Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Full-length Transcriptome and Gene Expression Analysis Reveal Genes and Molecular Elements Expressed During Seed Development in Gnetum Luofuense

Nan deng
Hunan academy of forestry https://orcid.org/0000-0002-7680-3163
Chen Hou
Guandong academy of forestry
Boxiang He
Guangdong Academy of Forestry
Fengfeng Ma
Hunan Academy of Forestry
Qingan Song
Hunan academy of forestry
Shengqing Shi
Chinese academy of forestry
Caixia Liu
Hunan academy of forestry
Yuxin Tian (tianyuxineco@163.com)
https://orcid.org/0000-0003-2066-9599

Research article

Keywords: Gnetales, full-length transcriptome, functional genes, seed, lncRNA

Posted Date: July 6th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-36058/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published on November 23rd, 2020. See the published version at https://doi.org/10.1186/s12870-020-02729-1.

Abstract

Background: *Gnetum* is an economically important tropical and subtropical gymnosperm genus with various dietary, industrial and medicinal uses. Many carbohydrates, proteins and secondary metabolic compounds accumulate during the ripening of *Gnetum* seeds.However, the molecular mechanisms related to this process remain unknown.

Results: We therefore assembled a full-length transcriptome from immature and mature *G. luofuense*seeds using PacBio sequencingreads. We identified a total of 5,726 novel genes, 9,061 alternative splicing events, 3,551 lncRNAs, 2,160 transcription factors, and 359 fusion genes, and we found that 8,512 genes possessed at least one poly(A) site. In addition, gene expression comparisons of six transcriptomes generated by Illumina sequencing showed that14,323 genes were differentially expressed from an immature stage to a mature stage with 7,891 genes upregulated and 6,432 genes downregulated. The expression of 14 differentially expressed transcription factors from the MADS-box, Aux/IAA and bHLH families was validated by qRT-PCR, suggesting that they may have important roles in seed ripening of *G. luofuense*.

Conclusions:These findings provide a valuable molecular resource for domestication and cultivation of *Gnetum* species.

Background

Gnetum in the order Gnetales is a genus of tropical and subtropical gymnosperm trees and shrubs distributed in South America, eastern Africa, and Asia [1–3]. Although molecular phylogenies have revealed the species numbers within the genus [4–6], the phylogenetic position of *Gnetum* within the seed plant lineage has been an important subject of debate in recent years [7–9]. Moreover, *Gnetum* possesses remarkable economic potential for dietary and industrial use: its leaves are used as a vegetable, its stems and bark are made into string, nets and paper, and its seeds are used in oil and drinks. In addition, a variety of chemicals extracted from the vegetative and reproductive organs of *Gnetum*, such as flavonoids and stilbenoids, have been shown to possess antioxidant, anticancer, and antibacterial effects [10–14]. Therefore, exploration of *Gnetum* germplasm resources and investigation of its seed development and germination offer considerable benefits for further domestication and cultivation.

A *Gnetum* seed originates from a female reproductive unit that is produced on the collar involucre of a female strobilus [4, 15, 16]. A *Gnetum* seed is composed of three layers of envelopes, the outermost of which gives rise to a seed coat-aril [16–18]. During seed development in *Gnetum*, the middle envelope is subject to sclerefication [18–20], while the inner envelope produces a micropylar tube that is blocked by closing tissues and is subsequently divided into two pieces [17, 18, 21]. Therefore, the *Gnetum* seed ripening process (during which the aril color changes from green to red, Fig. 1A) is of great importance, because primary (e.g. carbohydrate and protein) and secondary metabolites (e.g. flavonoid and

stilbenoid) are most likely to be synthesized during this time. However, the molecular mechanisms that underlie seed ripening in *Gnetum* have not been carefully investigated.

Previous investigations of transcription factors (TFs) provide valuable insight into the molecular mechanisms of seed development in *Gnetum*. For example, type I and type II MADS-box genes encode essential transcription factors that regulate reproductive organ development in seed plants [22–24]. Previous work has shown that type II MADS-box *AG*-like and *TM8*-like genes are highly expressed in *G. luofuense* seeds [25]. *Aux/IAA* genes also participate in the development of various organs in seed plants by responding to the hormone auxin/indole-3-acetic acid [26–28]. In *Gnetum*, six *Aux/IAA* genes (*GluIAA1-6*) were reported to be involved in female strobilus development [29]. Moreover, bHLH and MYB TFs are able to form a complex that regulates stamen development and seed production [30]. Nevertheless, genes that participate in primary and secondary metabolism during *Gnetum* seed ripening remain poorly understood.

In addition to transcription factors and essential genes, *Gnetum* probably also uses alternative splicing (AS) and alternative polyadenylation (APA) to enrich its transcriptomic complexity during seed development. AS generates multiple proteins from a single coding gene [31, 32], and APA enhances transcriptome complexity by initiating the expression of multiple 3' UTR isoforms from the same gene [33, 34]. Many examples of AS and APA have been documented in angiosperms [35–40], but investigations of AS and APA are rarer in gymnosperms [29, 41, 42]. In addition, lncRNAs, which are defined as possessing at least 200 nucleotides, are essential because they take part in transcriptional and post-transcriptional gene regulation [43–45]. To date, little attention has been paid to lncRNAs in gymnosperms [29, 41, 46].

To investigate AS, APA and lncRNAs, a full-length transcriptome was generated using single-molecule long-read sequencing technology from Pacific Biosciences (PacBio). PacBio sequencing provides better performance than Illumina sequencing because it is challenging to assemble transcripts from different isoforms that share the same exons using only short reads [35, 39, 41]. Compared with short read sequencing, single-molecule transcriptome sequencing provides greater sequence completeness with regard to the 5' and 3' ends of cDNA molecules, higher accuracy for the identification of alternative isoforms, and increased power to distinguish RNA haplotypes [47, 48]. Therefore, in the present study, we generated a full-length transcriptome from two developmental stages (immature and mature) of *G. luofuense* seeds using the reference genome of *G. luofuense* (= *G. montanum*) [8]. AS, APA, lncRNAs and relevant TFs were investigated using the single-molecule data. In addition, we generated separate transcriptomes for the two seed developmental stages using Illumina RNA sequencing to uncover key genes that regulate the seed ripening process in *Gnetum*.

Results

PacBio sequencing and error correction

The full-length transcriptome of mature and immature *G. luofuense* seeds comprised a total of 12,869,707 subreads (19.81 Gb) with an average length of 1,540 bp (Table 1, Fig. 1B). After self-correction with an accuracy value of ROIs > 0.8, 384,042 circular consensus sequences (CCSs) with an average length of 1,919 bp were generated, of which full-length, non-chimeric (FLNC) reads accounted for 81% (312,444, Fig. 1C). The FLNC reads were clustered using the ICE algorithm, and non- FLNC reads were polished. The FLNC reads and polished non- FLNC reads were merged, yielding 165,883 polished consensus isoforms ranging from 167 to 13,816 bp in length (Fig. 1D). The 165,883 polished consensus reads were further corrected using Illumina sequencing data with LoRDEC software. The mean length and N50 and N95 values changed slightly after correction (Table 2).

Table 1
Detail information in the processing of PacBio sequencing data

Terms	Numbers or ratio
Subreads base (G)	19.81
Number of subread	12,869,707
Average length of subread	1,540
N50 of subread	2,013
Number of CCS	384,042
Number of sequences with 5 'terminal primer	362,429
Number of sequences with 3 'terminal primer	362,580
Number of sequences with poly(A) tail	335,541
Number of full length sequence	317,094
Number of full-length non-chimeric reads (Flnc)	312,444
Average length of Flnc read	1,919
Percentage of Flnc (%)	0.81
Number of polished consensus read	165,883
Minimum length of consensus read	167
Maximum length of consensus read	13,816
Average length of consensus read	1,847
N50 of consensus read	2,245

Table 2
Detail information in PacBio sequencing data corrected by Illumina sequencing data

Type	Before correction	After correction
Total nucleotide	306,359,206	307,446,027
Total sequence	165,883	165,883
Mean length	1,847	1,854
Minimum length	167	155
Maximum length	13,816	14,509
N50	2,245	2,254
N90	1,175	1,179

Genome mapping and novel gene detection

The corrected polished consensus reads were mapped to the *G. luofuense* reference genome using GMAP. 162,887 (98.19%) reads were mapped to the reference (Fig. 2A); of these, 63,049 uniquely mapped reads (38.01% of total mapped reads) were mapped to the positive strand of the reference genome, 60,292 uniquely mapped (36.35%) reads were mapped to the negative strand, 39,546 (23.84%) were multiply mapped reads, and 2996 (1.81%) reads were unmapped. Over 98% of the mapped reads showed similarity to the reference genome, and coverage values of the mapped reads were all above 80% (Fig. 2B). A saturation curve revealed that there were sufficient mapped reads to identify most genes on the *G. luofuense* reference genome (Fig. 2C). The PacBio reads had higher percentages of exon numbers considering exon number > 5 (Fig. 2D). After deleting the unmapped and redundant reads, 41,151 reads remained, of which 7,899 were novel isoforms of known genes and 5,726 reads were from novel genes.

Annotation and classification of novel genes

The 5,726 novel genes were annotated by searching against six databases—NCBI NR, KEGG, GO, SwissProt, KOG, and Pfam. A total of 4,099 novel genes were annotated, of which 2,588 were annotated in the NR database (Table 3). Five species—*Picea sitchensis* (649 genes), *Amborella trichopoda* (116), *Vitis vinifera* (88), *Elaeis guineensis* (80), and *Nelumbo nucifera* (61)—produced the largest numbers of hits to the *G. luofuense* novel genes (Fig. 3A). 2,487 novel genes were annotated with KEGG pathways (Table 3), and the most enriched pathways were “signal transduction” (123 genes), “carbohydrate metabolism” (83 genes), and “translation” (69 genes, Fig. 3B). GO analysis classified 2,069 genes into three categories: “biological process”, “cellular components” and “molecular functions” (Fig. 3C). Novel genes classified in the biological process category were mainly annotated with the terms “metabolic process” (1,052), “cellular process” (1,037), and “single-organism process” (581). Novel genes classified in the cellular component category were mainly annotated with the terms “cell” (519), “cell part” (519), and “membrane” (367). Novel genes classified in the molecular function category were mainly annotated

with the terms “binding” (1,192), “catalytic activity” (942), and “transporter activity” (132). 1,930 genes, 1,315 genes and 2,069 genes were annotated with the Swiss Prot, KOG and Pfam databases, respectively (Table 3).

Table 3
Summary of annotated numbers of novel genes by the six databases

Databases	Novel gene numbers
NR	2,588
SwissProt	1,930
KEGG	2,487
KOG	1,315
GO	2,069
Pfam	2,069
All in databases	297
One in databases	2,942
Total annotated genes	4,099

AS, APA analysis and fusion genes

After mapping reads to the reference genome of *G. luofuense*, a total of 9,061 AS events were detected (Fig. 4A). These could be classified into seven types: retained intron (2,713, 29.94%), alternative 3' splice site (2,468, 27.24%), alternative 5' splice site (1,769, 19.52%), skipped exon (1305, 14.40%), alternative first exon (542, 5.98%), alternative last exon (217, 2.39%), and mutually exclusive exon (47, 0.52%) (Fig. 4B). The gene *TnS000292955g01* had the largest number of alternative splicing events: 27. A total of 8,512 genes from *G. luofuense* seeds had at least one supported poly(A) site. Of these, 3,654 (42.93%) had a single poly(A) site, and 640 (7.52%) had at least five poly(A) sites (Fig. 4A and 4C). The largest number of poly(A) sites—21—was found in the gene *TnS000670009g01*. In addition, 2,008 fusion genes were identified in the full-length transcriptome, of which 359 (17.9%) were mapped to scaffold149603 (Fig. 4A).

Identification of TFs and lncRNAs

A total of 2,160 transcription factors (TFs) from 86 gene families were detected using iTAK. The largest fraction of identified TFs came from the C3H (5.6%), bHLH (4.53%), and MYB-related (4.26%) families (Fig. 5A). In addition, 11,885, 5,958, 11,294 and 11,037 lncRNAs were identified using the CNCI, CPC, PFAM and PLEK methods, respectively. A total of 3,551 lncRNAs were identified by all four methods (Fig. 5B), with lengths ranging from 200 to 7,840 bp. The lncRNAs were further classified into four types:

1,422 (40.05%) sense intronic lncRNA, 1,149 (32.36%) long intergenic non-coding RNA, 547 (15.40%) antisense lncRNA, and 433 sense overlapping lncRNA (12.19%) (Fig. 5C). The length distribution of the identified lncRNAs was considerably narrower than that of mRNAs predicted from the *G. luofuense* genome (Fig. 4A and Fig. 5D). Most identified lncRNAs had five or fewer exons, whereas mRNAs predicted from the reference genome tended to have larger numbers of exons (Fig. 5E).

Illumina sequencing of seed samples at two developmental stages

To explore gene expression patterns during seed development of *G. luofuense*, 306,900,384 clean Illumina reads (46.04 Gb of raw data) with Q30 values from 93.54–94.07% were generated from three immature seed samples (IS) and three mature seed samples (MS) (Table 4). After the deletion of adaptors and low-quality reads, the average GC content of the six samples was 47.08%. PCA analysis showed that gene expression was highly correlated among the replicate samples of immature and mature seeds (correlation efficiency value = 0.95, cumulative proportion of variation explained by PC1 and PC2 = 78.7%) (Fig. 6A). After mapping to the *G. luofuense* genome, the mapping ratios of IS (average 89.44%, Table 5) were found to be significantly larger than those of MS (average 84.46%, Student's *t*-test *p*-value = 0.003). RNA-seq analysis of the two developmental stages yielded a total of 23,977 genes (19,010 in IS and 20,737 in MS), of which 2,970 were identified as novel genes.

Table 4
Detail information of Illumina sequenced data from the six *G. luofuense* seed samples

Sample name	Raw reads	Clean reads	Q20 (%)	Q30 (%)	GC content (%)	Genome mapping (%)
IS01	55,631,598	54,243,040	97.77	93.55	47.12	89.75
IS02	49,049,534	47,624,272	97.99	94.07	46.93	88.74
IS03	59,174,882	57,466,054	97.76	93.54	47.56	89.82
MS01	54,240,794	53,255,114	97.86	93.88	46.98	85.32
MS02	52,496,962	50,708,568	97.85	93.78	46.71	84.57
MS03	44,723,634	43,603,336	97.82	93.64	47.18	83.49
Total	315,317,404	306,900,384	97.84	93.74	47.08	86.95

Table 5
Detail information of genome mapping of the transcriptome from the six *G. luofuense* seed samples

Sample name	IS01	IS02	IS03	MS01	MS02	MS03
Total reads	54,243,040	47,624,272	57,466,054	53,255,114	50,708,568	43,603,336
Total mapped	48,642,362 (89.67%)	42,217,664 (88.65%)	51,570,801 (89.74%)	45,412,254 (85.27%)	42,897,372 (84.6%)	36,405,206 (83.49%)
Multiple mapped reads	1,210,358 (2.23%)	1,063,787 (2.23%)	1,296,609 (2.26%)	1,589,381 (2.98%)	1,416,262 (2.79%)	1,337,971 (3.07%)
Uniquely mapped reads	47,432,004 (87.44%)	41,153,877 (86.41%)	50,274,192 (87.49%)	43,822,873 (82.29%)	41,481,110 (81.8%)	35,067,235 (80.42%)
Reads map to '+' strands	23,746,949 (43.78%)	20,608,975 (43.27%)	25,166,379 (43.79%)	21,941,014 (41.2%)	20,771,109 (40.96%)	17,543,656 (40.23%)
Reads map to '-' strands	23,685,055 (43.66%)	20,544,902 (43.14%)	25,107,813 (43.69%)	21,881,859 (41.09%)	20,710,001 (40.84%)	17,523,579 (40.19%)

Enrichment analysis of DEGs and qRT-PCR validation

A total of 14,323 differentially expressed genes (DEGs) were identified between IS (control group) and MS: we found 7,891 upregulated genes and 6,432 genes downregulated (Fig. 6B) from IS to MS. The DEGs were also annotated with the three categories of GO terms, and multiple GO terms in the “biological process” category were significantly enriched with regard to *Z*-scores and adjusted *p*-values (Fig. 6C). The top five enriched GO terms were “single-organism cellular process” (GO:0044763), “single-organism process” (GO:0044699), “metabolic process” (GO:0008152), “cellular metabolic process” (GO:0044237), and “Organic substance metabolic process” (GO:0071704). The DEGs were also enriched in multiple KEGG pathways with reference to *Arabidopsis thaliana*. The top five enriched KEGG pathways were “metabolic pathways” (KEGG ID: ath01100, 1329 genes), “biosynthesis of secondary metabolites” (ath01110, 844 genes), “carbon metabolism” (ath01200, 179 genes), “ribosome” (ath03010, 164 genes), and “starch and sucrose metabolism” (ath00500, 154 genes) (Fig. 6D). qRT-PCR was used to validate the relative expression of 14 genes of interest: four MADS-box genes, four *Aux/IAA* genes, four *bHLH* genes, and two *MYB* genes. The relative expression of the 14 genes at the two seed developmental stages is presented in Fig. 6E.

Discussion

Structural analysis of the full-length transcriptome

AS and APA analysis A total of 8,512 genes were identified as having at least one poly(A) site, suggesting that alternative polyadenylation may enrich the proteomic complexity and affect the seed ripening process of *G. luofuense*. Previous studies have suggested that alternative polyadenylation affects the

development of *G. luofuense* leaves and female strobili [29, 41]. In addition, we found considerable differences in the number of AS events among different gymnosperm species and among different organs in the same species by comparing AS events in the leaves of *G. biloba* (12,209) [42] and the leaves (12,998) [41], female strobili (10,454) [29], and seeds of *G. luofuense* (9,061, the present study). It is noteworthy that intron retention constituted the largest proportion of identified AS events in this study. This finding is likely to occur in both angiosperms and gymnosperms [29].

lncRNA analysis lncRNAs tended to be shorter and possessed fewer exons than protein coding genes [46, 49], a finding that is consistent with previous studies in gymnosperms, such as *G. biloba* [46], *Picea abies* [49], and *G. luofuense* [41]. There were pronounced differences in the number of lncRNAs identified in *G. luofuense* leaves (1662) [41], female strobili (1992) [29] and seeds (3551, the present study), suggesting that lncRNA numbers are tissue-specific in *Gnetum*, similar to the scenario in *P. abies* [49].

Metabolism and secondary metabolites

The DEGs between immature and mature seeds were enriched in several KEGG pathways, e.g. carbon metabolism (179 genes), starch and sucrose metabolism (154 genes), glycolysis/gluconeogenesis (99 genes), and fructose and mannose metabolism (50 genes). The DEGs were also enriched in protein and amino sugar biosynthesis pathways, e.g. protein processing in endoplasmic reticulum (132 genes), amino sugar and nucleotide sugar metabolism (100 genes), and peroxisome (100 genes). The accumulation of carbohydrates and proteins in *Gnetum* seeds makes them palatable and nutritious, thereby attracting a variety of herbivores to promote seed dispersal [50–53]. These compounds in *Gnetum* seeds might have also become important food and oil resources [2, 3].

It is also notable that a large proportion of DEGs were enriched in the terms metabolic pathway (1329 genes), biosynthesis of secondary metabolites (844 genes), fatty acid metabolism (49 genes), and terpenoid backbone biosynthesis (48 genes). Moreover, the DEGs were annotated with several GO terms, e.g. cellular metabolic process (GO:0044237), organic substance metabolic process (GO:0071704), and small molecule metabolic process (GO:0044281). A previous study has shown that *G. parvifolium* seeds are rich in secondary metabolic products, particularly flavonoids and stilbenoids, and that the concentrations of these two compounds differ considerably between immature and mature seeds [10]. Moreover, resveratrol derivative components such as gnemonoside C and gnetin C were extracted from *G. gnemon* seeds, and these compounds were shown to possess antiangiogenic effects [54]. This is probably the reason why *Gnetum* has antioxidant, anticancer, and antibacterial medicinal properties [10, 11].

Various transcription factors and their functions

MADS-box genes Type I MADS-box genes are involved in the development of female gametophytes, embryos and seeds [55, 56]. However, the functions of type I MADS-box genes have been constantly neglected in gymnosperms, probably because type I MADS genes are considerably less numerous in gymnosperms than in angiosperms [25, 57]. To date, 11 type I MADS-box genes have been identified in *G. luofuense*[25], but their functions have not been investigated. In the present study, 27 type I MADS-box TFs

were identified in *G. luofuense* seeds, and gene *TnS000803113g11* was differentially expressed between immature and mature seeds, indicating that type I MADS-box genes may be involved in the seed ripening of *G. luofuense* but it requires further validation.

Compared to type I MADS-box genes, the functions of type II MADS-box genes in *Gnetum* reproductive organ development have received much more attention [23, 25, 57–61]. Thus far, 38 type II genes have been identified in *G. luofuense*, of which *TM8*-like genes constitute almost half of the identified gene numbers [25]. In mature seeds of *G. luofuense*, the type II *AG*-like MADS-box gene *TnS000064931g01* was strongly expressed, the *AGL6*-like gene *TnS000229425g02* was weakly expressed, and two *TM8*-like genes *TnS000061251g01* and *TnS000980857g01* were moderately expressed [25]. In the present study, four type II MADS-box genes were differentially expressed, suggesting that these type II MADS-box genes play an essential role in the regulation of seed ripening in *G. luofuense*. A previous study corroborates our results, showing that *AG*, *AGL6* and *TM8*-like genes regulate seed development of *Ginkgo biloba* and *Taxus baccata* [62].

Aux/IAA genes A previous study in angiosperms reported that *Aux/IAA* genes (i.e. *FaAux/IAA1* and *FaAux/IAA2*) were able to regulate the early developmental stages of strawberry fruits [63]. Another study showed that *IAA9* was involved in fruit and leaf morphogenesis in tomato [64]. By contrast, the functions of *Aux/IAA* genes remain poorly understood in gymnosperms, let alone in the Gnetales. A previous study identified six *Aux/IAA* genes in *G. luofuense* (*GluIAA1-6*) and examined their structure and phylogenetic relationships [29]. In the present study, we identified 30 Aux/IAA TFs in the full-length transcriptome of *G. luofuense* seeds. Four *Aux/IAA* genes, *TnS000653177g04* (*GluIAA2*), *TnS000867017g28* (*GluIAA3*), *TnS000053353g02* (*GluIAA4*), and *TnS000142615g19* (*GluIAA5*), were differentially expressed between the two developmental stages of *G. luofuense* seeds. These results suggest that *Aux/IAA* genes may also be of importance in *Gnetum* seed ripening.

bHLH and MYBs In angiosperms, *bHLH* genes have been shown to participate in seed coat color formation in *Brassica rapa* [65]. The expression of two *MYB*-related genes, i.e. *Osmyb1* and *Osmyb4*, reaches the level of saturation at 14 days after the anthesis, suggesting that they have an important role in the maturation of rice seeds [66]. Furthermore, a MYB–bHLH–WDR complex in angiosperms initiates flavonoid biosynthesis in response to environmental changes [67]. In gymnosperms, bHLH and MYB TFs have been reported to participate in flavonoid biosynthesis in the roots rather than the seeds of *Ginkgo biloba* [42]. Besides, bHLH and MYB TFs were both reported to be involved in the development of female strobili and leaves of *G. luofuense* [29, 41]. In the present study, 98 bHLH and 92 MYB TFs were identified in the full-length transcriptome of *G. luofuense* seeds, suggesting that bHLH and MYB TFs may play a role in color formation and seed ripening in *G. luofuense*.

Conclusions

Two development stage of *Gnetum* seed were used for analyzing molecular mechanisms throughput Next Generation Sequencing (NGS) and full-length transcriptome techniques. Novel genes, alternative

splicing events, lncRNAs and transcription factors of *Gnetum* seed were identified. The results indicated that transcription factors from the MADS-box, Aux/IAA and bHLH families may have important roles in seed ripening of *G. luofuense*. Our results improve our knowledge of Gnetum seed ripening mechanisms and laid solid foundation for domestication and cultivation of *Gnetum* species.

Methods

Plant material and RNA extraction

Gnetum luofuense seeds were collected at immature (IS) and mature (MS) developmental stages from a female individual (voucher number “CH003”, SYS) cultivated in the Bamboo Garden at Sun Yat-sen University on September 2nd and 28th 2018 (Fig. 1A) with the permissions of Sun Yat-sen University.. To obtain a full-length transcriptome for the two developmental stages, identical amounts (15 g) of mature and immature seeds with arils were pooled, incubated in liquid nitrogen, and frozen at − 20 ℃ for PacBio SMRT sequencing. In addition, six samples of *G. luofuense* seeds (“IS001–003” and “MS001–003”) were collected for Illumina sequencing, three from the immature stage (control group) and three from the mature stage. The RNA for each sample was extracted using an RNA kit (Qiagen, Valencia, CA, USA) following the manufacturer’s instructions. RNase-free DNase (Qiagen) was used to remove relic DNA, and the RNA concentration of samples was evaluated by 1% agarose gel electrophoresis. A NanoDrop spectrophotometer (ThermoFisher Scientific, Wilmington, DE, USA) and Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) were used to assess the purity and integrity of the extracted RNA. *G. luofuense* used in this research is wild plant resource and transplanted for teaching. The collection of seeds and the performance of experimental research on such plant were complied with the national guidelines of China.

Library construction and PacBio Sequel sequencing

When the integrity of extracted RNA met the minimum requirement (> 7.0), full-length cDNA was synthesized using a SMARTer PCR cDNA Synthesis kit (Clontech, Takara Bio Inc., Shiga, Japan). The synthesized cDNA was subjected to PCR amplification using a KAPA HIFI PCR kit (Kapa Biosystems, Boston, MA, USA). After PCR amplification, the cDNA was quality controlled and purified using a QIAquick PCR Purification kit (Qiagen, Hilden, Germany). The RNA samples were subjected to terminal repair and the attachment of SMRT dumbbell-type adapters. Before PacBio sequencing, two bins (1–4 kb, 4–6 kb) were established to preferentially sequence the smaller cDNAs. PacBio sequencing data from the merged seed sample were deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA622631.

Library construction and Illumina sequencing

Before Illumina sequencing, all six RNA samples that possessed poly(A) were enriched with oligo(dT) magnetic beads. The enriched RNA was randomly reduced to small pieces with a fragmentation buffer. First strand cDNA was generated using hexamers and reverse transcriptase (Superscript III, Invitrogen).

After purification with AMPure XP beads, second strand cDNA was synthesized using DNA polymerase I, RNase H and dNTPs (Sigma-Aldrich). The double-stranded cDNA was subjected to terminal repair and poly(A) tailing, followed by Illumina adaptor ligation. The final cDNA library was completed after a second round of purification and PCR amplification. The quality of the six cDNA libraries was assessed using a Qubit 2.0 fluorometer prior to sequencing on the Illumina HiSeq 4000 platform. RNA-seq data from the six samples were deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA622631.

PacBio data processing and error correction

PacBio sequencing data were analyzed using PacBio SMRTlink v. 5.1 software. First, we obtained reads of inserts (ROIs) from the BAM files generated from the platform using the following parameters: maximum drop fraction—0.8, minimum length—200, no polish, minimum *z*-score—9999, minimum passes —1, minimum predicted accuracy—0.8, and maximum length—18,000. The ROIs were classified into full-length reads (FLs) and non-full-length reads (nFLs) based on the presence and absence of 5' and 3' cDNA primers and a 3' poly (A) tail, see also in [29]. The FLs and nFLs were clustered to achieve consensus isoforms using an isoform-level clustering (ICE) algorithm. To obtain full-length non-chimeric (FLNC) isoforms, the high-quality isoforms from FLs were corrected using Quiver software with a post-correction accuracy above 99%. The low-quality consensus isoforms from nFLs were further corrected with LoRDEC [68] using two Illumina-sequenced samples (one from mature seeds and one from immature seeds).

Genome mapping and novel gene detection

All FLNCs and corrected nFLs were mapped to the reference genome of *G. luofuense* (= *G. montanum*) [8] using GMAP [69]. The GMAP output files were used for subsequent analyses. Redundant FLNCs were removed using the following parameters: minimum identity-0.9, minimum trimmed coverage-0.85, and allow close indel-0. Mapped FLNCs with different lengths at their 5' ends were not considered to be redundant. The FLNCs that mapped to annotated genes in the *G. luofuense* genome were considered to be known genes; otherwise, they were classified as novel genes and novel isoforms of known genes.

Functional annotation and classification

All identified novel genes were annotated by BLASTX v.2.2.26 searches (E-value $< 1 \times 10^{-5}$) of the gene ontology (GO, http://www.geneontology.org), Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/), Protein Family (Pfam), KOG/COG (Clusters of Orthologous Groups of proteins, http://www.ncbi.nlm.nih.gov/COG/), NCBI non-redundant protein sequence (NR, http//www.ncbi.nlm.nih.gov/), and Swiss-Prot (http// http://www.expasy.org/sprot/) databases and by HMMER v.3.1b2 searches (E-value $< 1 \times 10^{-10}$) of the Pfam (Protein Family, http://pfam.xfam.org/) database [70, 71]. In addition, GO enrichment analysis was performed using the GOseq package implemented in R [72] and KEGG enrichment analysis was performed using KOBAS version 2.0 [73].

AS, APA and fusion genes

Gene structure analysis was performed using the TAPIS pipeline [35]. First, seven types of alternative splicing (AS) event were identified: alternative 3' splice site, retained introns, alternative 5' splice site, skipped exon, alternative first exon, alternative last exon, and mutually exclusive exons. Second, alternative polyadenylation (APA) analysis was conducted, and genes were classified according to their poly(A) number. Third, fusion genes that derived from two or more genes distantly located on the *G. luofuense* genome were identified and further validated where at least two Illumina transcripts were mapped.

Identification of TFs and lncRNA

Coding sequences (CDS), which possess open reading frames (ORFs), were identified by searching against the Pfam database using TransDecoder [74]. Based on the identified CDS, transcription factors (TFs) were predicted by searching against the Plant Transcription Factor Database v.4.0 (http://planttfdb.cbi.pku.edu.cn) using iTAK version 15.03 [75]. In addition, four methods were used to identify lncRNAs: PC (Coding Potential Calculator) [76], CNCI (Coding-Non-Coding Index) [77], CPAT (Coding Potential Assessment Tool) [78], and Pfam. The lncRNAs, which are longer than 200 nt and possess at least two exons, do not encode proteins and are classified into four groups: lincRNA, sense intronic lncRNA, sense overlapping lncRNA, and antisense lncRNA.

Identification of DEGs and qRT-PCR validation

Illumina sequenced raw reads with poly(N) and low scores were removed, and the remaining reads were trimmed of adaptors at both ends. The cleaned reads were mapped to the *G. luofuense* genome using HISAT2 v.2.1.0 [79]. Mapped read numbers were counted and adjusted through one scaling normalized factor using the R package edgeR [80]. The numbers of mapped reads were converted to values of fragments per kilobase of transcript per million mapped fragments (FPKM). To identify differentially expressed genes (DEGs), RNA data from three replicate samples of mature and immature seeds were separately merged and then compared using the R package EBSeq v. 1.20.0 [81]. The DEGs met the following requirements: corrected *P*-value (adjusted by the Benjamini & Hochberg method)—0.005 and $\log_2$(fold change) value-1.

A total of 14 DEGs were selected for gene expression validation with qRT-PCR. The primer design for these target genes was performed using Primer Premier 5 [82], and primer sequences are provided in Table S1. Two micrograms of RNA were extracted from mature and immature seeds of *G. luofuens*e and subjected to cDNA synthesis according to the manufacturer's protocol. qRT-PCR was performed under the following conditions: 10 min at 95 ℃ (1 cycle), 10 s at 95 ℃, 30 s at 55 ℃ and 15 s at 72 ℃ (40 cycles), temperature reduction from 95 ℃ to 60 ℃ (0.5 ℃/10 s) and termination in 30 s at 25 ℃. The *G. luofuense* actin gene was used as an endogenous control to estimate the relative expression of target genes using the ΔΔCt-method [83]. For each sample, three replicates were performed, and the mean and standard deviation of the qRT-PCR gene expression values were calculated.

Abbreviations

DEGs
differentially expressed genes; CDS:coding sequences; ORFs:open reading frames; TFs:transcription factors; PC:coding potential calculator; CNCI:coding-non-coding index; CPAT:coding potential assessment tool; APA:alternative polyadenylation; FLs:full-length reads; nFLs:non-full-length reads; Aux/IAA:auxin/indole-3-acetic acid; bHLH:Basic helix-loop-helix.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The data that support the results are included within the article and its additional files. Other relevant materials are available from the corresponding authors on reasonable request.

Competing interests

The authors declare no competing financial interest.

Funding

This work was supported by the Special Fund for Hunan Forestry Science and Technology (XLK201812) and Forestry Science and Technology Innovation Project of Hunan Province (HNGYL-2019-01).

Authors' contributions

ND and CH conceived and designed the experiments and wrote the manuscript. BH, SS and SQ performed the experiments. CL and YT analyzed the data and wrote the manuscript. All authors have read and approved this manuscript

Acknowledgement

Not Applicable.

References

1. Markgraf F. Monographie der Gattung *Gnetum* Ser. 3. *Bulletin du Jardin Botanique de Buitzenzorg* **10**: 407–511 (1930).

2. Markgraf F. Gnetaceae. In: Steenis CGGJ, editor. Flora Malesiana Ser. 1 Vol 4, Vol. 4. Djakarta: Noordhoff-Kolff: Batavia 1951, pp. 336–47.
3. Kubitzki K. Gnetaceae. In: Kramer KU, Green PS, editors. *The families and genera of vascular plants*. Springer: Berlin, Heidelberg, Germany, 1990, pp 383–386.
4. Hou C, Humphreys AM, Thureborn O, Rydin C. New insights into the evolutionary history of *Gnetum*. (Gnetales) *Taxon*. 2015;64(2):239–53.
5. Kim JH, Won H. Identification of Cambodian Gnetum (Gnetaceae, Gnetales) species by DNA barcoding. Korean Journal of Plant Taxonomy. 2016;46(2):163–74.
6. Won H, Renner SS. Dating dispersal and radiation in the gymnosperm *Gnetum* (Gnetales) - clock calibration when outgroup relationships are uncertain. Syst Biol. 2006;55(4):610–22.
7. Ran JH, Shen TT, Wang MM, Wang XQ. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. *P Roy Soc B-Biol Sci* Jun. 2018;27(1881)):20181012.. ; **285**.
8. Wan T, Liu ZM, Li LF, et al. A genome for gnetophytes and early evolution of seed plants. *Nat Plants* Feb;. 2018;4(2):82–9.
9. Ickert-Bond SM, Renner SS. The Gnetales: recent insights on their morphology, reproductive biology, chromosome numbers, biogeography, and divergence times. J Syst Evol. 2015;54(1):1–16.
10. Deng N, Chang E, Li M, et al. Transcriptome characterization of *Gnetum parvifolium* reveals candidate genes involved in important secondary metabolic pathways of flavonoids and stilbenoids. *Front Plant Sci*;. 2016;7:174.
11. Deng N, Liu C, Chang E, et al. High temperature and UV-C treatments affect stilbenoid accumulation and related gene expression levels in *Gnetum parvifolium*. *Electronic Journal of Biotechnology*;. 2017;25:43–9.
12. Ali F, Assanta MA, Robert C. *Gnetum africanum*: a wild food plant from the african forest with many nutritional and medicinal properties. J Med Food. 2011;14(11):1289–97.
13. Seo C, Lym SH, Jeong W, et al. Flavonoids, stilbenoids, and phenolic derivatives from the stems of *Gnetum macrostachyum* (Gnetaceae). Biochem Syst Ecol. 2020;90:104033.
14. Saisin S, Tip-pyang S, Phuwapraisirisan P. A new antioxidant flavonoid from the lianas of *Gnetum macrostachyum*. Nat Prod Res. 2009;23(16):1472–7.
15. Hou C, Wikström N, Strijk J, Rydin C. Resolving phylogenetic relationships and species delimitations in closely related gymnosperms using high-throughput NGS, Sanger sequencing and morphology. Plant Syst Evol. 2016;302(9):1345–65.
16. Takaso T, Bouman F. Ovule and seed ontogeny in *Gnetum gnemon* L. J Plant Res. 1986;99(3):241–66.
17. Berridge EM. On some points of resemblance between gnetalean and Bennettitean seeds. New Phytol. 1911;10(4):140–4.

18. Thoday MG. The female inflorescence and ovules of *Gnetum africanum* with notes on *Gnetum scandens*. Ann Bot. 1911;25(4):1101–35.
19. Thoday MG. Anatomy of the ovule and seed in *Gnetum gnemon*, with notes on *Gnetum funiculare*. Ann Bot. 1921;35(1):37–53.
20. Rodin RJ, Kapil RN. Comparative anatomy of the seed coats of *Gnetum* and their probable evolution. Amer J Bot. 1969;56(4):420–31.
21. Berridge EM. The structure of the female strobilus in *Gnetum gnemon*. Ann Bot. 1912;26(4):987–92.
22. Gramzow L, Theißen G. Phylogenomics of MADS-box genes in plants-two opposing life styles in one gene family. Biology. 2013;2(3):1150–64.
23. Becker A, Winter KU, Meyer B, Saedler H, Theissen G. MADS-box gene diversity in seed plants 300 million years ago. Molec Biol Evol Oct. 2000;17(10):1425–34.
24. Melzer R, Wang YQ, Theissen G. The naked and the dead: The ABCs of gymnosperm reproduction and the origin of the angiosperm flower. Semin Cell Dev Biol Feb. 2010;21(1):118–28.
25. Hou C, Li L, Liu Z, Su Y, Wan T. Diversity and expression patterns of MADS-box genes in *Gnetum luofuense*—implications for functional diversity and evolution. Tropical Plant Biology. 2019;1:1–14.
26. Guilfoyle TJ. *Aux/IAA* proteins and auxin signal transduction. Trends Plant Sci. 1998;3(6):205–7.
27. Luo J, Zhou J-J, Zhang J-Z. *Aux/IAA* gene family in plants: molecular structure, regulation, and function. Int J Mol Sci. 2018;19(1):259.
28. Wu WT, Liu YX, Wang YQ, et al. Evolution analysis of the *Aux/IAA* gene family in plants shows dual origins and variable nuclear localization signals. *Int J Mol Sci* Oct;. 2017;18(10):107.
29. Hou C, Deng N, Su YJF. PacBio long-read sequencing reveals the transcriptomic complexity and *Aux/IAA* gene evolution in *Gnetum* (Gnetales). *Forests*; **10**(11): 1043 (2019).
30. Qi T, Huang H, Song S, Xie D. Regulation of jasmonate-mediated stamen development and seed production by a bHLH-MYB complex in *Arabidopsis*. Plant Cell. 2015;27(6):1620–33.
31. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. Genome Res. 2012;22(6):1184–95.
32. Reddy AS, Marquez Y, Kalyna M, Barta A. Complexity of the alternative splicing landscape in plants. Plant Cell. 2013;25(10):3657–83.
33. Gupta I, Clauder-Münster S, Klaus B, et al. Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA–protein interactions. Mol Syst Biol. 2014;10(2):719.
34. Blazie SM, Geissel HC, Wilky H, Joshi R, Newbern JM, Mangone M. Alternative polyadenylation directs tissue-specific miRNA targeting in *Caenorhabditis elegans* somatic tissues. Genetics. 2017;206(2):757–74.
35. Abdel-Ghany SE, Hamilton M, Jacobi JL, et al. A survey of the sorghum transcriptome using single-molecule long reads. *Nat Commun*;. 2016;7:11706.
36. Wang T, Wang H, Cai D, et al. Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (*Phyllostachys edulis*). Plant J. 2017;91(4):684–99.

37. Liu XX, Mei WB, Soltis PS, Soltis DE, Barbazuk WB. Detecting alternatively spliced transcript isoforms from single-molecule long-read sequences without a reference genome. Molec Ecol Resour Nov. 2017;17(6):1243–56.
38. Chao Q, Gao ZF, Zhang D, et al. The developmental dynamics of the Populus stem transcriptome. J Pl Biotech. 2019;17(1):206–19.
39. Chao Y, Yuan J, Li S, Jia S, Han L, Xu L. Analysis of transcripts and splice isoforms in red clover (*Trifolium pratense* L.) by single-molecule long-read sequencing. BMC plant biology. 2018;18(1):300.
40. Hu H, Yang W, Zheng Z, et al. Analysis of alternative splicing and alternative polyadenylation in *Populus alba* var. *pyramidalis* by single-molecular long-read sequencing. *Frontiers in genetics*;. 2020;11:48.
41. Deng N, Hou C, Ma F, Liu C, Tian Y. Single-molecule long-read sequencing reveals the diversity of full-length transcripts in leaves of *Gnetum* (Gnetales). Int J Mol Sci. 2019;20(24):6350.
42. Ye J, Cheng S, Zhou X, et al. A global survey of full-length transcriptome of *Ginkgo biloba* reveals transcript variants involved in flavonoid biosynthesis. Ind Crops Prod. 2019;139:111547.
43. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009;10(3):155–9.
44. Karlik E, Ari S, Gozukirmizi N. LncRNAs: genetic and epigenetic effects in plants. Biotechnology Biotechnological Equipment. 2019;33(1):429–39.
45. Liu J, Wang H, Chua NH. Long noncoding RNA transcriptome of plants. J Pl Biotech. 2015;13(3):319–28.
46. Wang L, Xia X, Jiang H, et al. Genome-wide identification and characterization of novel lncRNAs in *Ginkgo biloba*. Trees. 2018;32(5):1429–42.
47. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 2013;31(11):1009.
48. Minoche AE, Dohm JC, Schneider J, et al. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biol. 2015;16(1):184–4.
49. Nystedt B, Street NR, Wetterbom A, et al. The Norway spruce genome sequence and conifer genome evolution. Nature. 2013;497(7451):579–84.
50. Corlett RT. Characteristics of vertebrate-dispersed fruits in Hong Kong. J Trop Ecol. 1996;12(6):819–33.
51. Ridley HN. *Dispersal of plants throughout the world*. L. Reeve & Co., Ltd: Ashford, UK, 1930.
52. Forget P-M, Hammond DS, Milleron T, Thomas R. Seasonality of fruiting and food hoarding by rodents in Neotropical forests: consequences for seed dispersal and seedling recruitment. In: Levey DJ, Silva WR, Galetti M, editors. *Seed Dispersal and Frugivory: Ecology Evolution and Conservation*. CAB International: Wallingford, 2002, pp 241–256.
53. Kubitzki K. Ichthyochory in *Gnetum venosum*. Anais Acad Brasil Ci. 1985;57(4):513–6.

54. Kunimasa K, Ohta T, Tani H, et al. Resveratrol derivative-rich melinjo *(Gnetum gnemon* L.) seed extract suppresses multiple angiogenesis-related endothelial cell functions and tumor angiogenesis. Mol Nutr Food Res. 2011;55(11):1730–4.
55. Gramzow L, Theissen G. A hitchhiker's guide to the MADS world of plants. Genome Biol. 2010;11(6):214.
56. Masiero S, Colombo L, Grini PE, Schnittger A, Kater MM. The emerging importance of type I MADS box transcription factors for plant reproduction. Plant Cell. 2011;23(3):865–72.
57. Gramzow L, Weilandt L, Theissen G. MADS goes genomic in conifers: towards determining the ancestral set of MADS-box genes in seed plants. *Ann Bot* Nov;. 2014;114(7):1407–29.
58. Becker A, Kaufmann K, Freialdenhoven A, et al. A novel MADS-box gene subfamily with a sister-group relationship to class B floral homeotic genes. *Mol Genet Genomics* Feb;. 2002;266(6):942–50.
59. Becker A, Saedler H, Theissen G. Distinct MADS-box gene expression patterns in the reproductive cones of the gymnosperm *Gnetum gnemon*. Dev Genes Evol. 2003;Nov; 213(11):567–72.
60. Wang YQ, Melzer R, Theissen G. Molecular interactions of orthologues of floral homeotic proteins from the gymnosperm *Gnetum gnemon* provide a clue to the evolutionary origin of 'floral quartets'. *Plant J* Oct. 2010;64(2):177–90.
61. Winter KU, Becker A, Munster T, Kim JT, Saedler H, Theissen G. MADS-box genes reveal that gnetophytes are more closely related to conifers than to flowering plants. *Proc Natl Acad Sci USA* Jun. 1999;22(13):7342–7.. ; **96**.
62. Lovisetto A, Guzzo F, Tadiello A, Toffali K, Favretto A, Casadoro G. Molecular analyses of MADS-box genes trace back to Gymnosperms the invention of fleshy fruits. Molec Biol Evol. 2012;29(1):409–19.
63. Liu D, Chen J, Lu W. Expression and regulation of the early auxin-responsive *Aux/IAA* genes during strawberry fruit development. Mol Biol Rep. 2011;38(2):1187–93.
64. Wang H, Jones B, Li Z, et al. The Tomato *Aux/IAA* Transcription Factor *IAA9* Is Involved in Fruit Development and Leaf Morphogenesis. Plant Cell. 2005;17(10):2676–92.
65. Li X, Chen L, Hong M, et al A large insertion in *bHLH* transcription factor *BrTT8* resulting in yellow seed coat in *Brassica rapa*. *Plos One*; 7(9): (2012).
66. Suzuki A, Suzuki T, Tanabe F, et al. Cloning and expression of five myb-related genes from rice seed. Gene. 1997;198(1–2):393–8.
67. Xu W, Dubos C, Lepiniec L. Transcriptional control of flavonoid biosynthesis by MYB–bHLH–WDR complexes. Trends Plant Sci. 2015;20(3):176–85.
68. Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics. 2014;30(24):3506–14.
69. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21(9):1859–75.
70. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucl Acids Res. 2011;39(suppl_2):29–37.

71. Albert VA, Barbazuk WB, Der JP, et al. The *Amborella* genome and the evolution of flowering plants. Science. 2013;342(6165):1241089.
72. R Core Team. R: A language and environment for statistical computing version 3.2.0. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org. (2018).
73. Xie C, Mao X, Huang J, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucl Acids Res. 2011;39:316–22.
74. Haas BJ, Papanicolaou A, Yassour M, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494.
75. Zheng Y, Jiao C, Sun HH, et al. iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. *Mol Plant* Dec. 2016;5(12):1667–70. ; **9**.
76. Kong L, Zhang Y, Ye ZQ, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucl Acids Res* Jul;. 2007;35:345–9.
77. Sun L, Luo HT, Bu DC, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucl Acids Res* Sep;. 2013;41(17):e166.
78. Wang L, Park HJ, Dasari S, Wang SQ, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucl Acids Res Apr. 2013;41(6):e74.
79. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature Meth. 2015;12(4):357.
80. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
81. Leng N, Dawson JA, Thomson JA, et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013;29(8):1035–43.
82. Lalitha S. Primer premier 5. *Biotech Software & Internet Report: The Computer Software Journal for Scient*; 1(6): 270–272 (2000).
83. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the $2 - \Delta\Delta CT$ method. Methods. 2001;25(4):402–8.

Figures

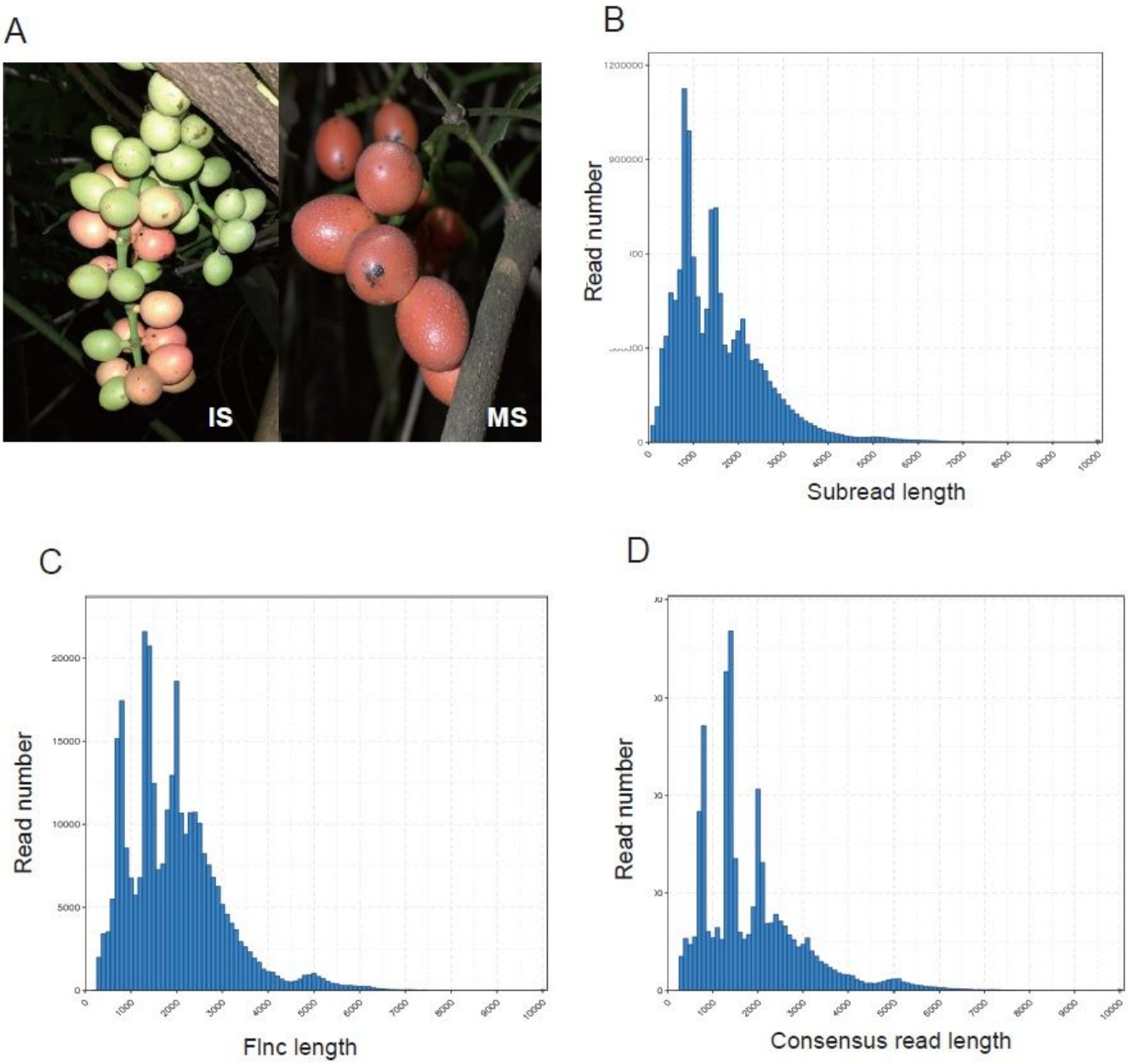


Figure 1

Results of PacBio sequencing and error correction. aImmature seeds (left) and mature seeds (right) of G. luofuense. b length distribution of subreads. c length distribution of FLNCs. d length distribution of consensus isoforms using an isoform-level clusteringalgorithm.

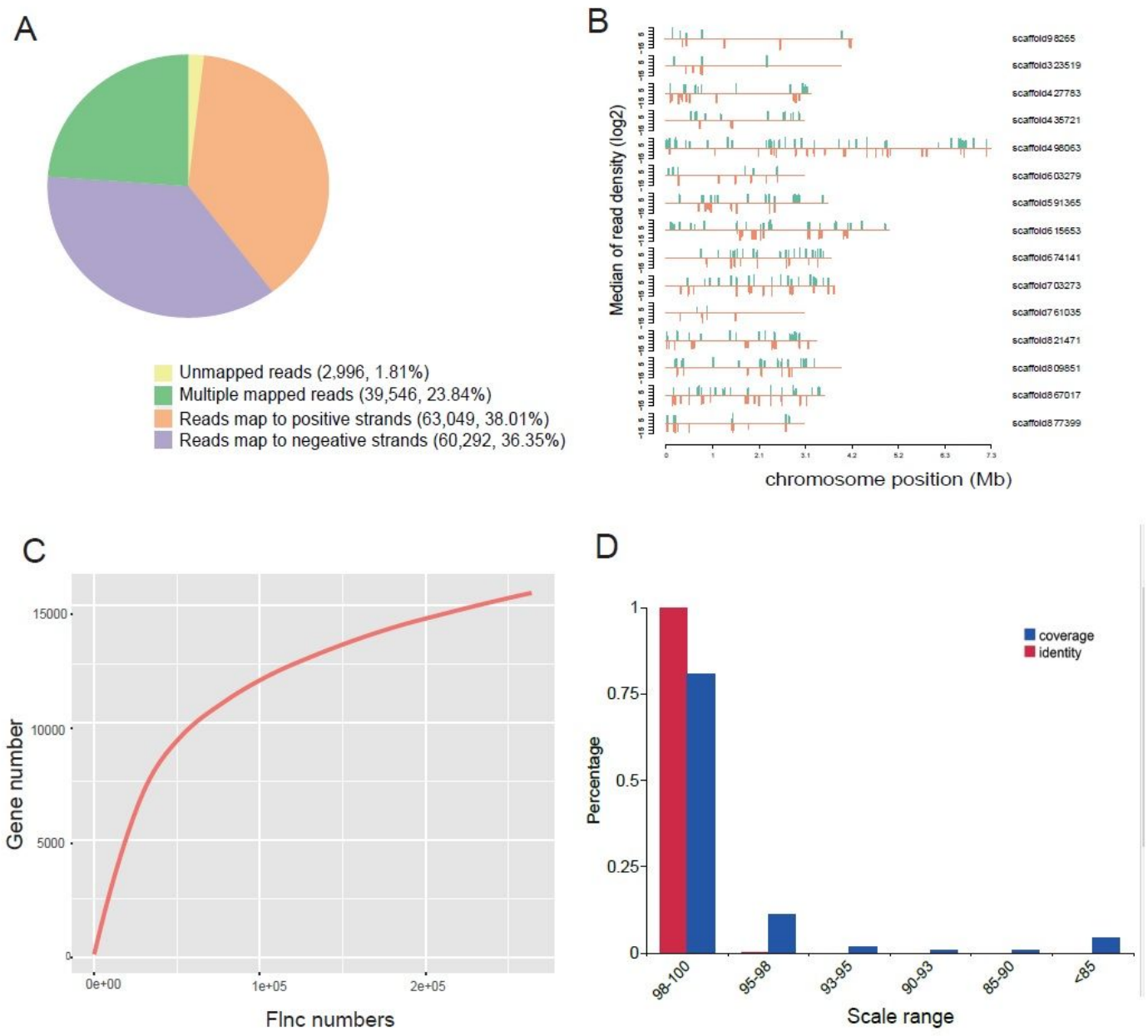


Figure 2

Genome mapping of the full-length transcriptome. a statistics and classification of full-length transcripts against the G. luofuense reference genome. bmapping read density on the G. luofuense scaffolds; the x-axis represents the scaffold position (Mb), the y-axis represents the median read density (log2), and the green and red lines represent the positive and negative strands of the scaffolds. c a saturation curve of consensus reads; the x-axis represents the number of full-length, non-chimeric reads, andthe y-axis represents the number of isoforms. dthe scale and identity range of the mapped full-length transcripts.The red and blue bars represent the coverage and identity of full-length reads.

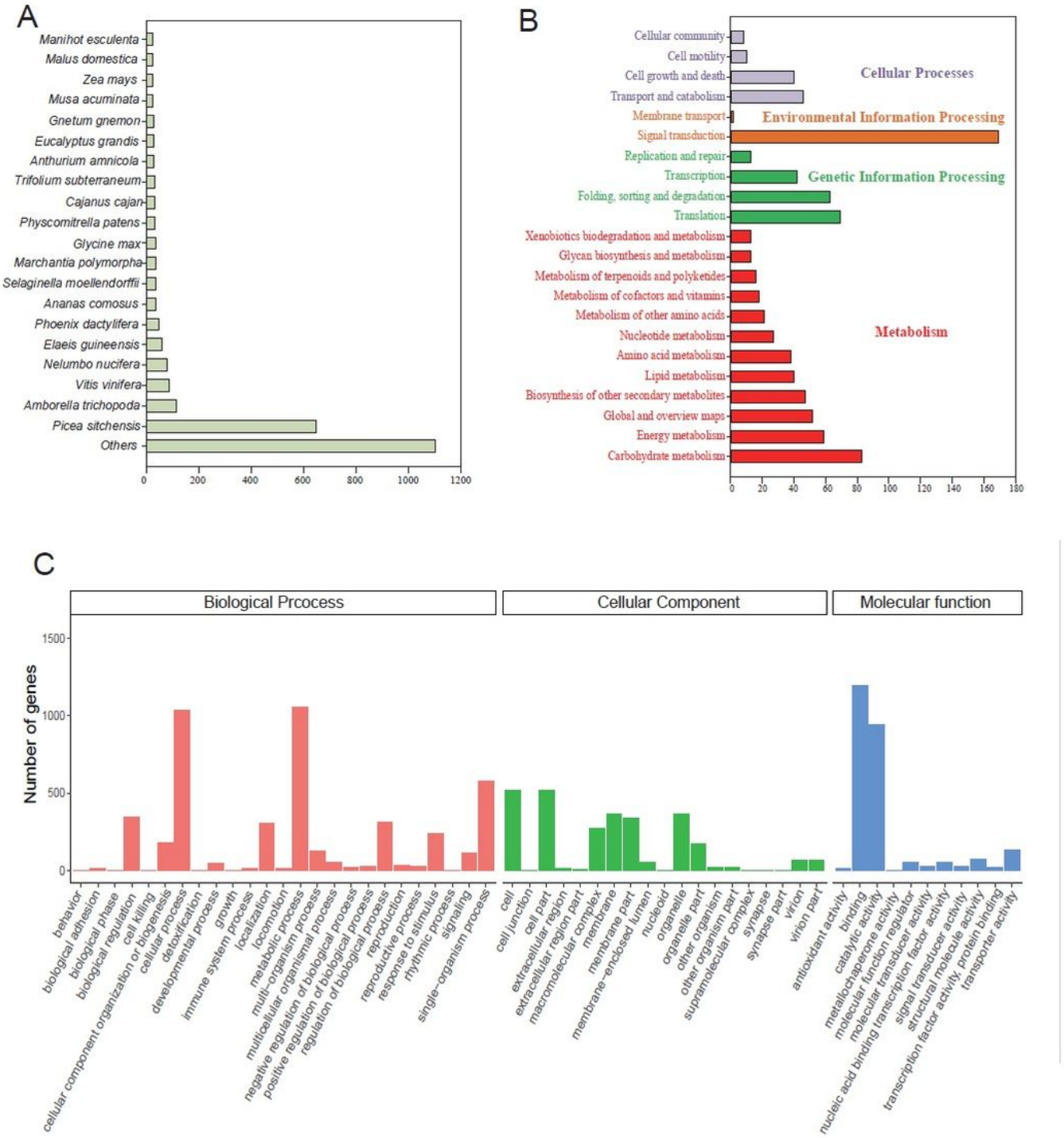


Figure 3

Annotation summary of novel genes from G. luofuenseseeds.athe distribution of NR annotations among different seed plant species, the x-axis represents the number of annotated reads.b KEGG enrichment of the annotated novel genes, the x-axis represents the number of annotated reads. c gene ontology (GO) annotation and categorization of full-length transcripts.

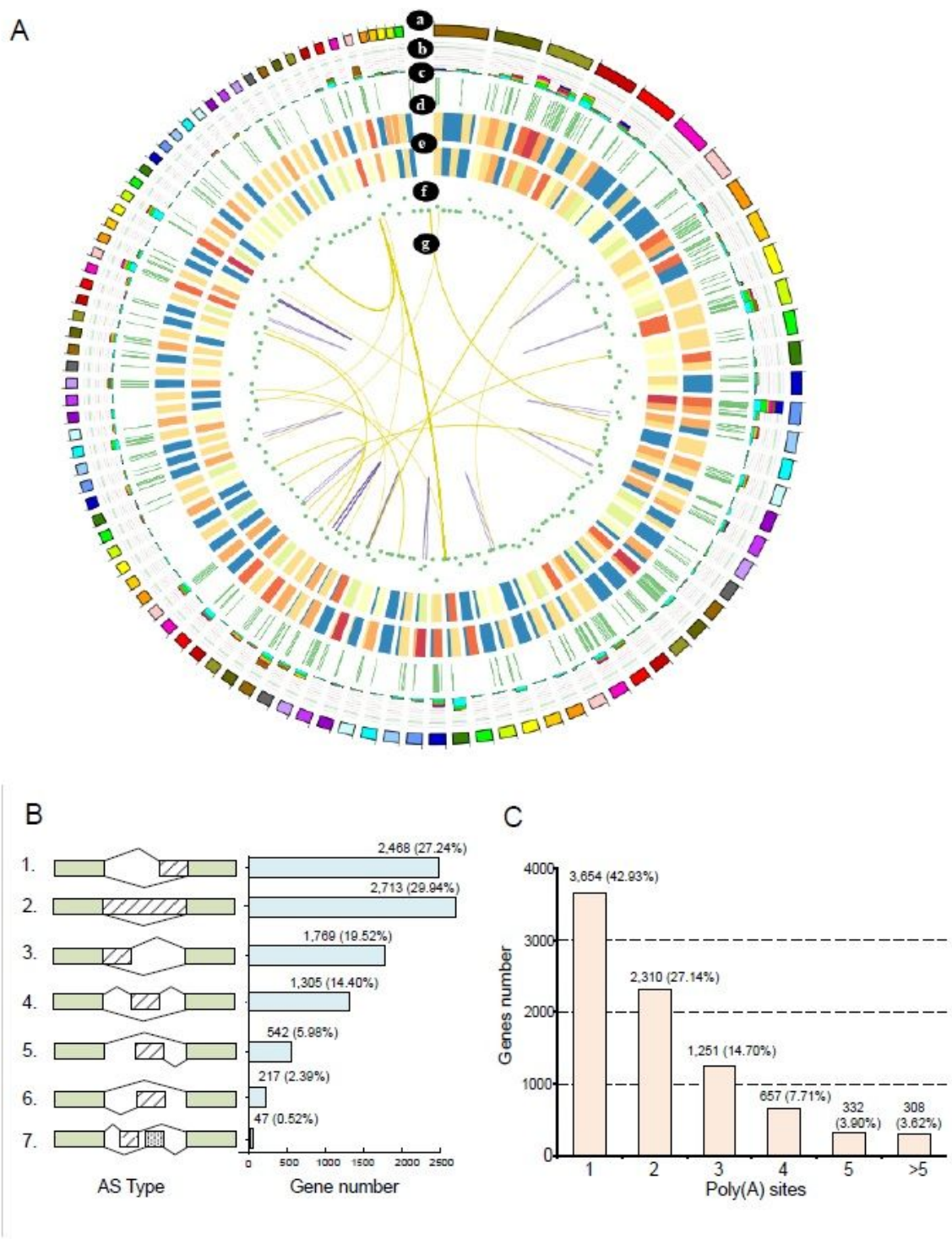


Figure 4

Structural analysis of full-length transcriptome. a Visualization of the full-length transcriptome at the genome-wide scale. a: scaffolds of the G. luofuensereference genome, b: alternative splicing sites; different colors represent various types of alternative splicing events,c: alternative polyadenylation sites, d: distribution of novel isoforms from known genes; high densityisrepresented by warm colors (e.g. red), and low density is represented by dark colors (e.g. blue),e: distribution of isoforms from novel genes;the

red color represents high density,f: distribution of lncRNAs, g: gene fusion events detected in the genome. bnumbers of alternative splicing events identified during seed ripening of G. luofuense.c genes with different numbers of alternative polyadenylation sites identified in the full-length transcriptome of G. luofuense seeds.

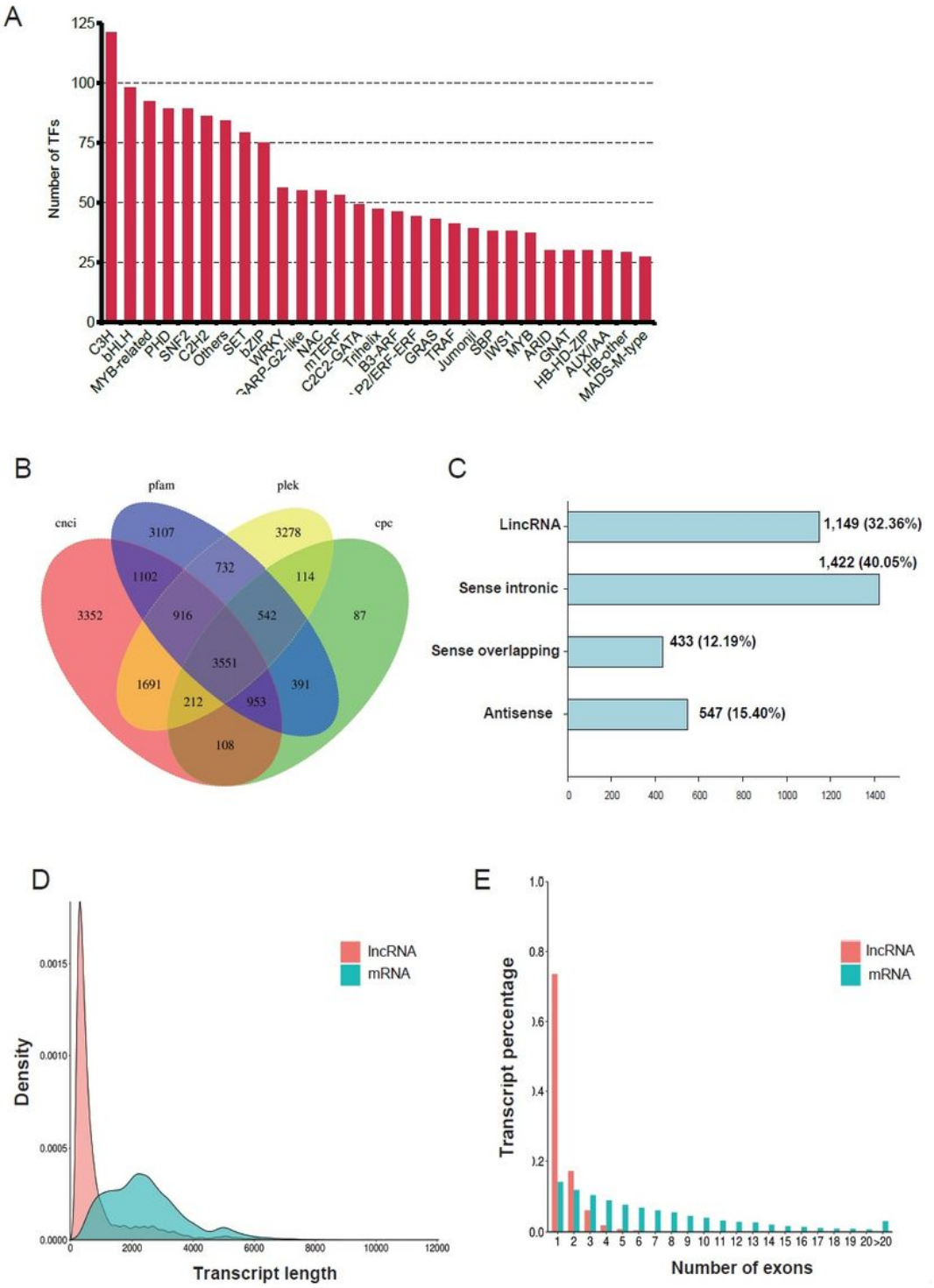
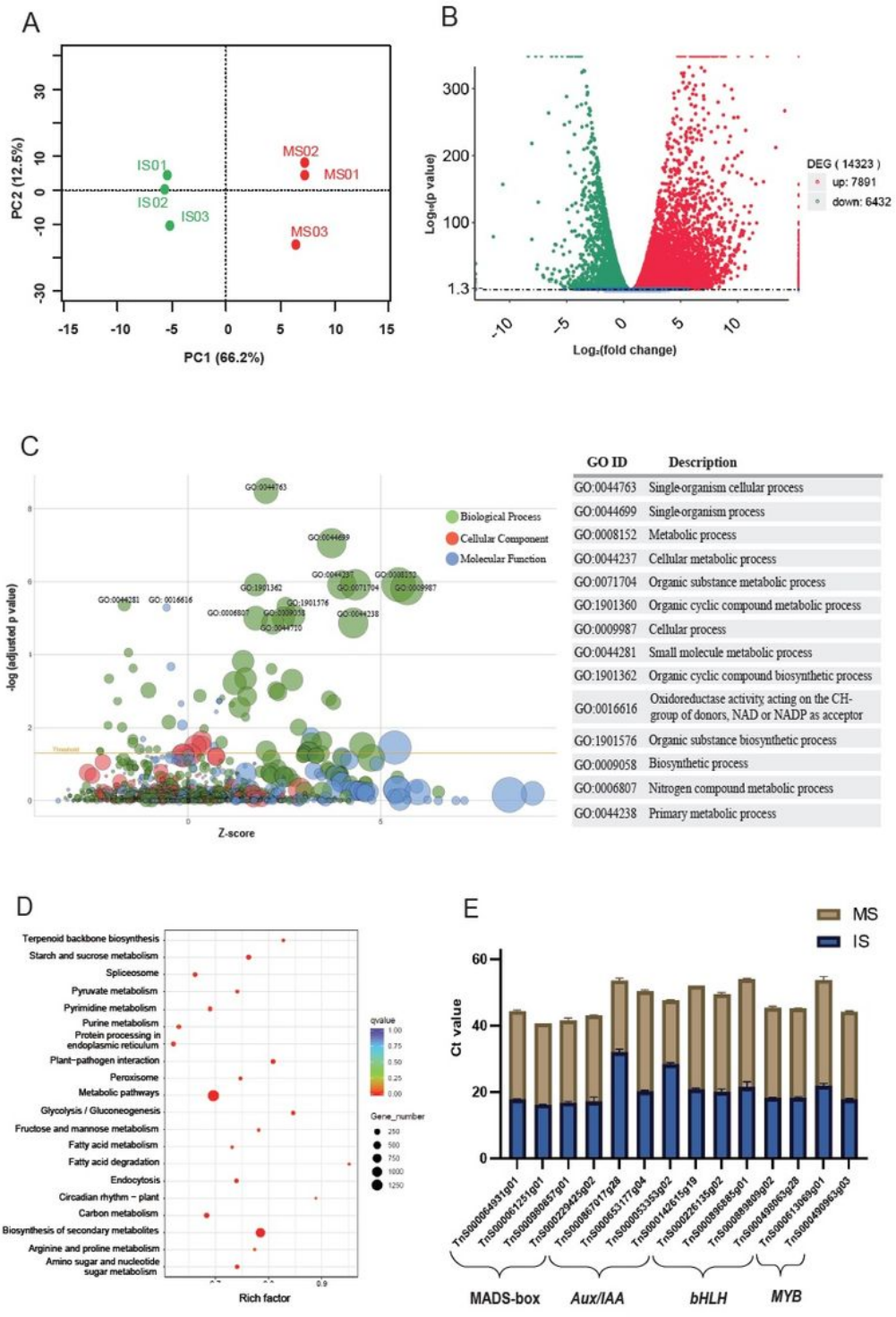


Figure 5

Transcription factors and lncRNAs. a A partial list of transcription factors (top 28 gene families) identified in the full-length transcriptome of G. luofuense seeds. bVenn diagram showing thenumber of lncRNAs identified using four different approaches: CPC (Coding Potential Calculator), CNCI (Coding-Non-Coding Index), CPAT (Coding Potential Assessment Tool), and Pfam (Protein Family). c functional classification and numbers of four lncRNA types. d the length density distribution of identified lncRNAs on the reference genome of G. luofuensecompared to that of identified lncRNA in the full-length transcriptome. e distribution of exon numbers in mRNAs predicted by the reference genome and identifiedlncRNAs in the full-length transcriptome.



GO ID	Description
GO:0044763	Single-organism cellular process
GO:0044699	Single-organism process
GO:0008152	Metabolic process
GO:0044237	Cellular metabolic process
GO:0071704	Organic substance metabolic process
GO:1901360	Organic cyclic compound metabolic process
GO:0009987	Cellular process
GO:0044281	Small molecule metabolic process
GO:1901362	Organic cyclic compound biosynthetic process
GO:0016616	Oxidoreductase activity acting on the CH-group of donors, NAD or NADP as acceptor
GO:1901576	Organic substance biosynthetic process
GO:0009058	Biosynthetic process
GO:0006807	Nitrogen compound metabolic process
GO:0044238	Primary metabolic process

Figure 6

Detection of DEGs and qRT-PCR validation. a PCA analysis of gene expression in the three immature seed samples (IS01–03) and three mature seed samples (MS01–03). b a volcano plot of differential gene expression between immature and mature seed samples of G. luofuense, with upregulated genes in red and downregulated genes in green from immature seeds to mature seeds. c a bubble plot of enriched GO terms; the x-axisrepresentsthe z-score, they-axis represents the negative logarithm of the adjusted p-

values, the circle sizes are proportional to the number of genes enriched in the GO terms, and the circle colors denote the three GO term categories.d a bubble plot of enriched KEGG terms; the x-axis represents rich factors, the circle sizes are proportional to enriched gene numbers, and the circle colors correspond to the negative logarithm of the adjusted p-values for each KEGG pathway.e The expression of 14 TF genes (i.e. MADS-box, Aux/IAA, bHLH and MYB genes) from immature and mature seeds of G. luofuense were verified by qRT-PCR, and the expression values were normalized with the △△Ct-method.