



Article

Full-Length Transcriptome Characterization and Comparative Analysis of *Chosenia arbutifolia*

Xudong He^{1,2,*} , Yu Wang^{1,3}, Jiwei Zheng^{1,2}, Jie Zhou^{1,2}, Zhongyi Jiao^{1,2}, Baosong Wang^{1,2} and Qiang Zhuge³ 

¹ Willow Engineering Technology Research Center of National Forestry and Grassland Administration, Jiangsu Academy of Forestry, Nanjing 211153, China; wangyu84yu@163.com (Y.W.); zjw932333@163.com (J.Z.); zjwin718@126.com (J.Z.); tianyaguke2009@163.com (Z.J.); baosongwang1965@163.com (B.W.)

² Willow Nursery of the Jiangsu Provincial Platform for Conservation and Utilization of Agricultural Germplasm, Jiangsu Academy of Forestry, Nanjing 211153, China

³ College of Biology and the Environment, Nanjing Forestry University, Nanjing 210037, China; qzhuge@njfu.edu.cn

* Correspondence: hxd_519@163.com; Tel.: +86-25-5274-4180

Abstract: As a unique tree species in the Salicaceae family, *Chosenia arbutifolia* is used primarily for construction materials and landscape planting in China. Compared with other Salicaceae species members, the genomic resources of *C. arbutifolia* are extremely scarce. Thus, in the present study, the full-length transcriptome of *C. arbutifolia* was sequenced by single-molecular real-time sequencing (SMRT) technology based on the PacBio platform. Then, it was compared against those of other Salicaceae species. We generated 17,397,064 subreads and 95,940 polished reads with an average length of 1812 bp, which were acquired through calibration, clustering, and polishing. In total, 50,073 genes were reconstructed, of which 48,174 open reading frames, 4281 long non-coding RNAs, and 3121 transcription factors were discovered. Functional annotation revealed that 47,717 genes had a hit in at least one of five reference databases. Moreover, a set of 12,332 putative SSR markers were screened among the reconstructed genes. Single-copy and special orthogroups, and divergent and conserved genes, were identified and analyzed to find divergence among *C. arbutifolia* and the five Salicaceae species. To reveal genes involved in a specific function and pathway, enrichment analyses for GO and KEGG were also performed. In conclusion, the present study empirically confirmed that SMRT sequencing realistically depicted the *C. arbutifolia* transcriptome and provided a comprehensive reference for functional genomic research on Salicaceae species.

Keywords: *Chosenia arbutifolia*; EST-SSR; *Salix*; SMRT sequencing; transcriptome



Citation: He, X.; Wang, Y.; Zheng, J.; Zhou, J.; Jiao, Z.; Wang, B.; Zhuge, Q. Full-Length Transcriptome Characterization and Comparative Analysis of *Chosenia arbutifolia*. *Forests* **2022**, *13*, 543. <https://doi.org/10.3390/f13040543>

Academic Editors: Giovanni Emiliani and Alessio Giovannelli

Received: 11 March 2022

Accepted: 30 March 2022

Published: 31 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The fast-growing tree *Chosenia arbutifolia* (Pall.) A. Skv. has a wide distribution range over northeastern China, northeastern North Korea, the Honshu and Hokkaido Islands of Japan, and the Russian Far East. It has been extensively utilized as a source of construction materials and landscape planting [1]. It has unique morphological features that differentiate it from poplar and willow, such as missing glands, unique flower structures, and special leaves. Hence, *C. arbutifolia* was segregated into the genus *Chosenia* by certain taxonomists [2,3]. However, there is molecular evidence that *Chosenia* should be categorized into *Salix* [4,5]. *C. arbutifolia* was defined as synonymic to *S. arbutifolia* in the Plant List (<http://www.theplantlist.org/tpl/record/tro-28300001>, accessed on 10 January 2022). Unlike other *Populus* and *Salix* species, *C. arbutifolia* is not readily propagated through cuttings. Furthermore, *C. arbutifolia* seed germination requires severe conditions, such as water flow and sediment accumulation [6]. Extensive deforestation and poor reproductive capacity have diminished indigenous *C. arbutifolia* populations and this tree species is now classified as endangered in China.

Whole-genome sequencing (WGS) is used to explore gene structure and variation. However, the universal application of this technology is constrained by high costs and long run times, especially for species with complex genetic backgrounds. Recently, RNA sequencing based on high-throughput sequencing platforms has provided vital advances and can be effectively implemented in non-model species lacking reference genomes. RNA sequencing realistically reflects gene copy number, type, expression level, and structure, and reveals new genes, potential metabolic pathways, and genetic mechanisms [7,8]. Despite some challenges such as length bias, read accuracy, and high cost, the third-generation single-molecular real-time sequencing (SMRT) technology has the momentous advantages of super-long reads and high throughput, no GC bias and the tendentiousness of PCR amplification, and directly detects DNA base modifications compared with short-read sequencing technology [9]. Thus, full-length transcriptome sequencing (Iso-seq) obtains entire transcript sequences without assembly and identifies multiple forms of alternative splicing [10,11]. It also discovers new genes and analyzes fusion, homologous, superfamily, and allelomorphic genes [12,13], and has been massively utilized in several woody species, including *Cinnamomum porrectum* [14], *Torreya grandis* [15], *Olea europaea* [16], *Vitis vinifera* [17], *Rhododendron lapponicum* [18], and *Cephalotaxus oliveri* [19].

Despite the importance and endangered status of *C. arbutifolia*, few studies have been conducted on it. Moreover, they focused primarily on the ecological and biological characteristics [1,6], propagation methods [20], population protection [21], molecular markers development [22], gene family analysis [23–25], whole-chloroplast genome sequencing [5], and population structure evaluation [26,27] of *C. arbutifolia*. Only one published report evaluated the *C. arbutifolia* transcriptome based on the Illumina HiSeq [28]. Several *Populus* species such as *P. deltoides* × *P. euramericana* cv. ‘Nanlin895’ [29], *P. wulianensis* [30], and *P. alba* var. *pyramidalis* [31] have already been subjected to full-length transcriptome sequencing. In contrast, there are few available transcriptome resources for *Salix* species and especially *C. arbutifolia*. In this study, we sequenced the full-length transcriptome of *C. arbutifolia* through SMRT technology and compared it against five other Salicaceae species to elucidate the gene functions and phylogenetic relationships in the Salicaceae. The results of this work could provide substantial resources for functional genomic research and may help clarify the complex evolutionary architecture of the Salicaceae.

2. Materials and Methods

2.1. Plant Materials and RNA Extraction

Twigs and flower branches were sampled from a single superior *C. arbutifolia* individual in Fusong, Jilin Province, China (41°47′10.55″ N, 127°55′56.13″ E) and brought to the laboratory. The plant material was hydroponically cultivated in a greenhouse. Young leaves, buds, male flowers, roots, and stem bark were sampled and their total RNAs were isolated with an Omega Plant RNA Kit (Omega Bio-tek, Norcross, GA, USA). Extracted RNA purity and concentration were evaluated through 0.8% agarose gels electrophoresis and an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).

2.2. cDNA Library Construction and Iso-Seq

Equivalent RNAs from diverse tissues were mixed as a single sample. Oligo (dT) was utilized for enrichment of mRNAs containing poly-A. The SMARTer PCR cDNA Synthesis Kit (TaKaRa Bio Inc., Kusatsu, Shiga, Japan) was used for cDNA reverse transcription. Partial cDNAs were screened with BluePippin for fragments > 4 kb and amplified by PCR. The full-length cDNAs were joined with a dumbbell-shaped SMRT adapter after damage and end repair. Sequences without adapter ligation were removed by exonuclease digestion. The complete SMRT bell library was constructed via primer ligation and DNA polymerase binding, evaluated by QC, and subjected to Iso-seq on a PacBio Sequel platform [9].

2.3. Data Processing

The PacBio official software packages, including *ccs*, *lima*, and *isoseq3*, were employed to process the raw offline data. The subread sequences were generated by filtration and removal of adapters and sequences < 50 bp. They were then calibrated in *ccs* to obtain circle consensus sequences (CCS). Based on the existence of 3'-primers, 5'-primers, or poly-A tails, the CCS were then separated with *lima* into full-length and non-full-length sequences. After removing poly-A with *isoseq3*, high-quality sequences were acquired by clustering and polishing the consensus sequences.

2.4. Sequences Analysis and Functional Annotation

The software programs PLEK [32], CNCI [33], CPC [34], and the Pfam database [35] were used to identify long non-coding RNAs (LncRNAs). The software program iTAK [36] was employed to predict transcription factors (TFs). The open reading frames (ORFs) were captured using software ANGEL [37]. For functional annotation, the reconstructed genes were compared against various protein databases, including non-redundant (NR), gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), clusters of orthologous groups (COG), and SWISS with the software program BlastX at $E \leq 1.0 \times 10^{-5}$. MISA (<http://pgrc.ipkgatersleben.de/misa/misa.html>, accessed on 5 February 2021) program in Perl script was run to screen SSRs with di-, tri-, tetra-, penta-, and hexa-nucleotide motifs at minimum repeats of 6, 5, 5, 5, and 5, respectively.

2.5. Comparative Transcriptome Analysis

Five whole-genome sequences for the Salicaceae species *S. suchowensis* (PRJNA668632), *S. brachista* (PRJNA472210), *P. alba* (PRJNA491245), *P. deltoides* (PRJNA598948), and *P. trichocarpa* (PRJNA10772) were accessed from the NCBI database to perform comparative analyses. The amino acid sequences were aligned using OrthoFinder v. 2.3.3 [38] under the default parameters to obtain homologous genes. Single-copy ortholog sequences in all six species were selected and subjected to multiple sequence alignment using software MAFFT [39]. The output was utilized to construct a phylogenetic with FastTree software [40]. Goatools [41] and KOBAS [42] based on Fisher's exact test ($p < 0.05$) were employed for the GO and KEGG enrichment analyses, respectively.

3. Results

3.1. Full-Length Transcriptome Sequencing

We filtered the raw offline sequencing data and generated 25.78 Gb of subreads by full-length transcriptome sequencing. There were 17,397,064 subreads and their lengths were in the range of 51–189,250 bp. The average length was 1482 bp (Table 1). Of these, 272,952 circle consensus sequences (CCS; mean, minimum, and maximum lengths 1759 bp, 62 bp, and 16,567 bp, respectively) were obtained after calibration (Table 1). The detection and elimination of 3'-primers, 5'-primers, and poly-A tails among the CCS yielded a net 232,288 full-length non-chimeric reads (FLNC), namely full-length transcripts. A total of 95,940 polished reads ranging from 51 bp to 13,247 bp were obtained for the following analysis after FLNC clustering and polishing. The average length was 1812 bp (Table 1). Overall, the N50 were 1616 bp, 1939 bp, and 2086 bp for the subreads, CCS, and polished reads, respectively (Table 1).

As there was alternative splicing, the genome coding area should be reconstructed with high-quality polished reads to form a complete coding area. The numbers of reconstructed genes and related transcripts were 50,073 (95,646,387 bp) and 95,940 (173,883,048 bp), respectively, and their average lengths were 1910 bp and 1812 bp, respectively (Table 1). A total of 48,174 ORFs and 4281 LncRNAs were predicted, and their average lengths were 986 bp and 1120 bp, respectively. In total, 3121 transcription factors (TFs) were identified, of which 2288 were true TFs and the remaining 833 were transcriptional regulators. Out of the 3121 TFs, 2942 were classified into 90 families, while the remaining 179 were sorted into other taxa. Table 2 shows the top 10 TF families.

Table 1. Characteristics of *Chosenia arbutifolia* transcriptome sequences.

Item	Number	Average Length (bp)	N50	Min_Length	Max_Length
Subreads	17,397,064	1482	1610	51	189,250
CCS	272,952	1759	1939	62	16,567
Polished reads	95,940	1812	2086	51	13,247
Cogent genes	50,073	1910	NA	51	13,835
ORF	48,174	986	NA	147	9693
LncRNA	4281	1120	NA	132	6754

Table 2. Top 10 TF families in *Chosenia arbutifolia* transcriptome sequences.

Family	Number	Percent (%)
bHLH	142	4.83
AP2-ERF	134	4.55
MYB	127	4.32
C3H	121	4.11
bZIP	117	3.98
C2H2	115	3.91
MYB-related	103	3.50
GRAS	102	3.47
NAC	92	3.13
WRKY	92	3.13

3.2. Functional Classification and Annotation

The putative functions of all 50,073 reconstructed genes were compared against the NR, GO, COG, KEGG, and SWSS databases (Figure 1A). We annotated 47,678 (95.22%) and 44,765 (89.40%) genes in the NR and GO databases, and 38,606 (77.10%), 33,587 (67.08%), and 21,499 (42.94%) in the SWSS, COG, and KEGG databases, respectively. Moreover, 18,096 (37.76%) genes were significantly correlated with sequences in all databases and 47,717 (95.29%) genes had hits in at least one database (Table 3).

Table 3. Functional annotation of reconstructed genes.

Database	Annotated Genes	Percent (%)
NR	47,678	95.22
GO	44,765	89.40
COG	33,587	67.08
KEGG	21,499	42.94
SWSS	38,606	77.10
In_all_DB	18,096	37.76
At_least_one_DB	47,717	95.29
Total_genes	50,073	100

For the GO term classification, all 44,765 genes assigned to three primary categories were divided into 59 subgroups (Figure 1B). ‘Biological process’ had 27 subgroups, and ‘cellular process’ was the most abundant (28,978), followed by ‘metabolic process’ (23,279) and ‘biological regulation’ (17,382). The ‘cellular component’ category had 19 subcategories and consisted mainly of ‘cell’ (38,644), ‘cell part’ (38,644), and ‘organelle’ (29,774). The ‘molecular function’ category contained 13 subcategories enriched with ‘catalytic activity’ (19,462), ‘binding’ (16,182), and ‘transcription regulator activity’ (3328). Compared with the KEGG database, the metabolic pathway analysis demonstrated that 21,499 genes were involved in 400 pathways. The most enriched pathways were ‘global and overview maps’ (5677), followed by ‘signal transduction’ (5984), and ‘carbohydrate metabolism’ (3817) (Figure 1C). Furthermore, 33,587 genes were separated into 25 categories for COG classification (Figure 1D). The ‘function unknown’ part explained the largest proportion (3937; 11.72%), followed by ‘posttranslational modification, protein turnover, chaperones’

(3818; 11.37%) and ‘signal transduction mechanisms’ (3254; 9.69%). Moreover, ‘cell motility’, ‘general function prediction only’, ‘extracellular structure’, and ‘nuclear structure’ were each <1% (Figure 1D).

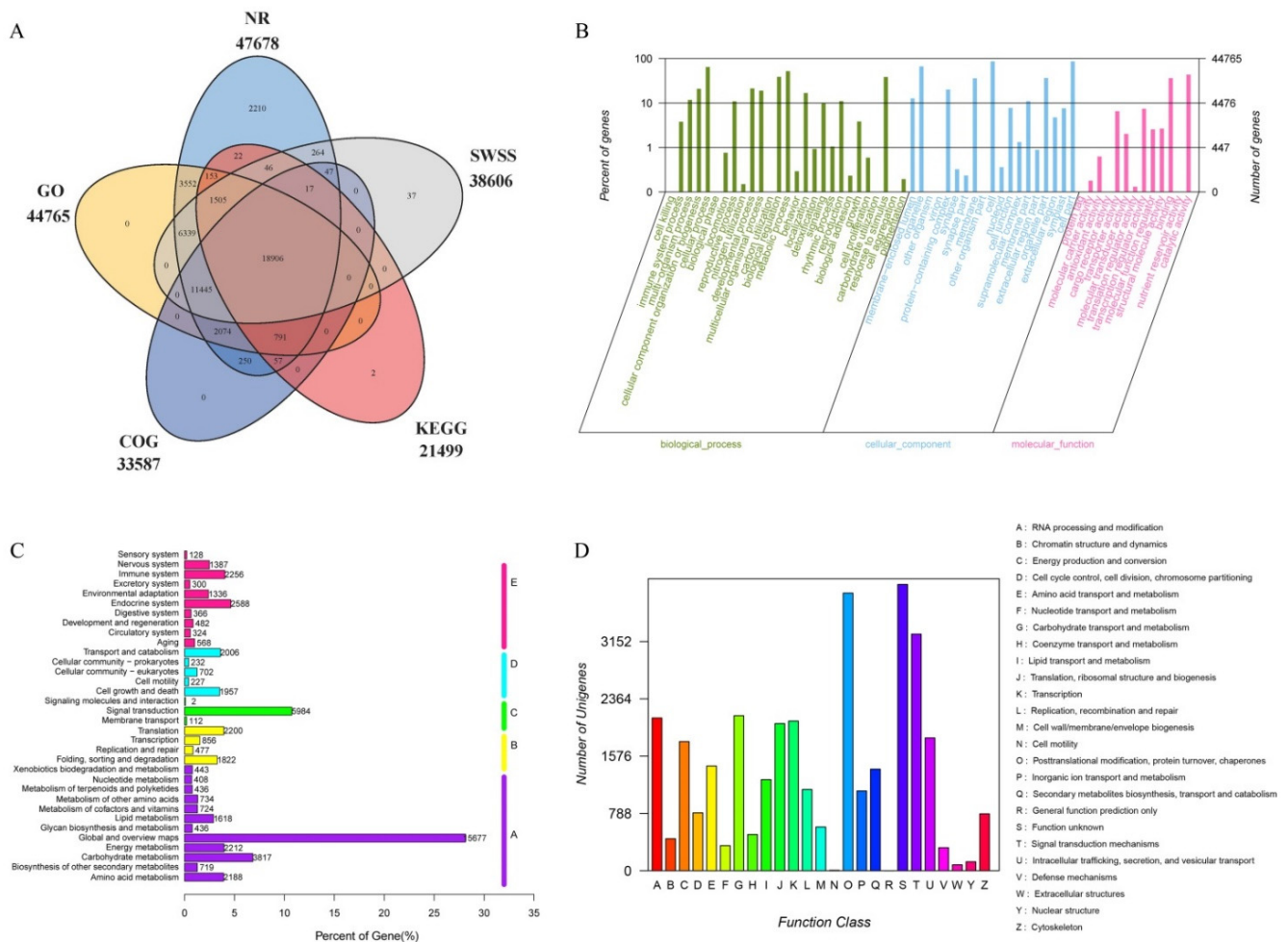


Figure 1. Gene annotation and classification for *Chosenia arbutifolia*. (A) Venn diagrams of reconstructed genes annotated in different databases. (B) Gene ontology term classifications for the reconstructed genes. (C) General pathway assignment based on KEGG database. (D) Histograms depicting clusters of orthologous groups.

3.3. SSR Mining

MISA screened 12,332 putative SSRs in 50,073 reconstructed genes. Of these, 2110 (4.2%) sequences contained at least two SSRs. The occurrence frequencies were one SSR per 4.06 genes and 7.76 kb in sequence length in *C. arbutifolia*. Table 4 shows that the dinucleotide repeat motif (6640) accounted for 53.8% and was the most abundant, followed by tri- (4996; 40.5%), tetra- (442; 3.5%), and hexa-nucleotide (141; 1.1%), whereas penta-nucleotide (113; 0.9%) was the least. Twelve different motifs were observed for the dinucleotide, and AG/CT was the most frequent (53.4%). Of the 60 types of tri-nucleotides and 94 types of tetra-nucleotides, AAG/CTT and AAAT/ATTTA were the most abundant, with frequencies of 20.0% and 8.6%, respectively. The penta- and hexa-nucleotides had 64 and 102 types of motifs, respectively. AGGGG/CCCCT and AAGGAG/CCTTCT were the most frequent, taking 11.5% and 9.9%, respectively.

Table 4. SSR occurrence and composition in *Chosenia arbutifolia*.

Repeat Motif	Number	No. Types	Dominant Motif	No. Dominant Motifs
Di-nucleotide	6640 (53.8%)	12	AG/CT	3546 (53.4%)
Tri-nucleotide	4996 (40.5%)	60	AAG/CTT	999 (20.0%)
Tetra-nucleotide	442 (3.5%)	94	AAAT/ATTT	38 (8.6%)
Penta-nucleotide	113 (0.9%)	64	AGGGG/CCCCT	13 (11.5%)
Hexa-nucleotide	141 (1.1%)	102	AAGGAG/CCCTCT	14 (9.9%)
Total	12,332 (100%)	336		

3.4. Comparative Analysis

The whole-genome sequences for three *Populus* and two *Salix* species were downloaded from NCBI, and included *P. alba*, *P. deltoides*, *P. trichocarpa*, *S. suchowensis*, and *S. brachista*. The amino acid sequences of all six species were clustered by similarity with Orthofinder. The homologous genes in each species are shown in Figure 2. A total of 2732 single-copy orthogroups were detected for all six species. The 19,704 most special orthogroups were screened in *C. arbutifolia*, followed by *P. deltoides* (6479), *S. suchowensis* (5181), *S. brachista* (3697), and *P. trichocarpa* (3063), whereas *P. alba* (1030) had the fewest special orthogroups (Figure 2A). The single-copy ortholog sequences of all species were selected for multiple sequence alignment using MAFFT. Based on the alignment results, a whole genome level phylogenetic tree was established using FastTree (Figure 2B). *C. arbutifolia* was assigned to one clade, while the other five species were relegated to another clade. The main clades were divided in the branches *Populus* and *Salix*. *P. alba*, *P. deltoides*, and *P. trichocarpa* were clustered in the former, while *S. brachista* and *S. suchowensis* were clustered in the *Salix* latter.

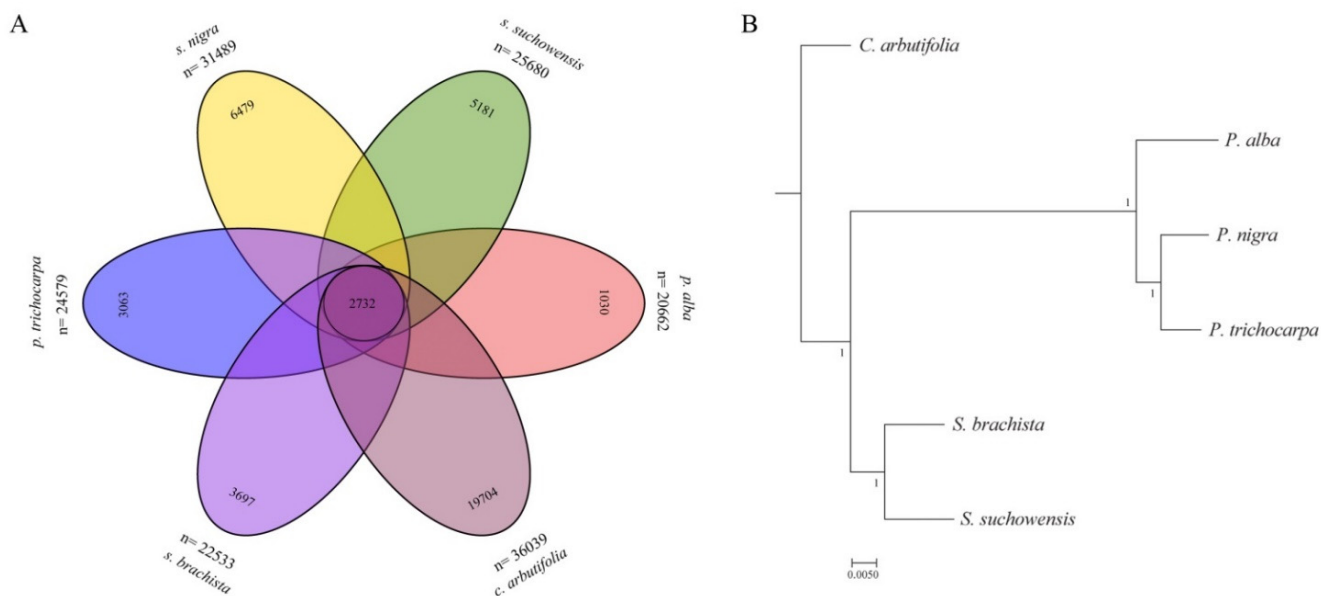


Figure 2. Analysis of orthologous genes in *Chosenia arbutifolia* and five other Salicaceae species. (A) Venn diagrams of orthologous genes. Numbers in middle circle represent total single-copy ortholog sequences for all six species. Numbers at interior edges of oval represent special orthogroups. Numbers outside oval represent total orthogroups. (B) Phylogenetic tree of *Chosenia arbutifolia* and five other Salicaceae species.

All 19,704 special orthogroups in *C. arbutifolia* were analyzed for GO and KEGG annotations. As shown in Figure 3, a total of 17,511 special orthogroups had hits and were assigned to 58 subcategories for GO term classification. ‘Cellular process’ (11,546; 66.04%), ‘cell and cell part’ (15,285; 87.29%), and ‘catalytic activity’ (8026; 45.83%) were the most plen-

tiful in the ‘biological process’, ‘cellular component’, and ‘molecular function’ categories, respectively. The KEGG pathway analysis demonstrated that 312 pathways were enriched in 9112 (46.24%) of the detected special orthogroups. The pathway involving the greatest number of special orthogroups was ‘metabolic pathways’ (3059), followed by ‘biosynthesis of secondary metabolites’ (1669) and ‘biosynthesis of antibiotics’ (854). Moreover, the results of the functional enrichment are displayed in Figure 4. For the GO enrichment, 13,301 genes involved in ‘molecular function’ were identified, followed by 11,145 genes related to ‘cytoplasm’, whereas the fewest 188 genes were detected in ‘fungal-type vacuole’ (Figure 4A). For the KEGG enrichment, ‘microbial metabolism in diverse environments’ was the most enriched pathway (759), followed by ‘glycolysis/gluconeogenesis’ (290) and ‘starch and sucrose metabolism’ (221) (Figure 4B).

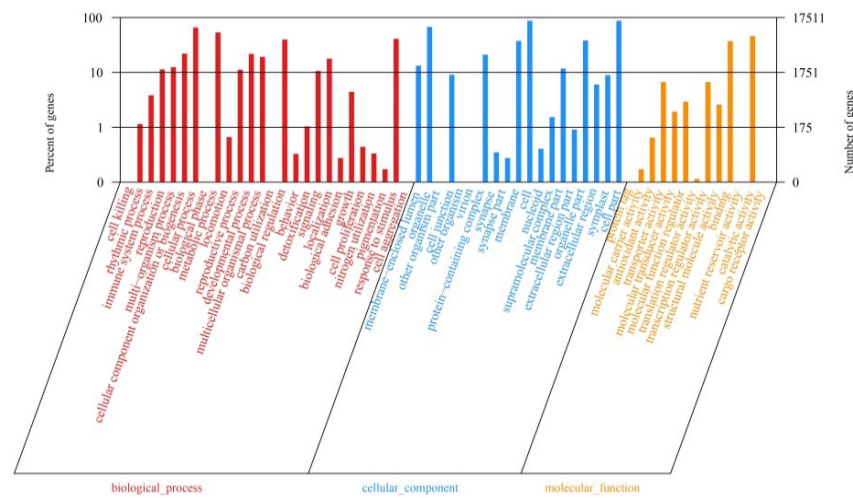


Figure 3. Gene ontology (GO) term classifications of special orthogroups in *Chosenia arbutifolia*.

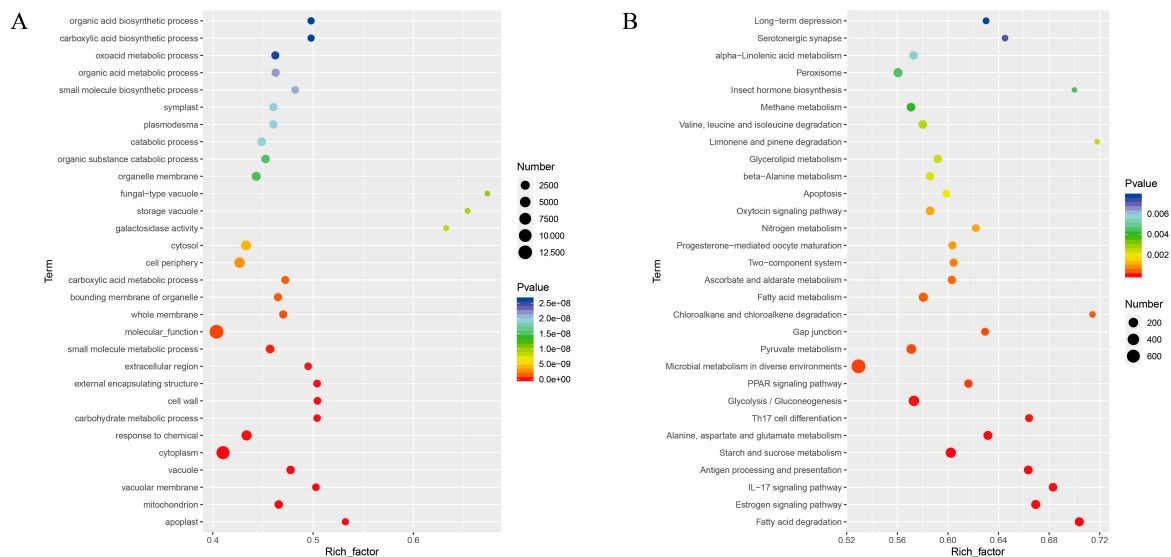


Figure 4. Bubble charts of enrichment for special orthogroups. (A) GO enrichment. (B) KEGG enrichment.

The divergent orthologous genes under positive selection ($Ka/Ks > 1$) and conserved orthologous genes under selection constraints ($Ka/Ks < 0.1$) were evaluated. Of the 2732 single-copy orthogroups in *C. arbutifolia*, 153 divergent and 41 conserved genes were identified. Of these, 122 (79.7%) divergent and 39 (95.1%) conserved genes had specific functions. The GO enrichment analysis demonstrated that for the divergent genes,

eight genes were involved in ‘nuclease activity’ and ‘endonuclease activity’, seven were associated with ‘RNA modification’, and six were related to ‘multicellular organism reproduction’ (Figure 5A). For the conserved genes, 16 genes were associated with ‘cellular component organization’ and ‘cellular component organization or biogenesis’, 14 were related to ‘protein-containing complex’, and 5 were associated with ‘vesicle-mediated transport’ (Figure 5C). The KEGG enrichment results disclosed that two genes were related to the ‘thermogenesis’, ‘retrograde endocannabinoid signaling’, and ‘nucleotide excision repair’ pathways, while the other 24 pathways only had one gene for the divergent genes (Figure 5B). For the conserved genes, four genes were involved in the ‘endocytosis’ pathway and three genes were associated with ‘necroptosis’ (Figure 5D).

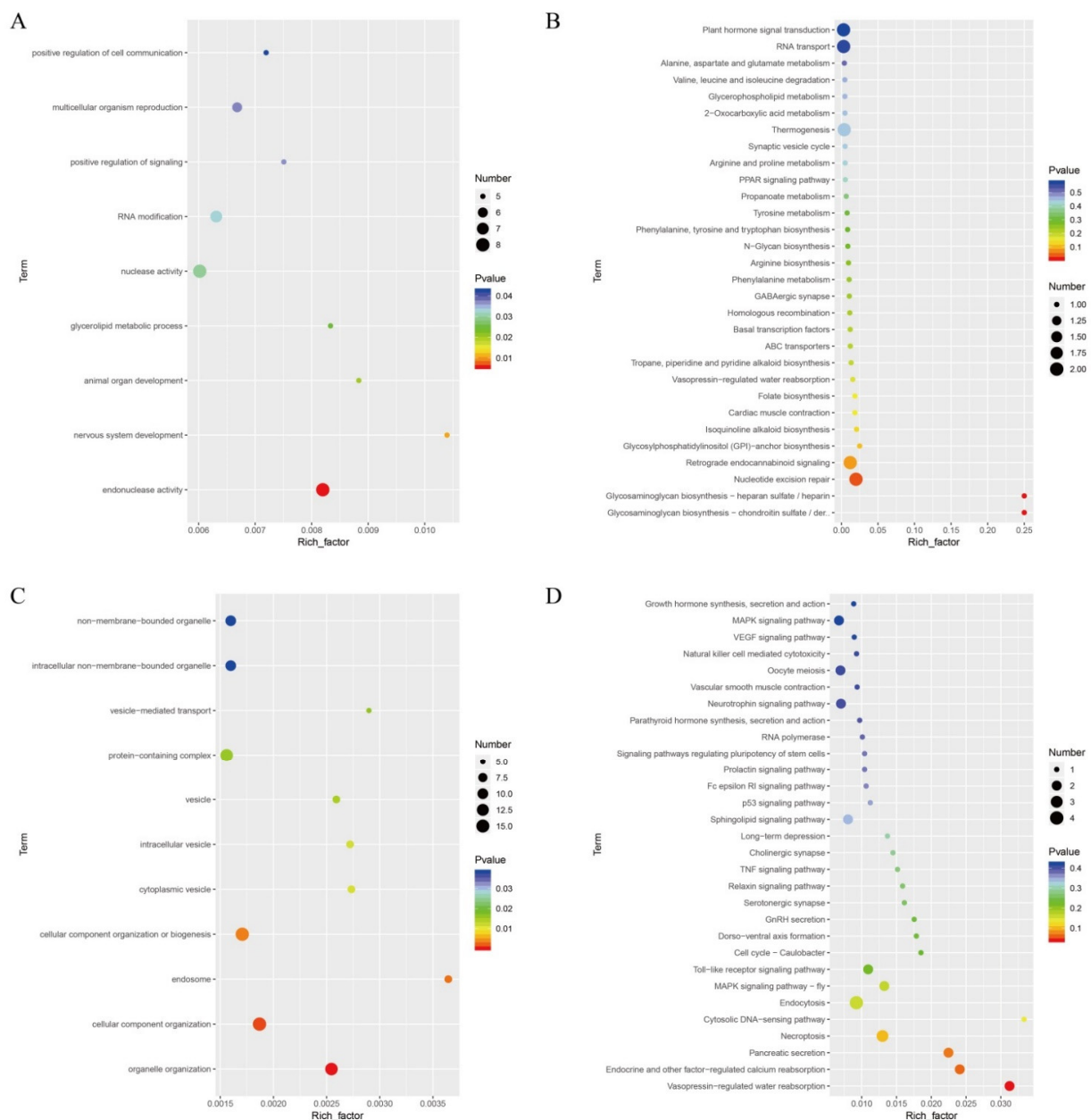


Figure 5. Bubble charts of enrichment for two types of orthologous genes. (A) GO enrichment of divergent genes. (B) KEGG enrichment of divergent genes. (C) GO enrichment of conserved genes. (D) KEGG enrichment of conserved genes.

4. Discussion

In the recent decade, high-throughput sequencing technology and bioinformatics have progressed and have been extensively applied in genetic and genomic research. Full-length transcriptome sequencing with SMRT technology is efficient and reliable in transcriptome reconstruction, especially for non-model organisms lacking de novo assembly [12,13]. Unlike short-read sequencing, Iso-seq provides much longer reads and can overcome certain limitations, such as assembly errors and artifacts [17]. Here, the full-length transcriptome of *C. arbutifolia* was sequenced on the third-generation PacBio Sequel platform. We obtained 17,397,064 subreads with a mean length of 1482 bp. In contrast, previous reports on second-generation sequencing for *Salix* spp. yielded only 398 bp for *S. suchowensis* [43], 432 bp for *S. babylonica* [43], 703 bp for *S. integra* [44], 713 bp for *S. psammophila* [45], 918 bp for *S. arbutifolia* [28], and 944 bp for *S. matsudana* [46]. For the clustered and polished reads, the sequence quality index N50 was 2086 bp (Table 1). In contrast, N50 was only 1191 bp for *S. integra* [44], 1274 bp for *S. psammophila* [45], 1315 bp for *S. babylonica* [43], 1445 bp for *S. suchowensis* [43], and 1468 bp for *S. matsudana* [46]. Nevertheless, the foregoing values were comparable to the N50 for *S. arbutifolia* (1952 bp) [28]. A comparison against other *Populus* species subjected to the same SMRT technology disclosed that the average subread length for *C. arbutifolia* was shorter than those of *P. wulianensis* (2177 bp) [30] and *P. × canadensis* (2057 bp) [47], possibly because of the diverse genetic backgrounds among these species.

LncRNAs are regulatory factors with important functions in several biological processes, including development, growth, secondary metabolism, and abiotic stress responses. They regulate gene expression by interacting with miRNA networks [48]. Here, a total of 4281 LncRNAs were identified, which is considerably greater than those of *P. alba* var. *pyramidalis* (3410) [31], *P. deltoides* × *P. euramericana* cv. 'Nanlin895' (1187) [29], and *P. × canadensis* (753) [47]. A few studies explored the functions of LncRNAs [47,49]. However, relatively little is known about those of *C. arbutifolia*. TFs play vital roles in regulating gene expression to biotic and abiotic stress responses. Of the 3121 TFs screened in the present study, 2942 were classified into 90 families. The latter number is comparable to that which was reported for *C. oliveri* [19]. The top 10 TF families, including bHLH, AP2-ERF, MYB, C3H, Bzip, C2H2, MYB-related, GRAS, NAC, and WRKY (Table 2), were similar to those reported for *C. oliveri* [19]. The properties of the foregoing TF families associated with plant immunity, regulation, and signaling pathways have been investigated [50].

In genetic and genomic studies, SSRs are the preferred ideal molecular markers. The development of SSR markers through high-throughput sequencing technology is highly efficient, and may be applied to those plants without available genomic resources. In our work, a total of 12,332 candidate SSRs were identified in 50,073 reconstructed genes. The coverage (24.6%) was extremely higher for the SSR markers derived from EST resources by short-read sequencing, such as *S. suchowensis* and *S. babylonica* (4.3%) [51], and *S. psammophila* (7.86%) [45]. The high frequency of occurrence for *C. arbutifolia* may be explained by the use of multiple tissues and advanced sequencing technology. However, the same mining criterion and SMRT technology returned a frequency of occurrence of 26.0% for *P. wulianensis* [30]. Substantial bias among repeat types can be observed for various plants. Hence, di- and tri-nucleotides are common but the other nucleotide types are rare. Tri-nucleotide repeats should be the most abundant because of genetic code selection [51]. Nevertheless, di-nucleotide repeats accounted for 53.8% of the total in *C. arbutifolia* (Table 4). These results are consistent with those for *S. psammophila* [45] but differ from those for *P. wulianensis* [30], *S. suchowensis*, and *S. babylonica* [40]. In the latter cases, tri-nucleotides were the most abundant. Motifs that differed in terms of their repeats were analyzed to disclose any bias in *C. arbutifolia*. For the di- and tri-nucleotides, AG/CT and AAG/CTT were the most frequent, respectively, as previously reported [30,51]. In contrast, GAA/TTC was the dominant motif in *S. psammophila* [45]. Although numerous EST-SSR markers have been developed for *Salix* [43], the SSRs in *C. arbutifolia* still have great potential to provide useful resources in genetic breeding research.

The taxonomically difficult genus *Salix* comprises over 500 species and the phylogenetic position of *C. arbutifolia* is still highly controversial. In the present study, we constructed a phylogenetic tree containing three *Populus* species and two *Salix* species. The construction was based on the alignment results for single-copy ortholog sequences (Figure 5B). *C. arbutifolia* occupied one branch, while another was occupied by *Populus* and *Salix* species. In the early stage, *Chosenia* was reckoned a transitional type during the genetic differentiation from *Populus* to *Salix*, based on certain morphological divergence indices. *Chosenia* was proposed as a separate genus in Salicaceae phylogeny [1–3], and this recommendation was consistent with our study. In contrast, whole cp genome sequences [4,52,53], ribosomal DNA [54,55], and *rbcL* or *matK* gene [55,56] sequences revealed that *C. arbutifolia* is closely related to *Salix* species and should be assigned to the genus *Salix*. As the resources in the database are limited, only five whole-genome sequences for Salicaceae species were downloaded and applied in the phylogeny analysis. Hence, the information returned was restricted and additional species should be used in future studies.

5. Conclusions

The present study is the first full-length transcriptome analysis of *C. arbutifolia*, which, to our knowledge, is classified as an endangered tree species in China. The *C. arbutifolia* transcriptome was sequenced and characterized through SMRT technology. The single-copy and special orthogroups, as well as divergent and conserved genes in *C. arbutifolia* and other five Salicaceae species were also compared and analyzed. Given the limited available information in databases, more species should be added for further analysis. Overall, the full-length transcriptome sequences generated for *C. arbutifolia* will be beneficial for understanding the genomic architecture and providing useful genetic resources for future research on Salicaceae species.

Author Contributions: X.H. and Q.Z. conceived and designed the experiments; X.H. wrote the paper; Y.W. performed the experiments; J.Z. (Jiwei Zheng) and J.Z. (Jie Zhou) analyzed the data; Z.J. and B.W. collected the samples. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 31670662) and the Independent Scientific Research Project of the Jiangsu Academy of Forestry (Grant No. ZZKY202101).

Data Availability Statement: All SMRT sequencing reads obtained for *C. arbutifolia* were deposited to the sequence read archive (SRA) database at NCBI under the accession number PRJNA788330.

Acknowledgments: The authors are grateful to Shanghai BIOZERON Biotechnology Co., Ltd. for their assistance with the bioinformatics analysis.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Kadis, I. *Chosenia*: An amazing tree of Northeast Asia. *Arnoldia* **2005**, *63*, 8–17.
- Nakai, T. *Chosenia*, a new genus of Salicaceae. *Bot. Mag.* **1920**, *34*, 66–69. [[CrossRef](#)]
- Skvortsov, A.K. *Willows of Russia and Adjacent Countries. Taxonomical and Geographical Revision (English Translation with Additions)*; University of Joensuu Faculty of Mathematics and Natural Sciences Report Series; University of Joensuu: Kuopio, Finland, 1999; Volume 39, pp. 1–307.
- Chen, J.H.; Sun, H.; Wen, J.; Yang, Y.P. Molecular phylogeny of *Salix* L. (Salicaceae) inferred from three chloroplast datasets and its systematic implications. *Taxon* **2010**, *59*, 29–37. [[CrossRef](#)]
- Feng, C.H.; He, C.Y.; Wang, Y.; Zeng, Y.F.; Zhang, J.G. Phylogenetic position of *Chosenia arbutifolia* in the Salicaceae inferred from whole chloroplast genome. *For. Res.* **2019**, *32*, 73–77. [[CrossRef](#)]
- Moskalyuk, T.A. *Chosenia arbutifolia* (Salicaceae): Life strategies and introduction perspectives. *Sib. Lesn. Zurnal (Sib. J. For. Sci.)* **2016**, *3*, 34–45. (In English)
- Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63. [[CrossRef](#)]
- Ozsolak, F.; Milos, P.M. RNA sequencing: Advances, challenges and opportunities. *Nat. Rev. Genet.* **2011**, *12*, 87–98. [[CrossRef](#)]
- Rhoads, A.; Au, K.F. PacBio sequencing and its applications. *Genom. Proteom. Bioinf.* **2015**, *13*, 278–289. [[CrossRef](#)]
- Liu, X.; Mei, W.; Soltis, P.S.; Soltis, D.E.; Barbazuk, W.B. Detecting alternatively spliced transcript isoforms from single-molecule long-read sequences without a reference genome. *Mol. Ecol. Resour.* **2017**, *17*, 1243–1256. [[CrossRef](#)]

11. Qiao, D.H.; Yang, C.; Chen, J.; Guo, Y.; Li, Y.; Niu, S.Z.; Cao, K.M.; Chen, Z.W. Comprehensive identification of the full-length transcripts and alternative splicing related to the secondary metabolism pathways in the tea plant (*Camellia sinensis*). *Sci. Rep.* **2019**, *9*, 2709. [[CrossRef](#)]
12. Byrne, A.; Cole, C.; Volden, R.; Vollmers, C. Realizing the potential of full-length transcriptome sequencing. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2019**, *374*, 20190097. [[CrossRef](#)] [[PubMed](#)]
13. Wang, B.; Kumar, V.; Olson, A.; Ware, D. Reviving the transcriptome studies: An insight into the emergence of single-molecule transcriptome sequencing. *Front. Genet.* **2019**, *10*, 384. [[CrossRef](#)] [[PubMed](#)]
14. Qiu, F.Y.; Wang, X.D.; Zheng, Y.J.; Wang, H.; Liu, X.; Su, X. Full-length transcriptome sequencing and different chemotype expression profile analysis of genes related to monoterpenoid biosynthesis in *Cinnamomum porrectum*. *Int. J. Mol. Sci.* **2019**, *20*, 6230. [[CrossRef](#)]
15. Lou, H.Q.; Ding, M.Z.; Wu, J.S.; Zhang, F.; Chen, W.; Yang, Y.; Suo, J.; Yu, W.; Xu, C.; Song, L. Full-length transcriptome analysis of the genes involved in tocopherol biosynthesis in *Torreya grandis*. *J. Agric. Food Chem.* **2019**, *60*, 1877–1888. [[CrossRef](#)]
16. Rao, G.D.; Zhang, J.G.; Liu, X.; Ying, L. Identification of putative genes for polyphenol biosynthesis in olive fruits and leaves using full-length transcriptome sequencing. *Food Chem.* **2019**, *300*, 125246. [[CrossRef](#)]
17. Minio, A.; Massonnet, M.; Figueroa-Balderas, R.; Vondras, A.M.; Blanco-Ulate, B.; Cantu, D. Iso-seq allows genome-independent transcriptome profiling of grape berry development. *G3 Genes Genomes Genet.* **2019**, *9*, 755–767. [[CrossRef](#)]
18. Jia, X.; Tang, L.; Mei, X.; Lui, H.; Luo, H.; Deng, Y.; Su, J. Single-molecule long-read sequencing of the full-length transcriptome of *Rhododendron lapponicum* L. *Sci. Rep.* **2020**, *10*, 6755. [[CrossRef](#)]
19. He, Z.P.; Su, Y.J.; Wang, T. Full-length transcriptome analysis of four different tissues of *Cephalotaxus oliveri*. *Int. J. Mol. Sci.* **2021**, *22*, 787. [[CrossRef](#)]
20. Zhang, H.Y.; Zhang, Y.; Yang, Y.Z. Study on breeding techniques of *Chosenia arbutifolia*. *J. Jilin For. Sci. Technol.* **2005**, *34*, 9–12. (In Chinese)
21. Ma, H.Y.; Cui, K.F.; Huang, B.J.; Chen, Q.H.; Huang, L.F.; Feng, X.C. Situation and protection of rare and endangered species *Chosenia arbutifolia* in Changbai Mountains. *J. Beihua Univ. (Nat. Sci.)* **2015**, *16*, 658–660. (In Chinese)
22. Hoshikawa, T.; Kikuchi, S.; Nagamitsu, T.; Tomaru, N. Eighteen microsatellite loci in *Salix arbutifolia* (Salicaceae) and cross-species amplification in *Salix* and *Populus* species. *Mol. Ecol. Resour.* **2009**, *9*, 1202–1205. [[CrossRef](#)] [[PubMed](#)]
23. Rao, G.D.; Sui, J.K.; Zeng, Y.F.; He, C.; Zhang, J. Genome-wide analysis of the AP2/ERF gene family in *Salix arbutifolia*. *FEBS Open Bio* **2015**, *5*, 132–137. [[CrossRef](#)] [[PubMed](#)]
24. Rao, G.D.; Sui, J.K.; Zhang, J.G. In silico genome-wide analysis of the WRKY gene family in *Salix arbutifolia*. *Plant Omics J.* **2015**, *8*, 353–360. [[CrossRef](#)]
25. Rao, G.D.; Zeng, Y.F.; He, C.Y.; Zhang, J. Characterization and putative posttranslational regulation of α - and β -tubulin gene families in *Salix arbutifolia*. *Sci. Rep.* **2016**, *6*, 19258. [[CrossRef](#)]
26. Nagamitsu, T.; Hoshikawa, T.; Kawahara, T.; Barkalov, V.Y.; Sabirov, R.N. Phylogeography and genetic structure of disjunct *Salix arbutifolia* populations in Japan. *Popul. Ecol.* **2014**, *56*, 539–549. [[CrossRef](#)]
27. Wang, Y.; Jiao, Z.Y.; Zhou, J.; Wang, B.S.; Zhuge, Q.; He, X.D. Population genetic diversity and structure of an endangered Salicaceae species in Northeast China: *Chosenia arbutifolia* (Pall.) A. Skv. *Forests* **2021**, *12*, 1282. [[CrossRef](#)]
28. Rao, G.D.; Zeng, Y.F.; Sui, J.K.; Zhang, J.G. *De novo* transcriptome analysis reveals tissue-specific differences in gene expression in *Salix arbutifolia*. *Trees* **2016**, *30*, 1647–1655. [[CrossRef](#)]
29. Chao, Q.; Gao, Z.F.; Zhang, D.; Zhao, B.G.; Dong, F.Q.; Fu, C.X.; Liu, L.J.; Wang, B.C. The developmental dynamics of the *Populus* stem transcriptome. *Plant Biotechnol. J.* **2019**, *17*, 206–219. [[CrossRef](#)]
30. Wu, Q.C.; Zang, F.Q.; Xie, X.M.; Ma, Y.; Zheng, Y.; Zang, D. Full length transcriptome sequencing analysis and development of EST-SSR markers for the endangered species *Populus wulianensis*. *Sci. Rep.* **2020**, *10*, 16249. [[CrossRef](#)]
31. Hu, H.Y.; Yang, W.L.; Zheng, Z.Y.; Niu, Z.M.; Yang, Y.Z.; Wan, D.S.; Liu, J.Q.; Ma, T. Analysis of alternative splicing and alternative polyadenylation in *Populus alba* var. *pyramidalis* by single-molecular long-read sequencing. *Front. Genet.* **2020**, *11*, 48. [[CrossRef](#)]
32. Li, A.M.; Zhang, J.Y.; Zhou, Z.Y. PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinform.* **2014**, *15*, 311. [[CrossRef](#)]
33. Sun, L.; Luo, H.T.; Bu, D.C.; Zhao, G.G.; Yu, K.T.; Zhang, C.H.; Liu, Y.N.; Chen, R.S.; Zhao, Y. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucl. Acid Res.* **2013**, *41*, e166. [[CrossRef](#)]
34. Kong, L.; Zhang, Y.; Ye, Z.Q.; Liu, X.Q.; Zhao, S.Q.; Wei, L.; Gao, G. CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucl. Acid Res.* **2007**, *36*, W345–W349. [[CrossRef](#)]
35. Finn, R.D.; Coghill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; et al. The Pfam protein families database: Towards a more sustainable future. *Nucl. Acid Res.* **2016**, *44*, D279–D285. [[CrossRef](#)]
36. Zheng, Y.; Jiao, C.; Sun, H.; Rosli, H.G.; Pombo, M.A.; Zhang, P.; Banf, M.; Dai, X.; Martin, G.B.; Giovannoni, J.J.; et al. iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* **2016**, *9*, 1667–1670. [[CrossRef](#)]
37. Shimizu, K.; Adachi, J.; Muraoka, Y. ANGLE: A sequencing errors resistant program for predicting protein coding regions in unfinished cDNA. *J. Bioinform. Comput. Biol.* **2006**, *4*, 649–664. [[CrossRef](#)]

38. Emms, D.M.; Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **2019**, *20*, 238. [[CrossRef](#)]
39. Katoh, K.; Kuma, K.I.; Toh, H.; Miyata, T. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acid Res.* **2005**, *33*, 511–518. [[CrossRef](#)]
40. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree: Computing large minimum-evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **2009**, *26*, 1641–1650. [[CrossRef](#)]
41. Klopfenstein, D.V.; Zhang, L.; Pedersen, B.S.; Tang, H. GOATOOLS: A Python library for gene ontology analyses. *Sci. Rep.* **2018**, *8*, 10872. [[CrossRef](#)]
42. Bu, D.C.; Luo, H.T.; Huo, P.P.; Wang, Z.H.; Zhang, S.; He, Z.H.; Wu, Y.; Zhao, L.H.; Liu, J.J.; Guo, J.C.; et al. KOBAS-i: Intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acid Res.* **2021**, *49*, W317–W325. [[CrossRef](#)]
43. Tian, X.Y.; Zheng, J.W.; Jiao, Z.Y.; Zhou, J.; He, K.Y.; Wang, B.S.; He, X.D. Transcriptome sequencing and EST-SSR marker development in *Salix babylonica* and *S. suchowensis*. *Tree Genet. Genomes* **2019**, *15*, 9. [[CrossRef](#)]
44. Shi, X.; Sun, H.J.; Chen, Y.T. Transcriptome sequencing and expression analysis of cadmium (Cd) transport and detoxification related genes in cd-accumulating *Salix integra*. *Front. Plant Sci.* **2016**, *7*, 1577. [[CrossRef](#)]
45. Jia, H.X.; Yang, H.F.; Sun, P.; Li, J.B.; Guo, Y.H.; Han, X.J.; Zhang, G.S.; Lu, M.J.; Hu, J.J. *De novo* transcriptome assembly, development of EST-SSR markers and population genetic analyses for the desert biomass willow, *Salix psammophila*. *Sci. Rep.* **2016**, *6*, 39591. [[CrossRef](#)]
46. Rao, G.D.; Sui, J.K.; Zeng, Y.F.; He, C.Y.; Zhang, J.G. *De novo* transcriptome and small RNA analysis of two Chinese willow cultivars reveals stress response genes in *Salix matsudana*. *PLoS ONE* **2014**, *9*, e109122. [[CrossRef](#)]
47. Xu, J.H.; Fang, M.; Li, Z.H.; Zhang, M.N.; Liu, X.Y.; Peng, Y.Y.; Wan, Y.L.; Chen, J.H. Third-generation sequencing reveals lncRNA-regulated HSP genes in the *Populus × Canadensis* Moench heat stress response. *Front. Genet.* **2020**, *11*, 249. [[CrossRef](#)]
48. Zhang, G.Y.; Chen, D.G.; Zhang, T.; Duan, A.G.; Zhang, J.G.; He, C.Y. Transcriptomic and functional analyses unveil the role of long non-coding RNAs in anthocyanin biosynthesis during sea buckthorn fruit ripening. *DNA Res.* **2018**, *25*, 465–476. [[CrossRef](#)]
49. Yuan, J.P.; Li, J.R.; Yang, Y.; Tan, C.; Zhu, Y.M.; Hu, L.; Qi, Y.J.; Lu, J.Z. Stress-responsive regulation of long non-coding RNA polyadenylation in *Oryza sativa*. *Plant J.* **2018**, *93*, 814–827. [[CrossRef](#)]
50. Tsuda, K.; Somssich, I.E. Transcriptional networks in plant immunity. *New Phytol.* **2015**, *206*, 932–947. [[CrossRef](#)]
51. He, X.D.; Zheng, J.W.; Zhou, J.; Shi, S.; Wang, B.S. Characterization and comparison of EST-SSRs in *Salix*, *Populus*, and *Eucalyptus*. *Tree Genet. Genomes* **2015**, *11*, 820. [[CrossRef](#)]
52. Chen, Y.N.; Hu, N.; Wu, H.T. Analyzing and characterizing the chloroplast genome of *Salix wilsonii*. *BioMed Res. Int.* **2019**, *2019*, 5190425. [[CrossRef](#)]
53. Lu, D.Y.; Zhang, L.; Zhang, G.S.; Hao, L. Chloroplast genome structure and variation of Salicaceae plants. *J. Northwest A&F U (Nat. Sci. Ed.)* **2020**, *48*, 87–94. [[CrossRef](#)]
54. Leskinen, E.; Alström-Rapaport, C. Molecular phylogeny of Salicaceae and closely related Flacourtiaceae: Evidence from 5.8 S, ITS 1 and ITS 2 of the rDNA. *Plant Syst. Evol.* **1999**, *215*, 209–227. [[CrossRef](#)]
55. Hardig, T.M.; Anttila, C.K.; Brunsfeld, S.J. A phylogenetic analysis of *Salix* (Salicaceae) based on *matK* and ribosomal DNA sequence data. *J. Bot.* **2010**, *2010*, 197696. [[CrossRef](#)]
56. Azuma, T.; Kajita, T.; Yokoyama, J.; Ohashi, H. Phylogenetic relationships of *Salix* (Salicaceae) based on *rbcl* sequence data. *Am. J. Bot.* **2000**, *87*, 67–75. [[CrossRef](#)]