 [22] Wang LH, Yu S, Tong CB, Zhao YZ, Liu Y, Song C, et al. Genome sequencing of
760 the high oil crop sesame provides insight into oil biosynthesis. Genome Biol
761 2014;15: R39.

762 [23] Pootakham W, Sonthirod C, Naktang C, Ruang-Areerate P, Yoocha T, Sangsrakru
763 D, et al. *De novo* hybrid assembly of the rubber tree genome reveals evidence of
764 paleotetraploidy in Hevea species. Sci Rep 2017;7: 41457.

765 [24] Wu J, Wang ZW, Shi ZB, Zhang S, Ming R, Zhu SL, et al. The genome of the
766 pear (*Pyrus bretschneideri* Rehd.). Genome Res 2013;23:396−408.

767 [25] Chen JQ, Lang CX, Hu ZH, Liu ZH, Huang RZ. Antisense PEP gene regulates to
768 ratio of and protein and lipid content in *Brassica Napus* seeds. J Agr Biotechnol
769 1999;7(7591):316-320.
770 [26] Cao H, Zhang L, Tan X, Long H, Shockey JM. Identification, classification and
771 differential expression of oleosin genes in tung tree (*Vernicia fordii*). PLoS One
772 2014;9:e88409.
773 [27] Baud S, Wuilleme S, To A, Rochat C, Lepiniec L. Role of WRINKLED1 in the
774 transcriptional regulation of glycolytic and fatty acid biosynthetic genes in
775 Arabidopsis. Plant J 2009;60:933−47.
776 [28] Sugliani M, Rajjou L, Clerkx EJM, Koornneef M, Soppe WJJ. Natural modifiers
777 of seed longevity in the *Arabidopsis* mutants abscisic acid insensitive3-5 (abi3-5)
778 and leafy cotyledon1-3 (lec1-3). New Phytol 2009;184:898−908.
779 [29] Kirkbride RC, Fischer RL, Harada JJ. LEAFY COTYLEDON1, a key regulator
780 of seed development, is expressed in vegetative and sexual propagules of
781 selaginella moellendorffii. PloS One 2013;8:e67971.
782 [30] Rikiishi K, Maekawa M. seed maturation regulators are related to the control of
783 seed dormancy in Wheat (*Triticum aestivum* L.). PloS One 2014;9:e107618.
784 [31] Huang MK, Hu YL, Liu X, Li YG, Hou XL. *Arabidopsis* LEAFY
785 COTYLEDON1 controls cell fate determination during post-embryonic
786 development. Front Plant Sci 2015;6:955.
787 [32] Guan R, Zhao Y, Zhang H, Fan G, Liu X, Zhou W, et al. Draft genome of the
788 living fossil *Ginkgo biloba*. Gigascience 2016;5:49.
789 [33] Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, et al. The
790 Norway spruce genome sequence and conifer genome evolution. Nature
791 2013;497:579−84.
792 [34] Patel RV, Nahal HK, Breit R, Provart NJ. BAR expressolog identification:
793 expression profile similarity ranking of homologous genes in plant species. Plant
794 J 2012;71:1038−50.
795 [35] Winter D, Baxter I, Vinegar B, Nahal H, Ammar R, Wilson GV, et al. An
796 “Electronic fluorescent pictograph” browser for exploring and analyzing

large-scale biological data sets. PLoS One 2007;2:e718.

[36] Kagale S, Nixon J, Khedikar Y, Pasha A, Provart NJ, Clarke WE, et al. The developmental transcriptome atlas of the biofuel crop *Camelina sativa*. Plant J 2016;88:879−94.

[37] Hawkins C, Caruana J, Li JM, Zawora C, Darwish O, Wu J, et al. An eFP browser for visualizing strawberry fruit and flower transcriptomes. Hortic Res 2017;4:17029.

[38] Meyers BC. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. Plant Cell 2003;15:809−34.

[39] Liu J, Liu X, Dai L, Wang G. Recent progress in elucidating the structure, function and evolution of disease resistance genes in plants. J Genet Genomics 2007;34:765−76.

[40] Gassmann W, Hinsch ME, Staskawicz BJ. The *Arabidopsis* RPS4 bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. Plant J 1999;20:265−77.

[41] Dyer JM, Chapital DC, Kuan JCW, Shepherd HS, Tang FQ, Pepperman AB. Production of linolenic acid in yeast cells expressing an omega-3 desaturase from tung (*Aleurites fordii*). J American Oil Chemists' Society 2004;81:647−51.

[42] Pastor S, Sethumadhavan K, Ullah AH, Gidda S, Cao H, Mason C, et al. Molecular properties of the class III subfamily of acyl-coenyzme A binding proteins from tung tree (*Vernicia fordii*). Plant Sci 2013;203-204:79−88.

[43] Hwang YT, Pelitire SM, Henderson MP, Andrews DW, Dyer JM, Mullen RT. Novel targeting signals mediate the sorting of different isoforms of the tail-anchored membrane protein cytochrome b5 to either endoplasmic reticulum or mitochondria. Plant Cell 2004;16:3002−19.

[44] Gidda SK, Shockey JM, Rothstein SJ, Dyer JM, Mullen RT. *Arabidopsis thaliana* GPAT8 and GPAT9 are localized to the ER and possess distinct ER retrieval signals: functional divergence of the dilysine ER retrieval motif in plant cells. Plant Physiol Biochem 2009;47:867−79.

[45] Shepherd HS, Dyer JM, Tang F, Shih DS, Pepperman AB. Nucleotide sequence

827 of a cDNA clone of a plastid-type omega-3 fatty acid desaturase (accession No.
828 AF200717) from tung (*Aleurites fordii*) seeds (PGR 00-009). Plant Physiol
829 2000;122:292.
830 [46] Tang F, Dyer JM, Lax AR, Shih DS, Chapital DC, Pepperman AB. cDNA cloning
831 of aquaporin (accession No. AF047173) and glutaredoxin (accession No.
832 AF047694) from *Aleurites fordii* seeds. Plant Physiol 1998;177:717−20.
833 [47] Long H, Tan XF, Yan F, Zhang L, Li Z, Cao HP. Molecular cloning and
834 expression profile of ß-ketoacyl-acp synthase gene from tung tree (*Vernicia*
835 *fordii* Hemsl.). Genetika 2015;47:143−59.
836 [48] Wendl MC, Barbazuk WB. Extension of Lander-Waterman theory for sequencing
837 filtered DNA libraries. BMC Bioinformatics 2005;6:245.
838 [49] GW V, FJ S, M N, CJ U, H F, J G, et al. GenomeScope: fast reference-free
839 genome profiling from short reads. Bioinformatics 2017;33:2202−4.
840 [50] Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core
841 genes in eukaryotic genomes. Bioinformatics 2007;23:1061−7.
842 [51] Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:
843 assessing genome assembly and annotation completeness with single-copy
844 orthologs. Bioinformatics 2015;31:3210−2.
845 [52] Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with
846 RNA-Seq. Bioinformatics 2009;25:1105−11.
847 [53] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T,
848 Telling A, et al. Comprehensive mapping of long-range interactions reveals
849 folding principles of the human genome. Science 2009;326:289−93.
850 [54] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic
851 genome annotation and pathway reconstruction server. Nucleic Acids Res
852 2007;35:W182−5.
853 [55] Zdobnov EM, Apweiler R. InterProScan--an integration platform for the
854 signature-recognition methods in InterPro. Bioinformatics 2001;17:847−8.
855 [56] Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al.
856 InterPro: the integrative protein signature database. Nucleic Acids Res

2009;37:D211−5.

[57] Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 2007;35:3100−8.

[58] Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1997;25:955−64.

[59] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;24:1586−91.

[60] De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics 2006;22:1269−71.

Figure legends

Figure 1 The genomic landscape of tung tree

The features from outside to inside are pseudochromosomes (**a**), gene density (0−1) (**b**), repeat density (0−1) (**c**), GC content (0%−50%) (**d**), expression (0−1) (**e**), retroelement (0−0.70) (**f**), DNA transposon (0−0.09) (**g**), tandem repeat (0−0.40) (**h**), genome synteny (**i**); Intra-genome collinear blocks with gene pairs numbering more than 20 are highlighted with arcs in the middle of the diagram. Circos was used to construct the diagram. All distributions were drawn using a window size of 1 Mb with the exception of expression, which was drawn using a window of 50 Kb. Chr, chromosome.

Figure 2 Evolution of tung tree genome

A. Phylogenetic tree of tung tree and 7 other plant species based on orthologues of single-copy gene families. The number in parentheses at each branch point represents the divergence time (Mya). The number at the root (15,662) represents the number of gene families in the common ancestor. The value above each branch indicates the number of gene family expansion/contraction at each round of genome duplication after divergence from the common ancestor. Bootstrap value for each node is 100. **B**. Density distribution of 4DTv and Ks for paralogous genes. The peak value is shown in the inset. “non” means no peak value. **C**. Collinear relationship of *V. fordii*, *M. esculenta* and *V. vinifera*. Gray line connects matched gene pairs. The chromosomes

of tung tree, cassava and grapevine were assigned with green, blue and purple, respectively. The annotated genes were clustered into gene families among eight sequenced whole genomes including *A. thaliana, P. trichocarpa, V. vinifera* and five Euphorbiaceae species *i.e. V. fordii, R. communis, M. esculenta, H. brasiliensis, and J. curcas.*

Figure 3 Analysis of the LTR Retrotransposons in the tung tree genome

A. The neighbor-joining tree based on 347 *Ty1/copia* sequences; **B**. The neighbor-joining tree based non 622 *Ty3/gypsy* sequences. **C**. Proportions of LTR Retrotransposon families by copy number in the tung tree genome. **D**. Heat map of expression patterns of 701 LTR Retrotransposons. All aligned sequences correspond to the RT domains without premature termination codon. LTR family names and their proportion are indicated. I, II, and III indicate high-copy families (>= 5 intact members), median-copy families (2–4 intact members) and single-copy families, respectively.

Figure 4 Functional conservation and diversification of tung tree homologs as visualized with the Tung Tree eFP Browser

eFP browser images showing conservation, sub-functionalization, neo-functionalization and non-functionalization of tung tree homologs. In each panel, the expression patterns of three homologs of each gene is shown. In all cases, red represents higher levels of transcript accumulation and yellow represents a lower level of transcript accumulation. From top to bottom, the genes are involved in feruloyl CoA ortho-hydroxylase (from left to right Vf03G0652, Vf00G0634, and Vf03G0623), Protein ECERIFERUM (from left to right Vf04G0546, Vf06G2858, and Vf06G2857), Purple acid phosphatase (from left to right Vf04G0305, Vf04G0306, and Vf11G0977), and Protein LYK5 (from left to right Vf09G1183, Vf03G0089, and Vf09G0959).WAF, week after flowering.

Figure 5 The NBS-encoding genes in tung tree genome

A. The maximum-likelihood phylogenetic tree based on 88 tung tree NBS encoding genes; dots in green, blue, pink, and orange indicate NBS subfamily, NBS-LRR subfamily, CC-NBS subfamily, and CC-NBS-LRR subfamily, respectively. Gene IDs

in red indicate tandem repeats. **B**. Heat map of expression patterns of tung tree NBS-encoding genes. FOE, FOM, and FOL represents early, middle, and late stage after *F. oxysporum* infection. Different colored arrows indicate NBS genes responding to Fusarium wilt.

Figure 6 Network of genes involved in tung oil biosynthesis

A. Tung oil biosynthesis pathway. Tung oil biosynthesis is catalyzed by 18 enzymatic steps with multiple isozymes in each step. Acetyl-CoA is converted into C16 and C18 fatty acids in the plastid. TAG is synthesized in the endoplasmic reticulum and packed in the oil bodies. The metabolites are described in the black box. The enzymes are circled between two metabolite boxes. The expression levels of oil-biosynthesis genes are presented with the heat map. The scale bar of relative expression levels are shown at the top left. **B**. Oil droplet development in tung tree seeds. **C**. Tung oil and fatty acid accumulation profiles. PEPC, phosphoenolpyruvate carboxylase. PK, pyruvate kinase. ACCase, acetyl CoA carboxylase. α/β-CT, acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha/ beta. BCCP, biotin carboxyl carrier protein. BC, biotin carboxylase. MAT, malonyl-CoA transacylases. KAS, ketoacyl-ACP synthase. KAR, ketoacyl-ACP reductase. HAD, hydroxyacyl-ACP dehydrase. EAR, enoyl-ACP reductase. FAT, fatty-acyl carrier protein thioesterase. SAD, stearoyl-ACP desaturase. FA, fatty acid. LACS, long-chain acyl-CoA synthetase. G3P, glycerol-3-phosphate. GPAT, glycerol-3-phosphate acyltransferase. LPA, lysophosphatidic acid. LPAT, lysophosphatidic acid acyltransferase. PA, phosphatidic acid. PP, phosphatidate Phosphatase DAG, diacylglycerol. PDCT, phosphatidylcholine. DAG-CPT, CDP-choline-diacylglycerol cholinephosphotransferase. PC, phosphatidylcholine. FAD, fatty-acid desaturase. DAGT, diacylglycerol O-acyltransferase. PDAT, phospholipid-DAG acyltransferase. LPC, lyso-phosphatidylcholine. TAG, triacylglycerol. Ole, oleosin. WAF, week after flowering.

Figure 7 Phylogeny of FAD2 and FADx proteins

A maximum-likelihood phylogenetic tree constructed from protein sequences. The taxon names in the phylogenetic tree are indicated after gene ID. The clades are marked by four different block colors in the tree. The last one (yellow), a basal angiosperm, *A. trichopoda*, used as an outgroup; the monocot FAD2, eudicot FAD2

and eudicot FADx clades are marked in red, blue and green, respectively.

Figure 8 Co-expression networks of tung tree oil biosynthesis-related genes and transcription factors at the transcriptome level

Oil biosynthesis-related genes are colored in red, and their adjacent transcription factors are colored in black.

Tables

Table 1 Statistics of tung tree genome assembly and annotation

Supplementary material

File S1 Self-pollination and heterozygosity estimation

File S2 Estimation of genome size and heterozygosity

File S3 Repeat sequence analysis

File S4 Transcriptome sequencing, assembly and eFP browser

File S5 Identification and expression of NBS-encoding gene families

File S6 Lipid analysis and electron microscopy observation

File S7 Oil biosynthesis-related gene family identification and phyleogenetic analysis

File S8 Gene co-expression analysis

File S9 Yeast two-hybrid assay of transcription factors

File S10 Whole-genome shotgun sequencing

File S11 Genome assembly and assessment

File S12 Gene prediction and functional annotation

File S13 Evolutionary analysis

Figure S4 The k-mer analysis to estimate the tung tree genome size

Figure S5 Distribution of length (A) and quality (B) of Pacbio raw reads

1014 **Table S2 Sequencing data for 500 kb-library used in genome survey**

1015 **Table S3 Sequencing data for 17-mer analysis**

1016 **Table S4 Statistics of sequencing data for tung tree genome**

1017 **Table S5 Statistics of the tung tree genome assembly**

1018 **Table S6 Sequencing data for Hi-C library**

1019 **Table S7 Mapped reads of Hi-C library to tung tree genome**

1020 **Table S8 Statistics of Hi-C sequencing data**

1021 **Table S9 Efficient coverage of Hi-C data to tung tree genome assembly**

1022 **Table S10 Data of tung tree genome after Hi-C assembly**

1023 **Table S11 Scaffold information after Hi-C assembly**

1024 **Table S12 Assessment for completeness of tung tree genome by CEGMA**

1025 **Table S13 Assessment for completeness of tung tree genome by BUSCO**

1026 **Table S14 Coverage of tung tree genome from male flower unigenes**

1027 **Table S15 Coverage of tung tree genome from female flower unigenes**

1028 **Table S16 Coverage of tung tree genome from seed 1 unigenes**

1029 **Table S17 Coverage bof tung tree genome from seed 2 unigenes**

1030 **Table S18 Coverage of tung tree genome from seed 3 unigenes**

1031 **Table S19 Coverage of tung tree genome assembly from merged seed unigenes**

1032 **Table S20 Transcriptomic reads mapped to tung tree genome**

1033 **Table S21 Comparison of gene modules between tung tree and other species**

1034 **Table S22 The GC content across the tung tree genome**

1035 **Table S23 The GC content in coding sequences in the tung tree genome**

1036 **Table S24 The GC content in intron regions in the tung tree genome**

1037 **Table S25 Assessment for completeness of predicted tung tree genes by BUSCO**

1038 **Table S26 Gene functional annotation of tung tree genome**

1039 **Table S27 Pathways based on KEGG annotation in tung tree genome**

1040 **Table S28 Non-coding RNAs in tung tree genome**

1041 **Table S29 Statistics of gene families of *V. forrdi* and other 7 species**

1042 **Table S30 GO analysis of unique gene families in the tung tree genome**

1043 **Table S31 KEGG analysis of unique gene families in the tung tree genome.**

1044 **Table S32 GO analysis of expanded gene families in tung tree genome**

1045 **Table S33 Swissprot annotation of PSGs in tung tree genome**

1046 **Table S34 GO analysis of PSGs in tung tree genome**

1047 **Table S35 Collinearity analysis in tung tree genome**

1048 **Table S36 Collinearity analysis in the *M.esculenta* genome**

1049 **Table S37 Collinearity analysis in the *V. vinifera* genome**

1050 **Table S40 Summary of repeat sequences in tung tree genome**

1051 **Table S41 Annotation of repeat sequences in tung tree genome**

1052 **Table S51 Four types of functional conservation and diversification of tung tree**

1053 **homologs**

1054 **Table S52 Cross-species comparison of NBS-encoding gene family number**

1055 **Table S53 Total oil-related gene families in tung tree genome**

1056 **Table S54 Cross-species comparison of oil-related gene family number**

1057 **Table S55 Expression quantity (FPKM value) and duplication type of 88**

1058 **important oil genes in tung tree genome**

1059 **Table S56 Colinear gene pairs in poplar (*P. trichocarpa*)**

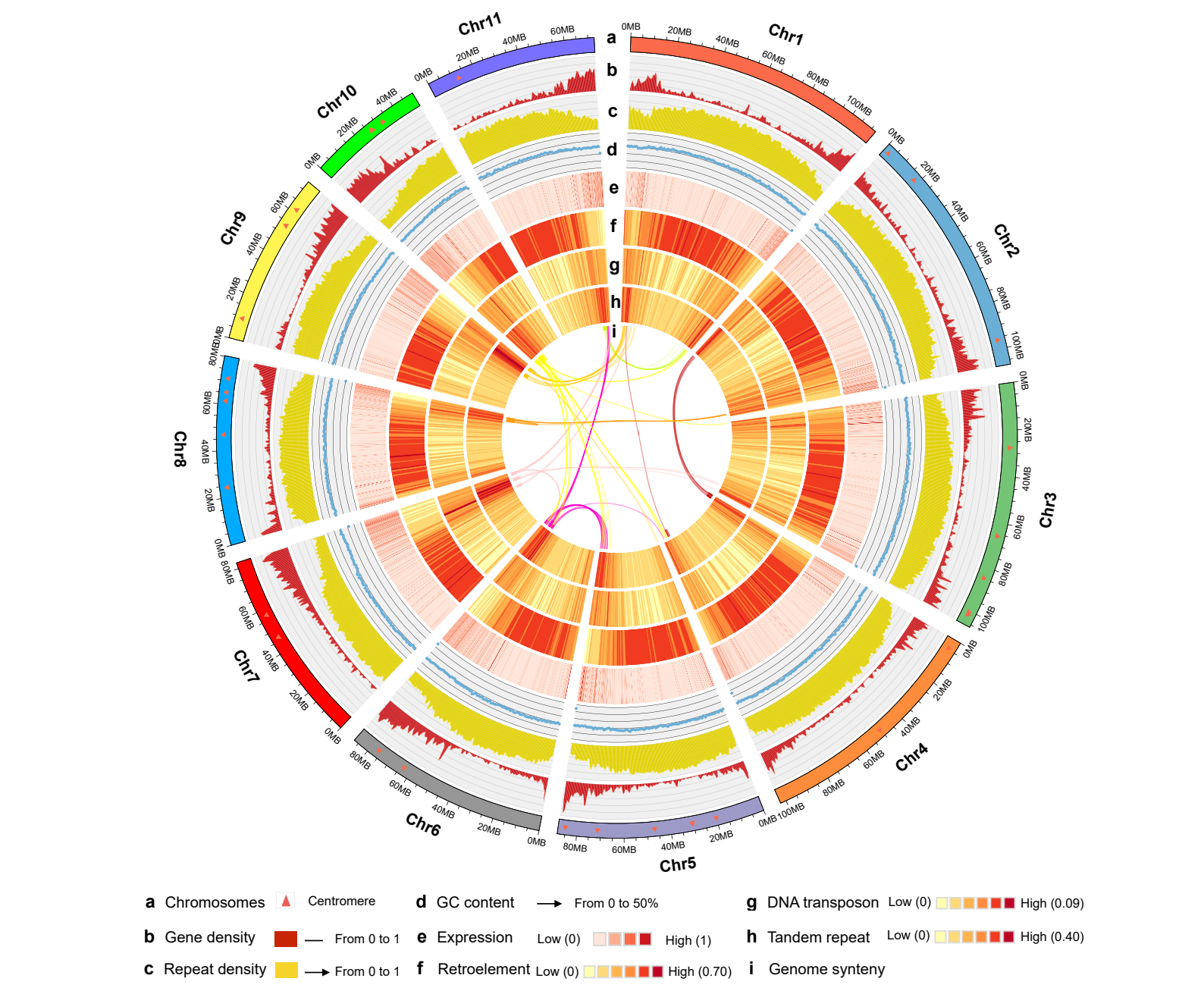
1060 **Table S57 Co-expression relationship among oil-related genes and transcription**

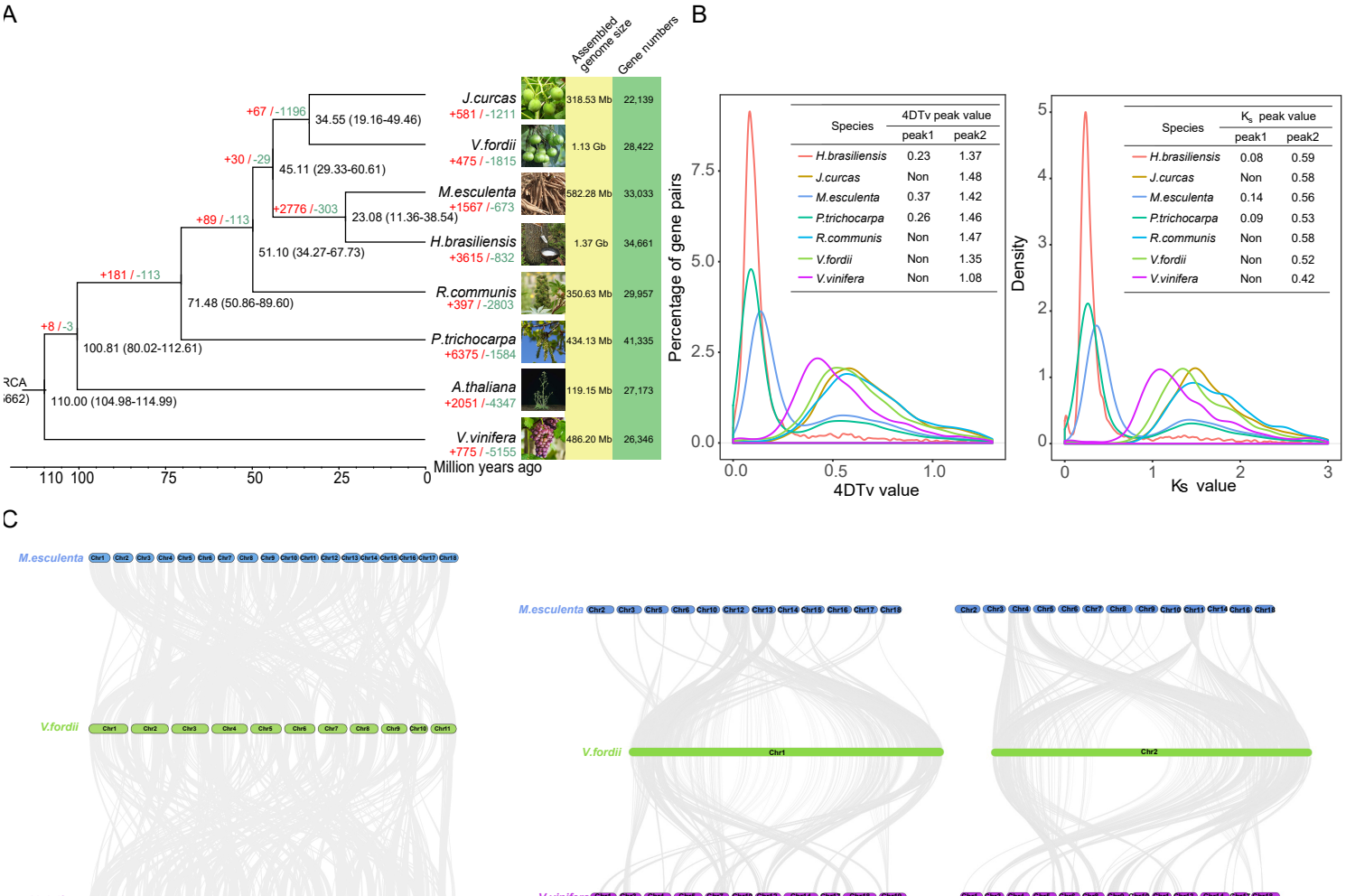
1061 **factors in yellow module**

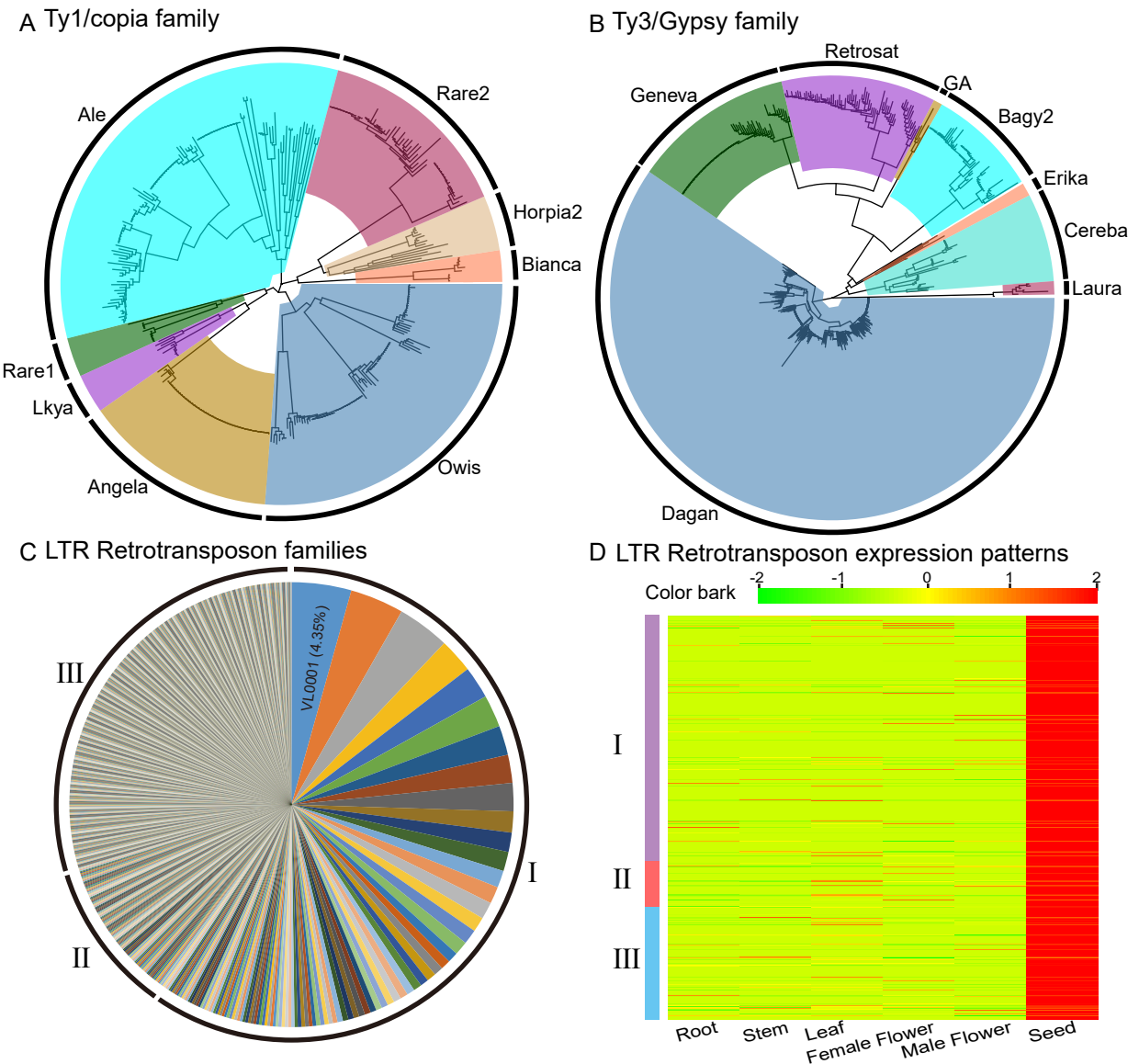
1062 **Table S58 Co-expression relationship among oil-related genes and transcription**

1063 **factors in brown module**

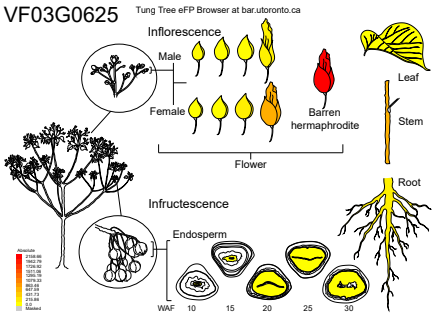
1064

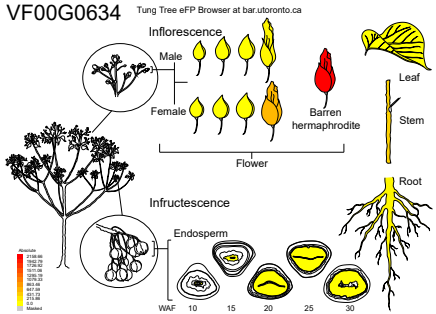

Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
a
b
c
d
e
f
g
h
i
a Chromosomes
Centromere
b Gene density
From 0 to 1
c Repeat density
From 0 to 1
d GC content
From 0 to 50%
e Expression
Low (0)
High (1)
f Retroelement
Low (0)
High (0.70)
g DNA transposon
Low (0)
High (0.09)
h Tandem repeat
Low (0)
High (0.40)
i Genome synteny

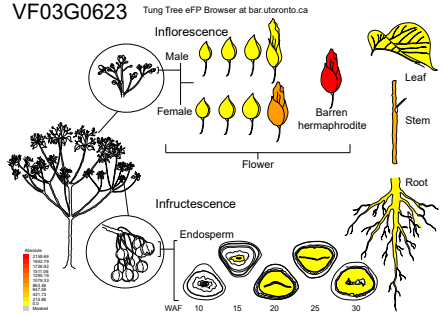

A
Assembled genome size
Gene numbers
J.curcas
+581 / -1211
318.53 Mb
22,139
V.fordii
+475 / -1815
1.13 Gb
28,422
M.esculenta
+1567 / -673
582.28 Mb
33,033
H.brasiliensis
+3615 / -832
1.37 Gb
34,661
R.communis
+397 / -2803
350.63 Mb
29,957
P.trichocarpa
+6375 / -1584
434.13 Mb
41,335
A.thaliana
+2051 / -4347
119.15 Mb
27,173
V.vinifera
+775 / -5155
486.20 Mb
26,346
+67 / -1196
34.55 (19.16-49.46)
+30 / -29
45.11 (29.33-60.61)
+2776 / -303
23.08 (11.36-38.54)
+89 / -113
51.10 (34.27-67.73)
+181 / -113
71.48 (50.86-89.60)
+8 / -3
100.81 (80.02-112.61)
110.00 (104.98-114.99)
110 100 75 50 25 0
Million years ago
B
Species
4DTv peak value
peak1 peak2
H.brasiliensis 0.23 1.37
J.curcas Non 1.48
M.esculenta 0.37 1.42
P.trichocarpa 0.26 1.46
R.communis Non 1.47
V.fordii Non 1.35
V.vinifera Non 1.08
Percentage of gene pairs
4DTv value
Species
Ks peak value
peak1 peak2
H.brasiliensis 0.08 0.59
J.curcas Non 0.58
M.esculenta 0.14 0.56
P.trichocarpa 0.09 0.53
R.communis Non 0.58
V.fordii Non 0.52
V.vinifera Non 0.42
Density
Ks value
C
M.esculenta
V.fordii
V.vinifera


A Ty1/copia family
Ale
Rare2
Horpia2
Bianca
Rare1
Lkya
Angela
Owis
B Ty3/Gypsy family
Retrosat
Geneva
GA
Bagy2
Erika
Cereba
Laura
Dagan
C LTR Retrotransposon families
III
VL0001 (4.35%)
I
II
D LTR Retrotransposon expression patterns
Color bark
-2
-1
0
1
2
I
II
III
Root
Stem
Leaf
Female Flower
Male Flower
Seed

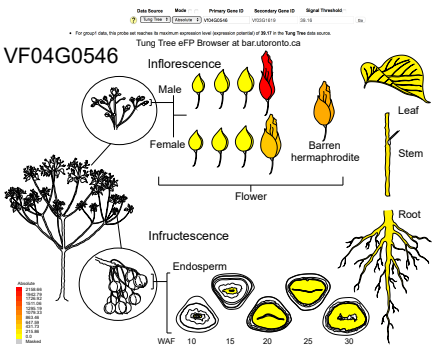
A Conservation of function

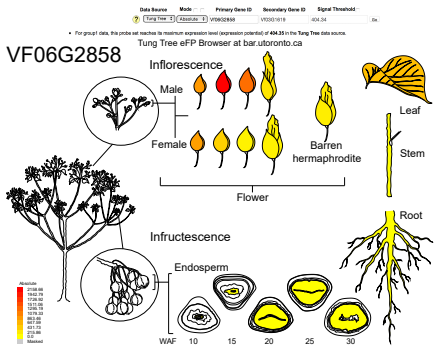


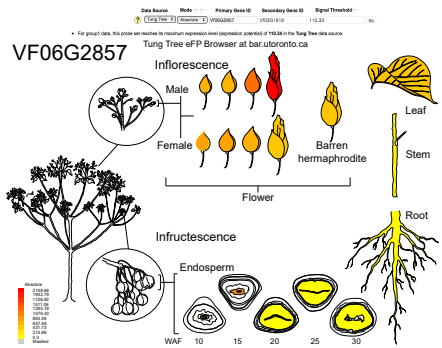




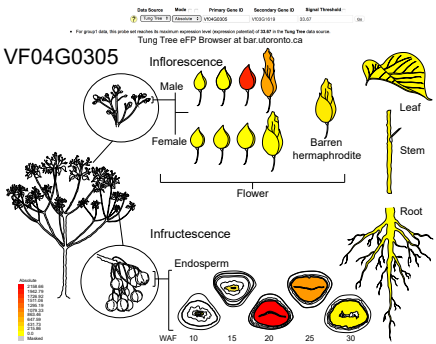
B Sub-functionalization

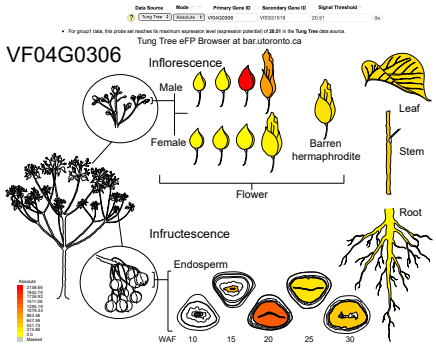


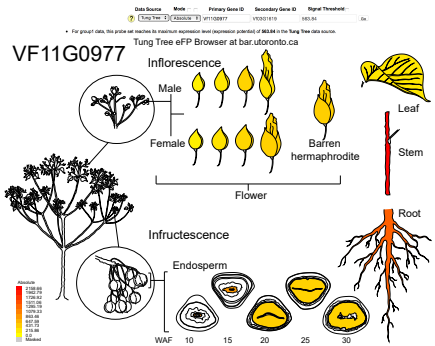




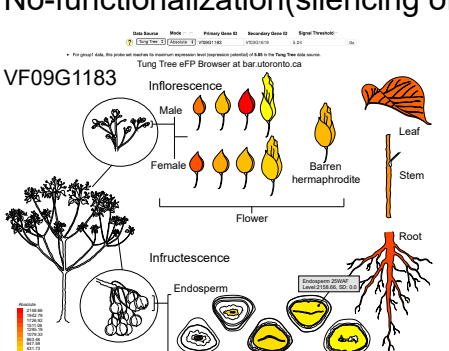
C Sub/neo- functionalization

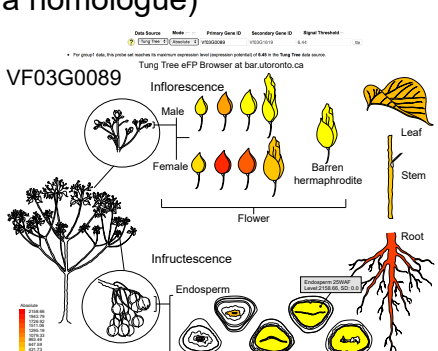


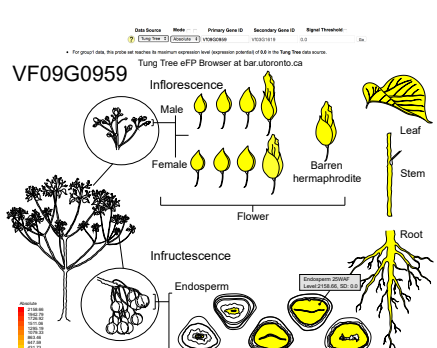


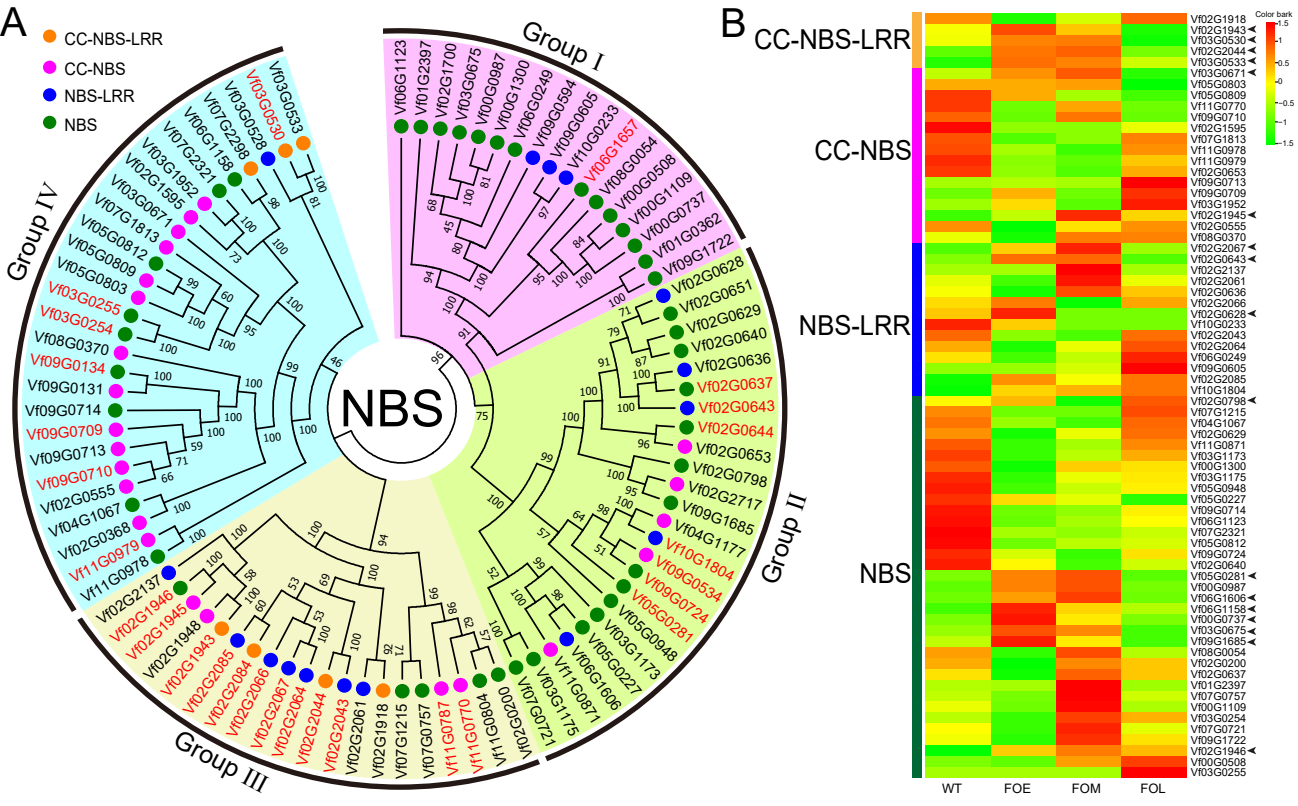


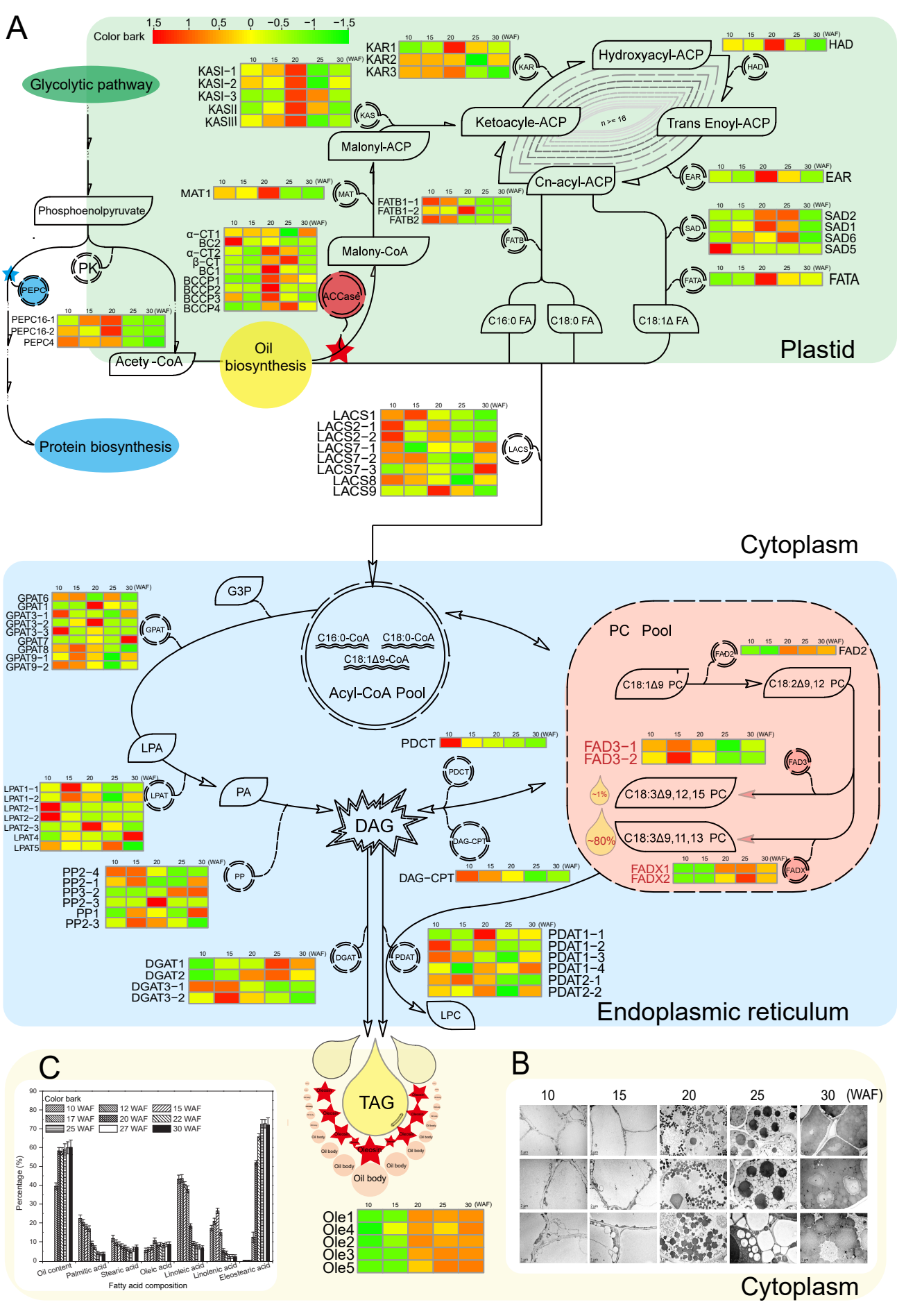
D No-functionalization(silencing of a homologue)

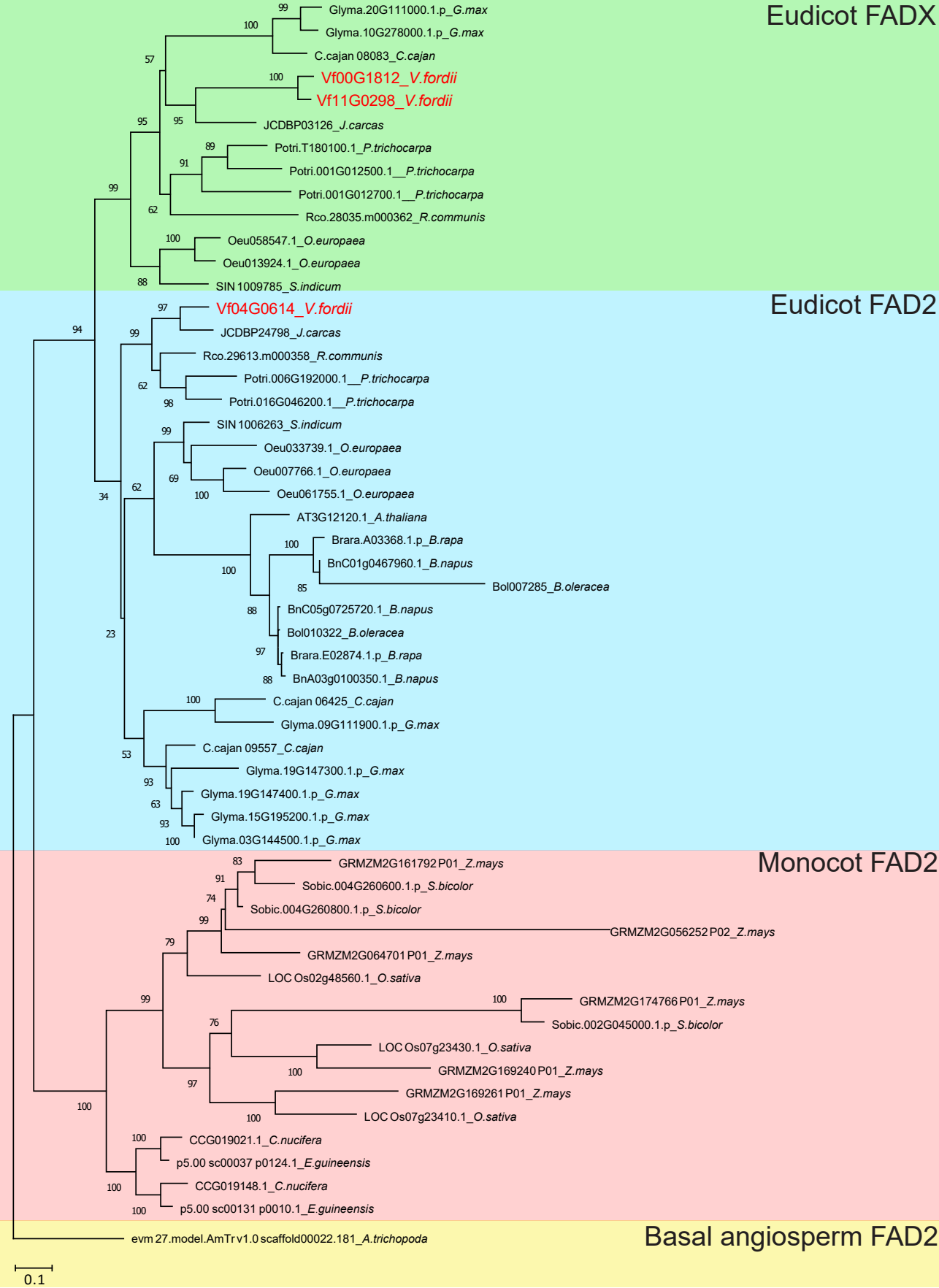


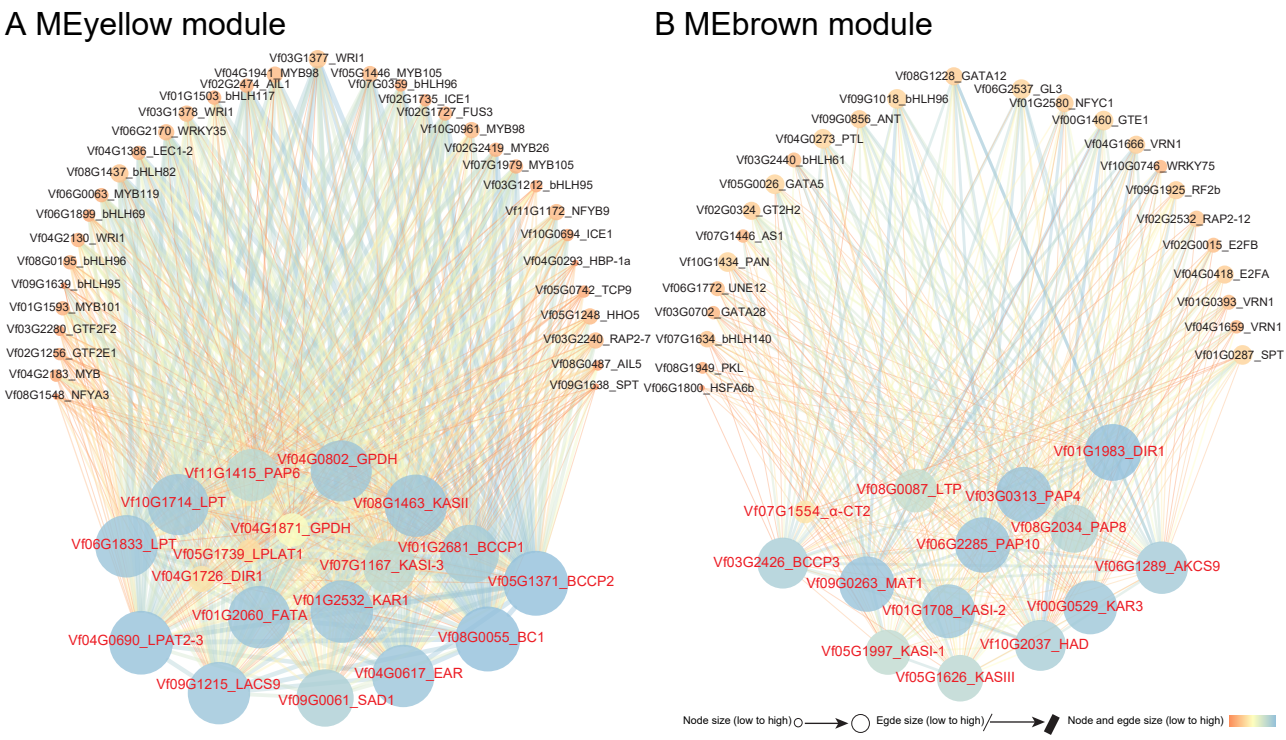





A
CC-NBS-LRR
CC-NBS
NBS-LRR
NBS
Group I
Group II
Group III
Group IV
NBS
Vf06G1123
Vf01G2397
Vf02G1700
Vf03G0675
Vf00G0987
Vf00G1300
Vf06G0249
Vf09G0594
Vf09G0605
Vf10G0233
Vf06G1657
Vf08G0054
Vf00G0508
Vf00G1109
Vf00G0737
Vf01G0362
Vf09G1722
Vf02G0628
Vf02G0651
Vf02G0629
Vf02G0640
Vf02G0636
Vf02G0637
Vf02G0643
Vf02G0644
Vf02G0653
Vf02G0798
Vf02G2717
Vf09G1685
Vf04G1177
Vf10G1804
Vf09G0534
Vf09G0724
Vf05G0281
Vf05G0948
Vf03G1173
Vf05G0227
Vf06G1606
Vf11G0871
Vf03G1175
Vf07G0721
Vf07G0757
Vf11G0804
Vf11G0770
Vf11G0787
Vf02G2200
Vf07G1215
Vf02G1918
Vf02G2061
Vf02G2043
Vf02G2044
Vf02G2064
Vf02G2067
Vf02G2066
Vf02G2084
Vf02G2085
Vf02G1943
Vf02G1948
Vf02G1945
Vf02G1946
Vf02G2137
Vf11G0978
Vf11G0979
Vf02G0555
Vf04G1067
Vf02G0368
Vf09G0710
Vf09G0713
Vf09G0714
Vf09G0131
Vf09G0134
Vf08G0370
Vf03G0254
Vf03G0255
Vf05G0803
Vf05G0809
Vf05G0812
Vf07G1813
Vf03G0671
Vf02G1595
Vf03G1952
Vf07G2321
Vf06G1158
Vf07G2298
Vf03G0528
Vf03G0530
Vf03G0533
B
CC-NBS-LRR
CC-NBS
NBS-LRR
NBS
Color bark
1.5
1
0.5
0
-0.5
-1
-1.5
Vf02G1918
Vf02G1943
Vf03G0530
Vf02G2044
Vf03G0533
Vf03G0671
Vf05G0803
Vf05G0809
Vf11G0770
Vf09G0710
Vf02G1595
Vf07G1813
Vf11G0978
Vf11G0979
Vf02G0653
Vf09G0713
Vf09G0709
Vf03G1952
Vf02G1945
Vf02G0555
Vf08G0370
Vf02G2067
Vf02G0643
Vf02G2137
Vf02G2061
Vf02G0636
Vf02G2066
Vf02G0628
Vf10G0233
Vf02G2043
Vf02G2064
Vf06G0249
Vf09G0605
Vf02G2085
Vf10G1804
Vf02G0798
Vf07G1215
Vf04G1067
Vf02G0629
Vf11G0871
Vf03G1173
Vf00G1300
Vf03G1175
Vf05G0948
Vf05G0227
Vf09G0714
Vf06G1123
Vf07G2321
Vf05G0812
Vf09G0724
Vf02G0640
Vf05G0281
Vf00G0987
Vf06G1606
Vf06G1158
Vf00G0737
Vf03G0675
Vf09G1685
Vf08G0054
Vf02G0200
Vf02G0637
Vf01G2397
Vf07G0757
Vf00G1109
Vf03G0254
Vf07G0721
Vf09G1722
Vf02G1946
Vf00G0508
Vf03G0255
WT
FOE
FOM
FOL


A
Color bark
1.5 1 0.5 0 −0.5 −1 −1.5
Glycolytic pathway
Phosphoenolpyruvate
PK
PEPC
PEPC16-1
PEPC16-2
PEPC4
Acety-CoA
Protein biosynthesis
Oil biosynthesis
KASI−1
KASI−2
KASI−3
KASII
KASIII
KAS
Malonyl-ACP
MAT1
MAT
α−CT1
BC2
α−CT2
β−CT
BC1
BCCP1
BCCP2
BCCP3
BCCP4
ACCase
Malony-CoA
KAR1
KAR2
KAR3
KAR
Ketoacyle-ACP
Hydroxyacyl-ACP
HAD
Trans Enoyl-ACP
n >= 16
Cn-acyl-ACP
EAR
FATB1−1
FATB1−2
FATB2
FATB
C16:0 FA
C18:0 FA
C18:1Δ FA
SAD2
SAD1
SAD6
SAD5
SAD
FATA
Plastid
LACS1
LACS2−1
LACS2−2
LACS7−1
LACS7−2
LACS7−3
LACS8
LACS9
LACS
Cytoplasm
G3P
GPAT6
GPAT1
GPAT3−1
GPAT3−2
GPAT3−3
GPAT7
GPAT8
GPAT9−1
GPAT9−2
GPAT
C16:0-CoA
C18:0-CoA
C18:1Δ9-CoA
Acyl-CoA Pool
LPA
LPAT1−1
LPAT1−2
LPAT2−1
LPAT2−2
LPAT2−3
LPAT4
LPAT5
LPAT
PA
PDCT
DAG
DAG-CPT
PP
PP2−4
PP2−1
PP3−2
PP2−3
PP1
PP2-3
PC Pool
FAD2
C18:1Δ9 PC
C18:2Δ9,12 PC
FAD3−1
FAD3−2
FAD3
~1%
C18:3Δ9,12,15 PC
~80%
C18:3Δ9,11,13 PC
FADX1
FADX2
FADX
DGAT1
DGAT2
DGAT3−1
DGAT3−2
DGAT
PDAT1−1
PDAT1−2
PDAT1−3
PDAT1−4
PDAT2-1
PDAT2-2
PDAT
LPC
Endoplasmic reticulum
B
10 15 20 25 30 (WAF)
Cytoplasm
C
Color bark
10 WAF
12 WAF
15 WAF
17 WAF
20 WAF
22 WAF
25 WAF
27 WAF
30 WAF
Percentage (%)
Oil content
Palmitic acid
Stearic acid
Oleic acid
Linoleic acid
Linolenic acid
Eleostearic acid
Fatty acid composition
TAG
Oleosin
Oil body
Ole1
Ole4
Ole2
Ole3
Ole5


Eudicot FADX
Glyma.20G111000.1.p_G.max
Glyma.10G278000.1.p_G.max
C.cajan 08083_C.cajan
Vf00G1812_V.fordii
Vf11G0298_V.fordii
JCDBP03126_J.carcas
Potri.T180100.1_P.trichocarpa
Potri.001G012500.1__P.trichocarpa
Potri.001G012700.1__P.trichocarpa
Rco.28035.m000362_R.communis
Oeu058547.1_O.europaea
Oeu013924.1_O.europaea
SIN 1009785_S.indicum
Eudicot FAD2
Vf04G0614_V.fordii
JCDBP24798_J.carcas
Rco.29613.m000358_R.communis
Potri.006G192000.1__P.trichocarpa
Potri.016G046200.1__P.trichocarpa
SIN 1006263_S.indicum
Oeu033739.1_O.europaea
Oeu007766.1_O.europaea
Oeu061755.1_O.europaea
AT3G12120.1_A.thaliana
Brara.A03368.1.p_B.rapa
BnC01g0467960.1_B.napus
Bol007285_B.oleracea
BnC05g0725720.1_B.napus
Bol010322_B.oleracea
Brara.E02874.1.p_B.rapa
BnA03g0100350.1_B.napus
C.cajan 06425_C.cajan
Glyma.09G111900.1.p_G.max
C.cajan 09557_C.cajan
Glyma.19G147300.1.p_G.max
Glyma.19G147400.1.p_G.max
Glyma.15G195200.1.p_G.max
Glyma.03G144500.1.p_G.max
Monocot FAD2
GRMZM2G161792 P01_Z.mays
Sobic.004G260600.1.p_S.bicolor
Sobic.004G260800.1.p_S.bicolor
GRMZM2G056252 P02_Z.mays
GRMZM2G064701 P01_Z.mays
LOC Os02g48560.1_O.sativa
GRMZM2G174766 P01_Z.mays
Sobic.002G045000.1.p_S.bicolor
LOC Os07g23430.1_O.sativa
GRMZM2G169240 P01_Z.mays
GRMZM2G169261 P01_Z.mays
LOC Os07g23410.1_O.sativa
CCG019021.1_C.nucifera
p5.00 sc00037 p0124.1_E.guineensis
CCG019148.1_C.nucifera
p5.00 sc00131 p0010.1_E.guineensis
Basal angiosperm FAD2
evm 27.model.AmTr v1.0 scaffold00022.181_A.trichopoda
0.1


A MEyellow module
Vf03G1377_WRI1
Vf04G1941_MYB98
Vf05G1446_MYB105
Vf02G2474_AIL1
Vf07G0359_bHLH96
Vf01G1503_bHLH117
Vf02G1735_ICE1
Vf03G1378_WRI1
Vf02G1727_FUS3
Vf06G2170_WRKY35
Vf10G0961_MYB98
Vf04G1386_LEC1-2
Vf02G2419_MYB26
Vf08G1437_bHLH82
Vf07G1979_MYB105
Vf06G0063_MYB119
Vf03G1212_bHLH95
Vf06G1899_bHLH69
Vf11G1172_NFYB9
Vf04G2130_WRI1
Vf10G0694_ICE1
Vf08G0195_bHLH96
Vf04G0293_HBP-1a
Vf09G1639_bHLH95
Vf05G0742_TCP9
Vf01G1593_MYB101
Vf05G1248_HHO5
Vf03G2280_GTF2F2
Vf03G2240_RAP2-7
Vf02G1256_GTF2E1
Vf08G0487_AIL5
Vf04G2183_MYB
Vf09G1638_SPT
Vf08G1548_NFYA3
Vf04G0802_GPDH
Vf11G1415_PAP6
Vf10G1714_LPT
Vf08G1463_KASII
Vf04G1871_GPDH
Vf06G1833_LPT
Vf01G2681_BCCP1
Vf05G1739_LPLAT1
Vf07G1167_KASI-3
Vf04G1726_DIR1
Vf05G1371_BCCP2
Vf01G2532_KAR1
Vf01G2060_FATA
Vf04G0690_LPAT2-3
Vf08G0055_BC1
Vf04G0617_EAR
Vf09G1215_LACS9
Vf09G0061_SAD1
B MEbrown module
Vf08G1228_GATA12
Vf06G2537_GL3
Vf09G1018_bHLH96
Vf01G2580_NFYC1
Vf09G0856_ANT
Vf00G1460_GTE1
Vf04G0273_PTL
Vf04G1666_VRN1
Vf03G2440_bHLH61
Vf10G0746_WRKY75
Vf05G0026_GATA5
Vf09G1925_RF2b
Vf02G0324_GT2H2
Vf02G2532_RAP2-12
Vf07G1446_AS1
Vf02G0015_E2FB
Vf10G1434_PAN
Vf04G0418_E2FA
Vf06G1772_UNE12
Vf01G0393_VRN1
Vf03G0702_GATA28
Vf04G1659_VRN1
Vf07G1634_bHLH140
Vf01G0287_SPT
Vf08G1949_PKL
Vf06G1800_HSFA6b
Vf01G1983_DIR1
Vf08G0087_LTP
Vf03G0313_PAP4
Vf07G1554_α-CT2
Vf08G2034_PAP8
Vf06G2285_PAP10
Vf03G2426_BCCP3
Vf06G1289_AKCS9
Vf09G0263_MAT1
Vf01G1708_KASI-2
Vf00G0529_KAR3
Vf10G2037_HAD
Vf05G1997_KASI-1
Vf05G1626_KASIII
Node size (low to high)
Egde size (low to high)
Node and egde size (low to high)