# Comparative Transcriptome and Chloroplast Genome Analyses of Two Related *Dipteronia* Species

Tao Zhou[1], Chen Chen[1†], Yue Wei[1†], Yongxia Chang[1†], Guoqing Bai[1,2], Zhonghu Li[1], Nazish Kanwal[1] and Guifang Zhao[1*]

[1] Key Laboratory of Resource Biology and Biotechnology in Western China (Ministry of Education), College of Life Sciences, Northwest University, Xi'an, China, [2] Shaanxi Engineering Research Centre for Conservation and Utilization of Botanical Resources, Xi'an Botanical Garden of Shaanxi Province, Xi'an, China

*Dipteronia* (order Sapindales) is an endangered genus endemic to China and has two living species, *D. sinensis* and *D. dyeriana*. The plants are closely related to the genus *Acer*, which is also classified in the order Sapindales. Evolutionary studies on *Dipteronia* have been hindered by the paucity of information on their genomes and plastids. Here, we used next generation sequencing to characterize the transcriptomes and complete chloroplast genomes of both *Dipteronia* species. A comparison of the transcriptomes of both species identified a total of 7814 orthologs. Estimation of selection pressures using Ka/Ks ratios showed that only 30 of 5435 orthologous pairs had a ratio significantly >1, i.e., showing positive selection. However, 4041 orthologs had a Ka/Ks < 0.5 ($p < 0.05$), suggesting that most genes had likely undergone purifying selection. Based on orthologous unigenes, 314 single copy nuclear genes (SCNGs) were identified. Through a combination of *de novo* and reference guided assembly, plastid genomes were obtained; that of *D. sinensis* was 157,080 bp and that of *D. dyeriana* was 157,071 bp. Both plastid genomes encoded 87 protein coding genes, 40 tRNAs, and 8 rRNAs; no significant differences were detected in the size, gene content, and organization of the two plastomes. We used the whole chloroplast genomes to determine the phylogeny of *D. sinensis* and *D. dyeriana* and confirmed that the two species were highly divergent. Overall, our study provides comprehensive transcriptomic and chloroplast genomic resources, which will be valuable for future evolutionary studies of *Dipteronia*.

Keywords: *Dipteronia*, transcriptome, positive selection, purifying selection, chloroplast genome, phylogenetic relationship

## INTRODUCTION

*Dipteronia* Oliver (order Sapindales) is an endangered genus endemic to China; it has two living species, *D. sinensis* Oliver and *D. dyeriana* Henry, and is a sister genus of *Acer* (Peng and Thomas, 2008). The genus *Dipteronia* has been documented in the fossil record with specimens found in Tertiary sediments in North America (McClain and Manchester, 2001). Both extant species are perennial woody plants with different natural ranges; *D. sinensis* occupies a relatively extensive range in central and southwestern China, while *D. dyeriana* is located in a limited area of Yunnan province. The latter species is grown as an ornamental species and for oil. Although the two species of *Dipteronia* are allopatric at the present time, they share some morphological similarities

such as leaf shape and fruit characteristics. However, comparatively little is known of the genetic differentiation of the two species or their evolutionary dynamics.

As relic species of the Tertiary period, both species of *Dipteronia* have experienced long complex evolutionary histories to result in their current distributional status. Previous research based on analyses of chloroplast simple sequence repeats (cpSSRs) and amplified fragment length polymorphisms (AFLPs) revealed that significant genetic differences are present between *D. sinensis* and *D. dyeriana*; these analyses also suggested that the populations of *Dipteronia* may have suffered a genetic bottleneck (Yang et al., 2007, 2008). However, a clear understanding of the causes of their genetic divergence and speciation has still not been achieved; this is largely because of the lack of genomic resources. To date, comprehensive genome sequences and complete chloroplast genomes have not been described for either species. Nor has any attempt been made at comparative transcriptomics to identify possible causes of genome divergence and selection in these two species.

Next-generation sequencing (NGS) has greatly advanced our ability to obtain genome resources in non-model species. Transcriptome sequencing (RNA-seq) offers both a convenient means of rapidly obtaining information on expressed genomic regions and also provides an opportunity to resolve comparative genomic-level problems for non-model organisms (Logacheva et al., 2011; Zhang L. et al., 2013). With the advent of NGS, transcriptome sequencing has become more effective. Transcriptome sequencing also provides an alternative method for whole-genome sequencing for use in analyzing adaptive evolution and genetic divergence (Zhang L. et al., 2013; Chen et al., 2015; Mu et al., 2015). For closely related species, comparative transcriptome analyses can not only provide useful genomic resources, such as SSRs and single copy nuclear gene (SCNG) markers, but can also provide insights into speciation and adaptive evolution.

In plants, the chloroplast genome is more conserved than the nuclear genome; it usually has a circular structure of a pair of inverted repeat (IR) regions separated by large single-copy (LSC) and small single-copy (SSC) regions (Bendich, 2004). Because of its conserved nature, many plastid molecular markers have been used to infer phylogeographic history as well as to resolve the phylogenetic relationships of different species. The availability of NGS technology has enabled the generation of large amounts of sequence data at relatively low cost. Thus, it is comparatively simple to obtain comprehensive chloroplast sequences for plant species with this new technology. Sequencing of the complete chloroplast genome has been used in phylogenetic analyses and has proved effective in clarifying difficult phylogenetic relationships (Ma et al., 2014; Carbonell-Caballero et al., 2015). However, until now, only two chloroplast genomes have been reported for the Aceraceae (Yang J. B. et al., 2014; Li Z. H. et al., 2015). Thus, the present study on the chloroplast genomes of *Dipteronia* will provide valuable plastid resources to resolve phylogenetic relationships in *Acer* and *Dipteronia*. Furthermore, the chloroplast genome data will aid development of plastid genetic markers for phylogeographic research in *Dipteronia*.

In the present study, we compared the transcriptomes and chloroplast genomes of the two *Dipteronia* species. We also carried out pairwise comparisons of orthologous sequences from these species to identify candidate genes under positive selection. Our sequencing analysis of the transcriptomes identified a large number of single-copy nuclear gene markers from both species. Additionally, we used the information on the chloroplast genomes to analyze the phylogenetic relationships of species within the order Sapindales. Overall, our study provides new insights into the evolutionary history of the two *Dipteronia* species and has produced resources for further evolutionary studies on *Dipteronia* and related species in the Aceraceae.

## MATERIALS AND METHODS

### Transcriptome Sequence Datasets, *De novo* Assembly, Gene Expression Levels, and Functional Annotation

Two transcriptome datasets (SRR2127986/SRR2127991) from a previous study (Zhou et al., 2016) were used for the present comparative transcriptome analysis. Before assembly, the raw reads were filtered to obtain high-quality clean reads by removing adapters, low-quality sequences (reads with unknown bases "N"), and reads with more than 20% low-quality bases (quality value ≤10). High quality reads were assembled as transcripts using Trinity (r2013-02-25) with default parameters (Grabherr et al., 2011). After assembly, the resultant transcripts were processed by CD-HIT version 4.6 with a sequence identity threshold of 0.95 to remove redundancies (Li and Godzik, 2006). We used RSEM-1.2.29 software (Li and Dewey, 2011) to estimate gene expression levels in each species. First, the clean reads of each species were mapped back onto the transcripts to obtain the read count values of all genes. Then we calculated the fragments per kilobase of transcript per million mapped reads (FPKM), which is the most commonly used method to estimate gene expression levels (Trapnell et al., 2010). For evaluating the function of *D. sinensis* and *D. dyeriana* transcriptome sequences, we separately aligned the unigene sequences of these two species with public protein databases such as the NCBI non-redundant protein database (Nr), Cluster of Orthologous Group (COG), Swiss-Prot, and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database using blastx with an *E*-value threshold of 1E-5. Gene ontology (GO) annotation was performed by Blast2GO software with a cut-off *E*-value of 1E-5 and then plotted with functional classification using Web Gene Ontology Annotation Plot (WEGO) (Conesa et al., 2005; Ye et al., 2006).

### Identification of Orthologous Genes in *D. sinensis* and *D. dyeriana*, Estimation of Substitution Rates, and Mining of Single Copy Nuclear Genes

Open reading frames (ORF) of unigene sequences were predicted by the Getorf program with a minimum length of 150 amino acids (Rice et al., 2000). The predicted coding DNA sequence regions of the *D. sinensis* and *D. dyeriana* transcriptomes were then used to identify orthologous groups between the two species.

OrthoMCL v2.0.9, based on a protein similarity graph method (Li et al., 2003), was employed to retrieve the groups of homologous protein coding genes with the default parameters. InParanoid 7 was also used to search the orthologous groups with the genome of _Theobroma cacao_ as an outgroup (Ostlund et al., 2010). Finally, we compared the results from both methods and orthologs shared between the two methods were retrieved as the orthologous genes of two species. The remaining protein coding genes that could not be assigned to orthologous groups were considered as species-specific expressed genes. The obtained orthologous pairs were aligned and formatted by ParaAT1.0 with default parameters (Zhang Z. et al., 2012). The nonsynonymous (Ka), synonymous (Ks), and Ka/Ks values were calculated using KaKs_Calculator v. 1.2 based on the YN algorithm (Zhang et al., 2006) and Fisher's exact test was performed to justify the validity of the Ka and Ks values. For the purpose of finding SCNGs in _Dipteronia_, the 959 APVO genes (959 SCNGs shared by _Arabidopsis_, _Populus_, _Vitis_, and _Oryza_) were used for our analysis (Duarte et al., 2010). We retrieved the protein sequences encoded by the APVO genes from the TAIR10 database and then queried these sequences against the orthologous genes of _D. sinensis_ and _D. dyeriana_ using BLASTP with a threshold _E_-value of 1E-10. All the queries with hits were considered to be SCNGs in the _Dipteronia_ species.

## The Chloroplast Genome Sequencing, Assembly, and Annotation of _D. sinensis_ and _D. dyeriana_

Total genomic DNA was isolated from leaf tissues using the modified CTAB method (Doyle, 1987). The DNA library was constructed using TruSeq DNA sample preparation kits and then a paired-end library with insert sizes of 200 bp was sequenced using Illumina HiSeq$^{TM}$ 2500 with the average read length of 125 bp. In order to conduct comparative chloroplast genome analyses of two _Dipteronia_ species, the raw Illumina sequencing reads of _D. sinensis_ from our previous study (Zhou et al., 2015) were retrieved in the present study. Illumina raw reads were first quality trimmed using NGS QC Toolkit_v2.3.3 with default cut-off values (Patel and Jain, 2012). After trimming of low quality reads and adapters, the clean reads were assembled using MIRA 4.0.2 (Chevreux et al., 2004) with the chloroplast genome of _Acer buergerianum_ subsp. _ningpoense_ (Yang J. B. et al., 2014) as a reference (parameters: job = genome, mapping, accurate; technology = solexa; segment_placement = FR). Subsequently, the resultant contigs were further assembled using a baiting and iteration method based on Perl script MITObim_1.8.pl (Hahn et al., 2013). After assembly, the obtained contigs were ordered with the reference chloroplast genome of _A. buergerianum_ subsp. _Ningpoense_. The gaps were filled by realignment of input reads using Geneious R8 v 8.0.2 (Biomatters Ltd., Auckland, New Zealand) and some ambiguous regions with low coverage were confirmed by PCR-based Sanger sequencing using primers designed for gap-flanking regions (**Table S9**). Eventually, the complete chloroplast genome was annotated by the online software DOGMA (Wyman et al., 2004) with default parameters and manual adjustment of the start and stop codons in

Geneious R8 v 8.0.2. The annotated GenBank files were used to draw circular plastid genome maps with the online program OrganellarGenome DRAW (OGDRAW) (Lohse et al., 2013).

## Repeat Structure and Sequence Divergence of Chloroplast Genomes

Dispersed and palindromic repeats in each chloroplast genome were identified using REPuter with a minimum repeat size of 30 bp and a sequence identity >90% (Kurtz et al., 2001). The Tandem Repeats Finder program was used to identify tandem repeat sequences with the following parameters: 2 for alignment parameters match, 7 for mismatch and indel, respectively (Benson, 1999). SSR loci in both chloroplast genomes were detected using MISA with the SSR identification parameters of ten for mono-, five for di-, four for tri-, and three for tetra-, penta, and hexa-nucleotide motifs. Construction of multiple alignments of complete cpDNA sequences was carried out by the mVISTA comparative genomics tool with the annotation of _A. buergerianum_ subsp. _Ningpoense_ as reference (Frazer et al., 2004). The percentages of variable characters for each coding and noncoding regions were calculated as described in a previous study of Poaceae species (Zhang et al., 2011). In order to detect whether selective pressure exists for plastid genes, we calculated the nonsynonymous (Ka), synonymous (Ks), and Ka/Ks values of each protein coding gene in the two chloroplast genomes.

## Phylogenetic Analyses

The phylogeny of the _Dipteronia_ species was investigated using the complete plastid genomes of species in the order Sapindales, including _A. buergerianum_ subsp. _Ningpoense_ (KF753631), _Acer morrisonense_ (KT970611), _Sapindus mukorossi_ (KM454982), _Citrus aurantifolia_ (NC_024929), _Citrus sinensis_ (NC_008334), _Azadirachta indica_ (NC_023792), and _Zanthoxylum piperitum_ (NC_0279390); these sequences were downloaded as ingroup taxa. _Populus trichocarpa_ (NC_009143) and _T. cacao_ (HQ244500) were used as outgroup taxa. The complete chloroplast genomes with one IR region removed were aligned by MAFFT v7.017 software with default parameters (Katoh and Standley, 2013) and then the sequences were manually adjusted using ClustalX (Larkin et al., 2007). The choice of substitution model for each partition was primarily determined using Modeltest 3.7 (Posada and Crandall, 1998) with the Akaike information criterion (AIC) (Posada and Buckley, 2004). Phylogenetic analysis was conducted based on the maximum likelihood (ML) method using RAxML version v 7.2.8 (Stamatakis, 2006). The ML tree was constructed with a combined rapid bootstrap of 1000 replicates and a search for the best tree in a single run under the GTR + G model. In parallel, phylogeny was also inferred from the plastid genomes using MrBayes v 3.1.2 (Ronquist and Huelsenbeck, 2003) with the TVM + I +G model. The Markov chain Monte Carlo (MCMC) algorithm was run for one million generations with trees sampled very 100 generations. Convergence of the parallel runs was determined by examining the average standard deviation of split frequencies, which fell below 0.01. The first 25% of trees generated were discarded as burn-in and the remaining trees were used to build a

majority-rule consensus tree. The ML and Bayesian analyses were separately conducted based on the three plastid genomic regions (LSC, IR, and SSC).

## RESULTS

### *De novo* Assembly and Annotation of the *Dipteronia* Transcriptome

Using Trinity software, short reads were assembled to generate transcripts, which were further clustered to obtain unigenes. A total of 91,340 transcripts (N50 = 1777 bp, average length = 1055 bp) and 52,351 unigenes (N50 = 1351 bp, average length = 749 bp) were recovered for *D. sinensis*. For *D. dyeriana*, 101,628 transcripts (N50 = 2071 bp, average length = 1248 bp) and 53,983 unigenes (N50 = 1519 bp, average length = 809 bp) were obtained (**Table 1**). After calculating the FPKM values, our results showed 154/104 unigenes (*D. sinensis*/*D. dyeriana*) with FPKM values >500 (**Table S1**). To annotate the *D. sinensis* and *D. dyeriana* sequences, searches were conducted against the Nr, Swiss-Prot, COG, KEGG, and GO databases. There were 30,834 unigenes (58.9%) for *D. sinensis* and 27,796 (51.5%) for *D. dyeriana* with at least one significant match to the above databases (**Table 2**). For Nr annotation of both species, a BLASTX top-hit species distribution showed highest homology to *T. cacao* (8049 hits in *D. sinensis*/7863 hits in *D. dyeriana*) followed by *Vitis vinifera* (3901/3850) and *P. trichocarpa* (2741/2606). GO terms were assigned to 25,591 annotated sequences from *D. sinensis* and 23,003 annotated sequences from *D. dyeriana*. The annotated sequences belonged to three GO categories: "cellular component," "molecular function," and "biological process" (**Figure 1**). We found that the assigned gene functions were similarly distributed in both species. In the "cellular component" category, "cell" (20.1%/21.1%) and "cell part" (21.2%/21.3%) was prominent, while in the "molecular function" category "binding" (43.4%/44.0%) and "catalytic activity" (36.7%/36.8%) were overrepresented. In the "biological process" category, "cellular process" (14.2%/14.3%) was most representative followed by "metabolic process" (13.9%/14.1%). All of the *D. sinensis* and *D. dyeriana* unigenes were subjected to functional prediction and classification using the COG database. The unigenes were assigned to 25 COG categories (**Figure 2**). The category "cluster of general function" represented the largest group (18.6%/18.7%) in both species. The next most represented category was "translation, ribosomal structure, and biogenesis" for *D. sinensis* (8.8%), while for *D. dyeriana*, "replication, recombination and repair" was the next most represented category (9.4%). Only a few unigenes in both species were assigned into the "nuclear structure" category (4 genes for *D. sinensis* and 1 gene for *D. dyeriana*) and no genes were found in either species in the "extracellular structures" category. To identify the biological pathways of these two species, the annotated unigene sequences were mapped to reference pathways in the KEGG database. The results showed that 7182 unigenes from *D. sinensis* mapped to 120 pathways and 6225 *D. dyeriana* unigenes mapped to 118 pathways. Interestingly, the representative pathways were "ribosome" (673

**TABLE 1 | Summary of statistics for the transcriptomes of *D. sinensis* and *D. dyeriana*.**

|  | *D. sinensis* | *D. dyeriana* |
|---|---|---|
| Total number of reads | 40,615,432 | 53,620,610 |
| Total number of transcripts | 91,340 | 101,628 |
| Total number of unigenes | 52,351 | 53,983 |
| Min length length of unigenes (bp) | 201 | 201 |
| Max length length of unigenes (bp) | 14,265 | 14,906 |
| N50 of unigenes (bp) | 1351 | 1519 |
| Mean length of unigenes (bp) | 749 | 809 |
| Mapping rates of unigenes[a] | 74.6% | 76.8% |

[a]Mapping rates were generated by mapping clean reads to the assembled unigenes using the Bowtie mode of RSEM 1.2.29.

**TABLE 2 | Annotation information of *D. sinensis* and *D. dyeriana*.**

|  | *D. sinensis* | | *D. dyeriana* | |
|---|---|---|---|---|
|  | Number | Percentage (%) | Number | Percentage (%) |
| COG | 10,637 | 20.32 | 9411 | 17.43 |
| GO | 25,591 | 48.88 | 23,003 | 42.61 |
| KEGG | 7182 | 13.72 | 6225 | 11.53 |
| Swiss-Prot | 23,321 | 44.55 | 20,936 | 38.78 |
| Nr | 30,689 | 58.62 | 27,738 | 51.38 |
| All | 30,834 | 58.90 | 27,796 | 51.49 |

genes/564 genes, ko03010), "oxidative phosphorylation" (300 genes/252 genes, ko00190), and "glycolysis/gluconeogenesis" (281 genes/269 genes, ko00010) in both species (**Figure S1**). We also searched the highly expressed unigenes in the GO annotation results and found that many of them were involved in functions related to environmental adaption such as "response to salt stress," "response to cadmium ion," "defense response," "response to water deprivation," "response to high light intensity."

### Putative Orthologs, Substitution Rates, and Single Copy Nuclear Genes in *D. sinensis* and *D. dyeriana*

By utilizing OrthoMCL and InParanoid methods, we obtained an initial set of 9480 and 9190 putative orthologous pairs in *D. sinensis* and *D. dyeriana*, respectively. After comparing the results of both approaches, 7814 orthologs pairs were found to be common to both methods and were used in subsequent analyses. Synonymous (Ks) and nonsynonymous (Ka) substitution rates were calculated for the orthologous unigene pairs. We excluded orthologous pairs that only had either synonymous or nonsynonymous substitutions; this step left 7699 orthologous unigene pairs that could be used for the calculation of Ka/Ks ratios. In order to avoid paralogs in our analyses, we excluded candidate orthologs with a synonymous (Ks) substitution value >0.1, as these may be paralogs (Zhang J. et al., 2013). Finally, a total of 5435 orthologous pairs was selected and used to calculate Ka/Ks ratios (**Table S2**). Of these
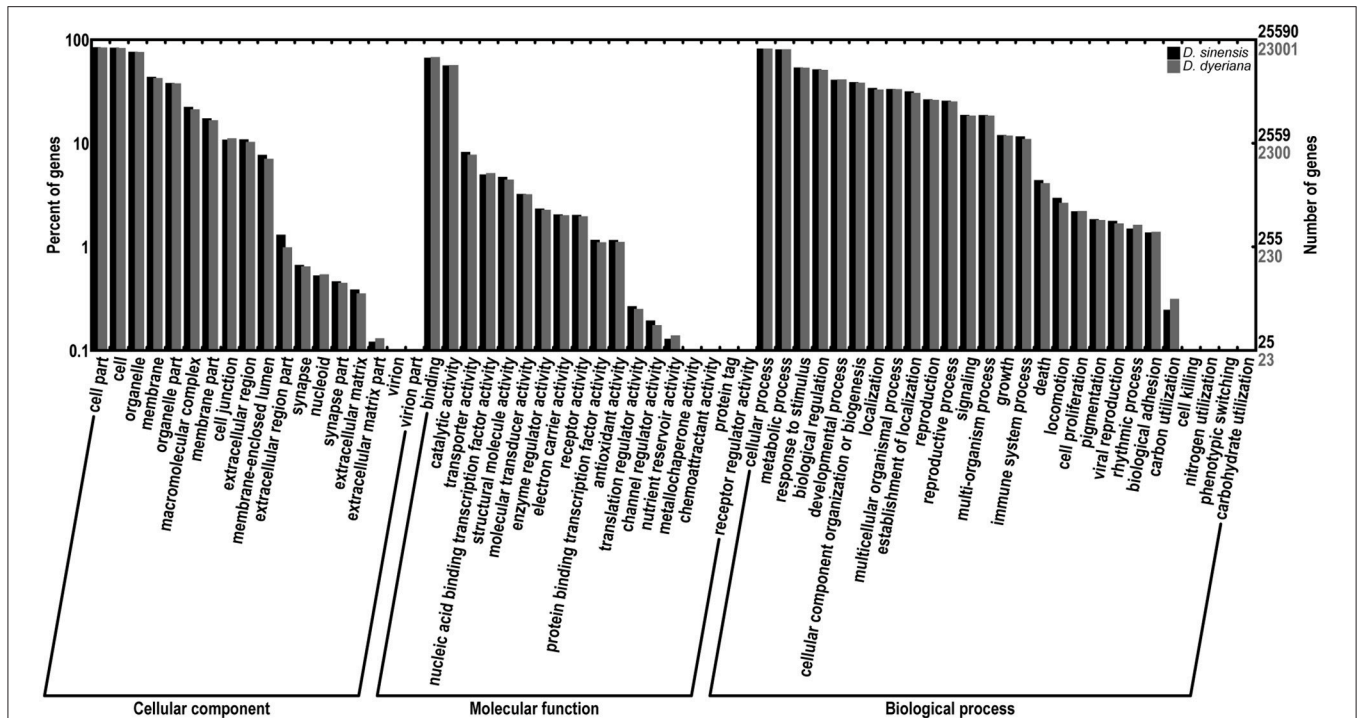
**FIGURE 1 | Comparison of gene ontology (GO) terms distributions between *D. sinensis* and *D. dyeriana* transcriptome.** GO terms were annotated according to three main categories (biological process, cellular component, molecular function) and 63 sub-categories.
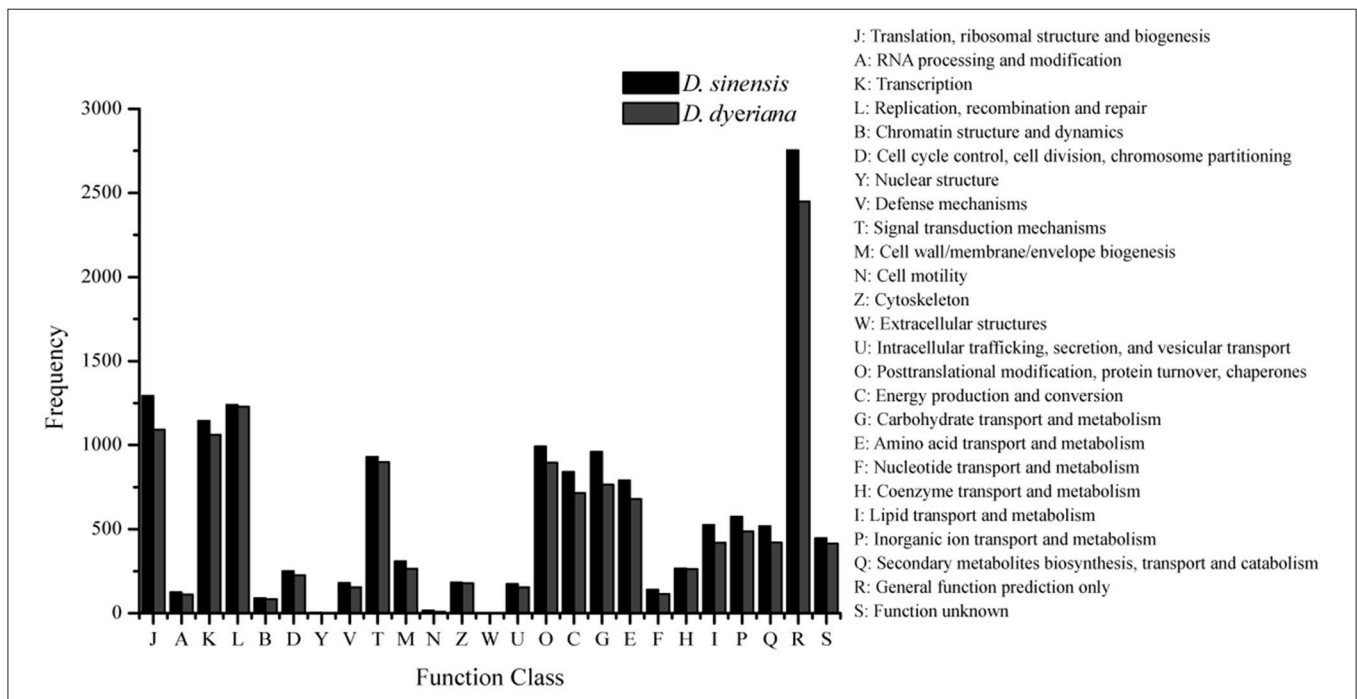


**FIGURE 2 | Clusters of orthologous group (COG) classifications for *D. sinensis* and *D. dyeriana* transcriptome.** All unigenes were aligned to COG database to predict and classify possible functions.

orthologs, 283 had a Ka/Ks ratio >1 indicating positive selection, and 857 had a Ka/Ks ratio between 0.5 and 1, indicating weak purifying selection. The annotation information of orthologs which showed a Ka/Ks ratio significantly >1 ($p < 0.05$) indicated that some of these genes were related to "abiotic and biotic stress response," "metabolism," and "enzyme" (**Table S3**). Using

the APVO gene sets (Duarte et al., 2010) to implement BLASTP queries against the 7814 orthologs. Three hundred and fourteen of the APVO genes were found to give hits against orthologous unigenes between *D. sinensis* and *D. dyeriana*; these are most likely the SCNGs of *Dipteronia* species. A total of 54 pairs of orthologs were extracted with more than 600 bp length and >75% identity to *Arabidopsis thaliana* peptide sequences (**Table S4**).

## Chloroplast Genome Sequencing, Assembly, and Annotation

Illumina pair-end sequencing produced 25,566,606 and 29,304,216 raw reads with a sequence length of 125 bp for *D. sinensis* and *D. dyeriana*, respectively. The total length of the reads was approximately 7.38 Gb for *D. sinensis* and 6.3 Gb for *D. dyeriana*. After quality trimming of the raw reads, 25,562,204 and 29,221,800 clean reads were collected for *D. sinensis* and *D. dyeriana*, respectively. Based on a combination of *de novo* and reference guided assembly, the complete plastid nucleotide sequences for the two species were recovered. The final chloroplast genome sequences have been deposited in GenBank (Accession numbers: KT878501 and KT985457). The *D. sinensis* and *D. dyeriana* chloroplast genomes were composed of 157,080 bp and 157,071 bp, respectively (**Table 3**). After annotation, a total of 135 unique genes included 87 protein coding genes, 40 tRNAs, and 8 rRNA operons were obtained for both species (**Table S5**). The gene map of both species is shown in **Figure 3**.

## Comparative Analyses of Chloroplast Genomes of *Dipteronia*

Both chloroplast genomes exhibited a typical quadripartite structure, consisting of a pair of IRs (26,766 bp in *D. sinensis*/26,730 bp in *D. dyeriana*) separated by an LSC (85,455 bp/85,529 bp) and an SSC (18,093 bp/18,082 bp); there was no significant difference in the lengths of the three regions in the two species. The two chloroplast genomes shared identical complements of genes with similar orders. The GC content of *D. sinensis* and *D. dyeriana* were similar (37.8%/38.0%) (**Table 3**). The two genomes encode an identical set of 135 genes and 19 are duplicated in the IR regions (**Table 3**). Of these 135 genes, 15 genes (*rpl2, ndhB, trnI-GAU, trnA-UGC, ndhA, rpl16, petD, petB, trnV-UAC, trnL-UAA, rpoC1, atpF, trnG, rps16, trnK-UUU*) harbored one intron and three genes (*clpP, rps12, ycf3*) harbored two introns (**Table S5**). Two genes (*infA, rps2*) were inferred to be pseudogenes in *A. buergerianum* subsp. *Ningpoense* (Yang J. B. et al., 2014). The sequence identity of the two *Dipteronia* chloroplast genomes was plotted with mVISTA software (**Figure 4**). The whole aligned chloroplast genome sequences indicated that they were relatively conserved in the two *Dipteronia* species and *A. buergerianum*, although some highly divergent regions were found. Similar to most plant species, the chloroplast gene coding regions were more conserved than those of their noncoding counterparts. According to the alignment results, several intergenic regions were found to display high divergence, including *trnS(GCU)-trnG*, *trnT(UGU)-rps4*, *trnL(UAA)-trnT(UGU)*, *psbE-petL*, and *rpl32-trnL(UAG)*. Additionally, we found that the level of variation in the noncoding regions (1.96%) was 2.5-fold greater than that in the coding regions (0.79%) and that the IRs and coding regions were more conserved than single copy and noncoding regions, respectively (**Figure S2**).

Analyses of repeat sequences in the genomes using the REPuter program showed that the characteristics of repeat sequences were similar in the two genomes: 27 repeats were >30 bp in *D. sinensis* and 28 repeats were >30 bp in *D. dyeriana*. Using the Tandem Repeats Finder program, 11 and 15 tandem repeats were identified in *D. sinensis* and *D. dyeriana*, respectively (**Tables S6, S7**). Most of the repeats were distributed in intergenic (IGS) or intronic regions; a few were located in genic regions (*psaA, psaB*, rps2, rps19, *ycf1, ycf2, trnS-GCU, trnS-UGA, trnS-GGA*) (**Tables S6, S7**). A total of 118 and 80 microsatellite loci were detected in *D. sinensis* and *D. dyeriana* chloroplast genomes, respectively. The most abundant repeat type in both genomes was mononucleotide repeats (**Figure S3**). In order to investigate the evolutionary characteristics of cpDNA genes, nonsynonymous (Ka) and synonymous substitution rates (Ks), and the ratio Ka/Ks were calculated for the 87 individual protein coding genes in the two species. The Ka values ranged from 0 to 0.08, the Ks values ranged from 0.007 to 0.03, and most Ka/Ks ratios were less than 1, suggesting that cpDNA genes were under purifying selection. Only four genes (*rpl32, rpl22, rpl33, cemA*) had Ka/Ks ≥ 1 indicating that they had undergone positive selection or neutral selection (**Table S8**).

## Phylogenetic Analyses Based on the Complete Chloroplast Genome

The plastid genomes (with one IR region removed) of 11 species, including *D. sinensis* and *D. dyeriana*, were used to construct a phylogenetic tree. The data set comprised of 152,721 nucleotide

**TABLE 3 | Summary of two complete chloroplast genomes of *Dipteronia*.**

|  | D. sinensis | D. dyeriana |
|---|---|---|
| Total cp DNA size (bp) | 157,080 | 157,071 |
| Length of large single copy (LSC) region (bp) | 85,455 | 85,529 |
| Length of inverted repeat (IR) region (bp) | 26,766 | 26,730 |
| Length of small single copy (SSC) region (bp) | 18,093 | 18,082 |
| Total GC content (%) | 37.8 | 38.0 |
| LSC | 35.9 | 36.1 |
| IR | 42.7 | 42.8 |
| SSC | 32.1 | 32.5 |
| Total number of genes | 135 | 135 |
| Protein encoding[a] | 87 (8) | 87 (8) |
| tRNA[a] | 40 (7) | 40 (7) |
| rRNA[a] | 8 (4) | 8 (4) |

[a]*The numbers in parenthesis indicate the genes duplicated in the IR regions.*

positions with 10,179 informative sites for the ingroup taxa. However, there were only 458 informative sites for the four aceraceous species. ML analyses resulted in a fully resolved tree with 9 of the 10 nodes supported by 100% bootstrap values; all the species of Aceraceae formed a monophyletic clade (**Figure 5**). With respect to the Bayesian analysis, the identical topology was obtained with a posterior probability of 1.0. ML and Bayesian analyses were separately conducted using the LSC, IR, and SSC genomic regions; these analyses yielded an identical topology with all aceraceous species in a monophyletic clade (**Figures S4–S6**). The two *Dipteronia* species did not cluster in the same clade except when the SSC region was used to construct the phylogenetic tree, indicating that there is considerable divergence between *D. sinensis* and *D. dyeriana*.

## DISCUSSION

### Transcriptome Sequencing, *De novo* Assembly, and Annotation for *D. sinensis* and *D. dyeriana*

Illumina-based transcriptome sequencing has been proven to be an efficient and cost-effective way to retrieve transcriptome data. Recently, many assembled transcriptomes of non-model species have been obtained and employed for studies of differential gene expression, genetic marker development (Huang et al., 2015), and phylogenomic analysis (Yang X. et al., 2014), as well as for detecting selection and inferring adaptive evolution in closely related species (Chen et al., 2015; Guo et al., 2016). To date, however, most transcriptome studies have been carried out on single species. Here, 40.6 million and 53.6 million clean reads were assembled into 52,351 unigenes with a mean length 749 bp for *D. sinensis* and 53,983 unigenes with a mean length of 809 bp for *D. dyeriana*. These results are comparable to those reported previously using the same technology (Li S. S. et al., 2015; Rong et al., 2016). Therefore, the transcriptome datasets produced in the present study will boost the previously meager genomic resources for Aceraceae species.
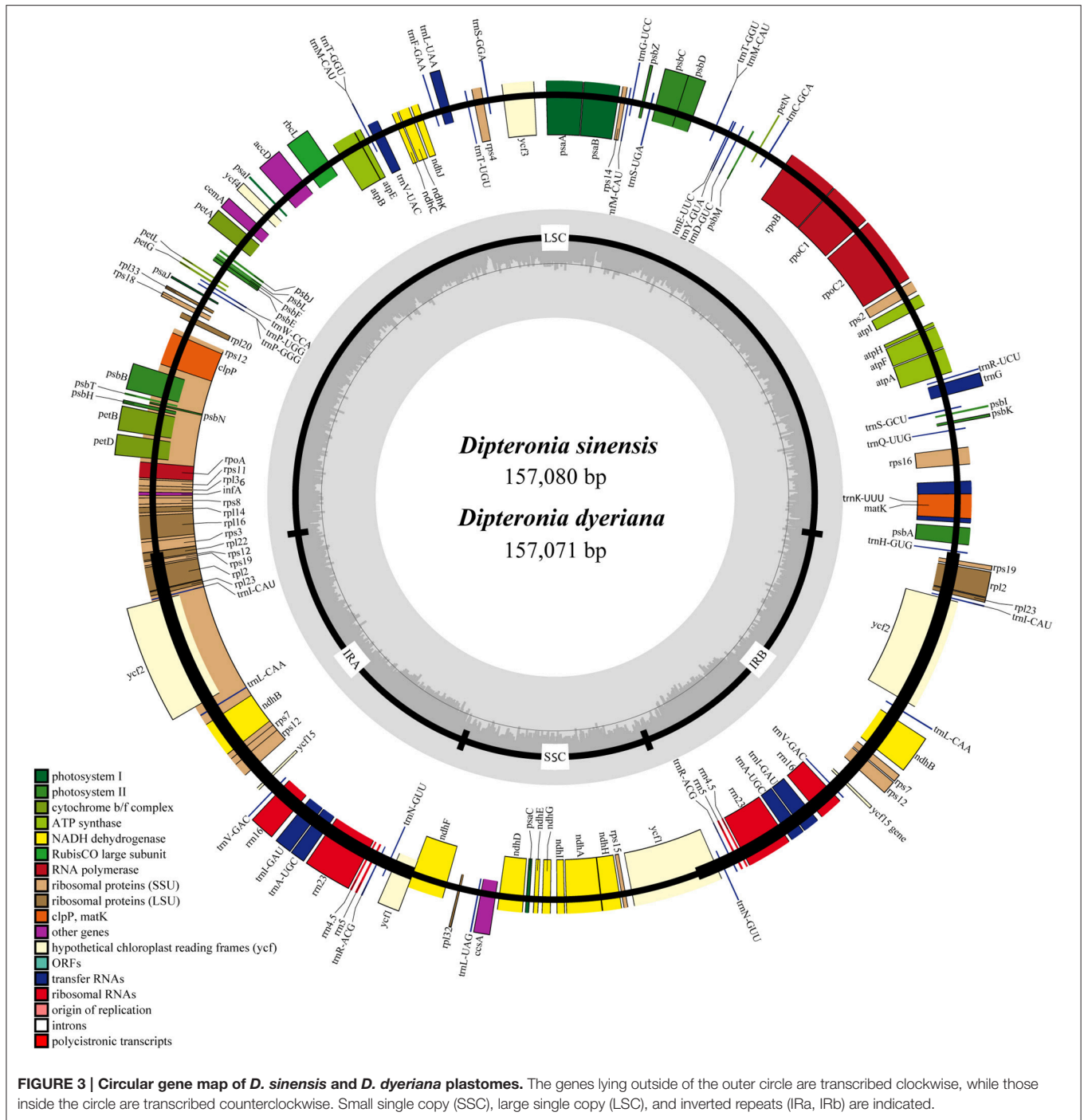
More than half of the unigenes of both species (58.9%/51.5%) could be annotated using five public protein databases and most involved plant proteins. However, a significant number of unigenes had no BLAST hits to these databases. This may be because there are no comprehensive genomic resources for *Dipteronia* and also because of the lack of a reference genome for Aceraceae; these unigene sequences might therefore represent novel transcripts. Comparative analyses of the functional annotation for the two species showed that they had a similar distribution of functional categories in different protein databases. This may be due in part to use of the same tissues from both species for transcriptome sequencing; alternatively, there may be no significant differences in the protein coding genes of the two species. Intriguingly, a higher number of unigenes were obtained for *D. dyeriana*, although the number of annotated unigenes for *D. sinensis* was greater than for *D. dyeriana*. This difference suggested that the unigene sequences of *D. dyeriana* might include a greater proportion of novel transcripts. Highly expressed genes in both species did not show identical functions,

although most of these genes were involved in functions related to environmental adaption. We presume that the different habitat preferences of the two species stimulated this genetic divergence.

### Orthologous Genes, Substitution Rates, and SCNGs Markers in *D. sinensis* and *D. dyeriana*

Ka/Ks values are widely used to distinguish protein coding genes under positive or purifying selection (Hurst, 2002). Orthologs under positive selection contain interesting candidate genes that are usually related to "abiotic and biotic stress response," "biosynthesis," and "metabolism and enzyme" (Zhao et al., 2013). In the present study, 5435 orthologous pairs were analyzed and 30 were found to have a Ka/Ks ratio significantly >1; some of these orthologs were related to the above-mentioned functions such as "response to stress" (GO:0006950), "response to salt stress" (GO:0009651), "metabolic process" (GO:0008152), and "oxidative stress" (GO:0034599). We thus deduced that such genes suffered significant positive selection during evolution; these results are in line with those reported in previous studies on non-model species (Zhang J. et al., 2013; Zhang L. et al., 2013; Zhao et al., 2013). One orthologous pair was found to be a member of the subtilase protein family which is involved in seed coat development (GO:0048359) (Rautengarten et al., 2008). Therefore, we infer that these genes were also under significant positive selection and would result in differences in seed characters in the two species. Additionally, some orthologs were detected and annotated with a function in response to UV (GO:0071492). *D. dyeriana* is generally found in locations at comparatively high altitudes; we speculated that this species is subject to more intense ultraviolet light exposure that might affect expression of genes related to UV response. The remaining 5151 orthologous pairs had a Ka/Ks < 1; 4041 orthologs had a Ka/Ks < 0.5 ($p$ < 0.05), suggesting that most genes are likely to undergo purifying selection with stronger selective constraints for nonsynonymous changes than for synonymous ones (Tiffin and Hahn, 2002). If a Ka/Ks ratio >0.5 is considered an indicator of positive selection, as in previous studies (Swanson et al., 2004), then 1140 pairs with a Ka/Ks ratio between 0.5 and 1 were detected. This indicates a large number of orthologous pairs in *D. sinensis* and *D. dyeriana* with a relatively high Ka/Ks value. One factor that increases Ka/Ks value as well as weakening the strength of purifying selection is a decrease in the effective population size (Fay and Wu, 2003). Both *D. sinensis* and *D. dyeriana* are listed as endangered Tertiary relic species. Thus, in our study, reduced effective population sizes may have contributed to the relatively high Ka/Ks ratios.

Previous studies described genetic markers, such as SSRs, in *Dipteronia* (Chen et al., 2011; Su et al., 2012; Zhou et al., 2016) but no SCNGs markers have been developed. SCNGs with heterogeneous rates of variation are generally thought to provide a higher level of discrimination than chloroplast and nuclear ribosomal (nrDNA) spacer sequences (Salas-Leiva et al., 2014; Mao et al., 2016). Recently, single copy or low copy nuclear genes have been increasingly used to clarify phylogenetic relationship in some angiosperms and to determine the dynamics of speciation

**FIGURE 3 | Circular gene map of *D. sinensis* and *D. dyeriana* plastomes.** The genes lying outside of the outer circle are transcribed clockwise, while those inside the circle are transcribed counterclockwise. Small single copy (SSC), large single copy (LSC), and inverted repeats (IRa, IRb) are indicated.

(Curto et al., 2012; Zhang N. et al., 2012; Du et al., 2015; Guo et al., 2015). Until now, only nrDNA and chloroplast markers have been used to probe phylogenetic relationships between *Dipteronia* and related genera (Yang et al., 2010). The large number of SCNGs developed in the present study will contribute substantially to the elucidation of phylogenetic relationships and to investigation of population demographic history in *Dipteronia* and Aceraceae species.

## Comparative Analyses of Complete Chloroplast Genome Sequences

The present study produced complete chloroplast genomes for each of the *Dipteronia* species using Illumina sequencing technology. Apart from the plastid genomes of *A. buergerianum* subsp. *Ningpoense* and *A. morrisonense*, no published chloroplast genomes have been reported for Aceraceae. Therefore, our determination of the whole plastid genomes for the two
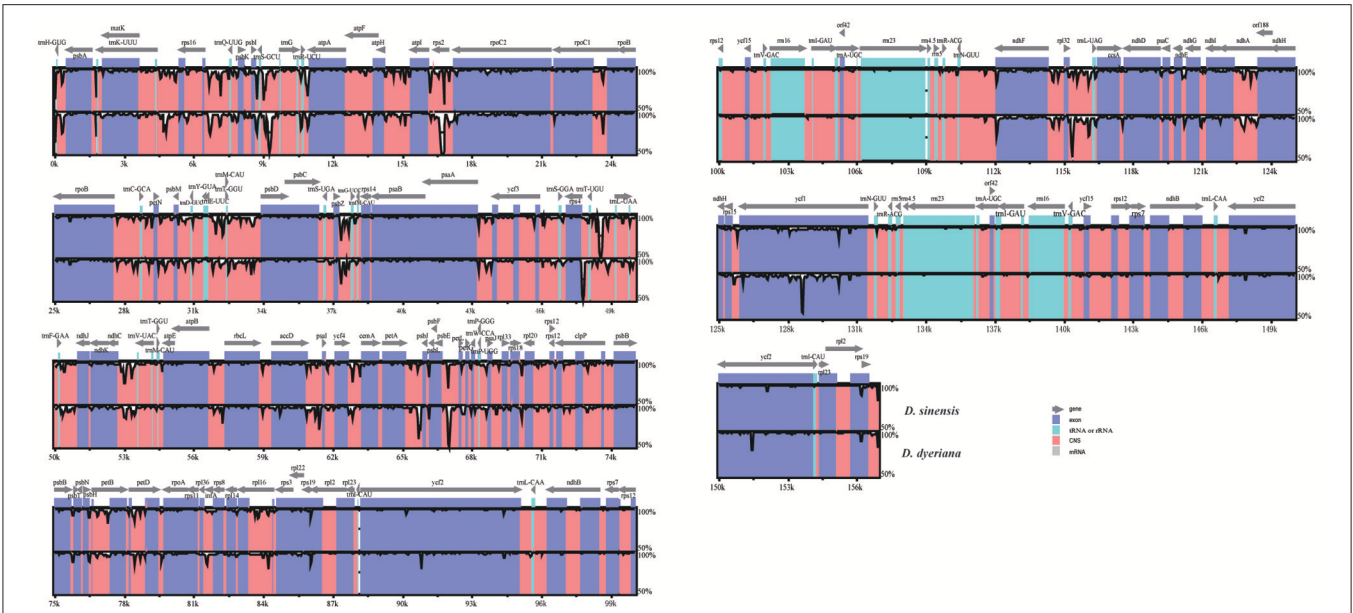
**FIGURE 4 | mVISTA percent identity plot comparing the two *Dipteronia* plastid genomes with *Acer buergerianum* subsp. *Ningpoense* as a reference.** The top line shows genes in order (transcriptional direction indicated by arrows). The sequence similarity of the aligned regions between *Dipteronia* species and *Acer buergerianum* subsp. *Ningpoense* is shown as horizontal bars indicating the average percent identity between 50 and 100% (shown on the y-axis of the graph). The x-axis represents the coordinate in the chloroplast genome. Genome regions are color coded as protein-coding (exon), tRNA or rRNA, and conserved noncoding sequences (CNS).
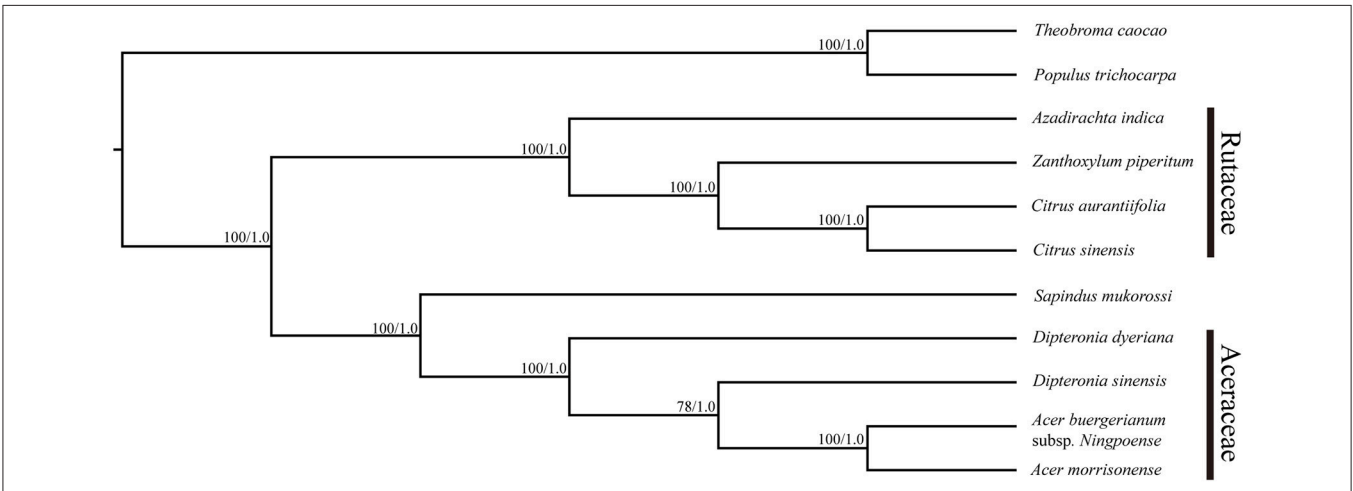


**FIGURE 5 | Maximum likelihood phylogeny of the nine Sapindales species based on the complete plastid genome sequences.** The numbers associated with the nodes are bootstrap support and posterior probability values.

*Dipteronia* species will be a significant aid to filling in the gap in our knowledge of plastid genome evolution in *Dipteronia* and *Acer* species. The two plastid genomes described here possess the typical angiosperm quadripartite structure with two short inverted repeat regions separated by two single copy regions. The size, gene content, and organization of the plastomes of *Dipteronia* are similar to that of *A. buergerianum* subsp. *Ningpoense* and no significant structural rearrangements, such as inversions or gene relocations, were detected. The chloroplast

genomes of both species in this study were relatively well conserved, and most variations were detected in intergenic regions; a similar effect was seen in two other species of Aceraceae (**Figure 5**). One of the aims of this type of study is to identify genomic "barcodes"; these are DNA sequences with a sufficiently high mutation rate to identify a species within a given taxonomic group (Li X. et al., 2015). Here, we found highly variable regions in *accD, rpl33, rpl22, psaC, rps16/trnQ-UGG, trnS(GCU)/trnG-GCC,* and *trnL-UAA/trnF-GAA;* this variation may be sufficient

to suggest these are candidate gene regions for developing more specific DNA barcodes for the Aceraceae family. Such variable markers could also be used to further clarify phylogenetic relationships in aceraceous plants.

As repeat elements are correlated with plastome rearrangement (Weng et al., 2013), we decided to investigate the large, tandem, dispersed, and palindromic repeat sequences in the plastomes of *Dipteronia*. We identified a similarly low number of repeats in the two chloroplast genomes; these repeats were usually located in the same genes (*ycf1, ycf2*) or in genes with similar functions (*psaA, psaB*; *trnS-GCU, trnS-UGA, trnS-GGA*) in both species. Low numbers of repeats have also been found in other species of Geraniaceae and Chloridoideae (Weng et al., 2013; Rousseau-Gueutin et al., 2015). Additionally, SSRs were also distributed similarly in two chloroplast genomes and most of these were located in the same regions of both genomes. For protein coding genes in both species, sequence divergence was evaluated by comparing the synonymous (Ks) substitution rates; all of the genes showed a low sequence divergence (Ks < 0.1) except for *psaC* (Ks = 0.114). For all protein coding genes, most Ka/Ks value were <1 which indicated that most chloroplast genes were under purifying selection; this is consistent with previous studies (Rousseau-Gueutin et al., 2015; Xu et al., 2015). Only three genes (*rpl32, rpl22, cemA*) had a Ka/Ks ratio >1 as expected of genes under positive selection. Of these genes, *rpl32* and *rpl22* encoded ribosomal proteins. A previous study also found that ribosomal proteins have more divergent protein sequences than genes for photosynthesis (Xu et al., 2015). Interestingly, the *cemA* gene is related to the PPR7 protein. We speculated that *cemA* may have coevolved with nuclear genes (Jalal et al., 2015).

## The Phylogenetic Position of *Dipteronia* Chloroplast Genome Sequences

Plastid genomes have been proven to be effective in resolving difficult phylogenetic relationships (Ma et al., 2014; Carbonell-Caballero et al., 2015). In the present study, 11 complete chloroplast genomes of five taxa were used to resolve the still-debated phylogenetic position of *Dipteronia* species (Yang et al., 2010). In our analyses, all the species of Aceraceae formed a monophyletic clade with a high-resolution value and clustered with *S. mukorossi* (Sapindaceae) in the same clade. This result is compatible with the proposal that the *Dipteronia-Acer* clade is a subfamily (Aceroideae) or lower rank within the Sapindaceae (McClain and Manchester, 2001). Although traditional plant taxonomy considers *Dipteronia* and *Acer* as sister taxa, *D. sinensis* and *D. dyeriana* were not clustered into a monophyletic clade and did not show a paraphyletic relationship with *Acer* in the current study. Both the BI and ML analyses showed coincident topology based on different plastid regions (except for the SSC) and this was used to construct the phylogeny in which *D. sinensis* and *D. dyeriana* were clustered into a monophyletic clade. The phylogenetic trees based on complete chloroplast genomes and three different plastid regions in this study indicated that *D. sinensis* is usually in parallel with *Acer* but not with *D. dyeriana*, as was suggested in a previous study

(Yang et al., 2010). This significant discrepancy in phylogenetic placement of *D. dyeriana* should be interpreted with caution. First, only a few chloroplast gene fragments were utilized in the previous study to construct the phylogenetic relationships. Since phylogenomics has been proven a robust method for tackling difficult phylogenies, the results of the present study may therefore provide a more reliable conclusion (Bewick et al., 2012; Zhou et al., 2012; Ma et al., 2014; Carbonell-Caballero et al., 2015). Second, as tertiary species with allopatric distribution ranges have undergone a long-term complex evolutionary history, involving different geological and climate events over a long period, this may be the cause of the high genetic divergence between *D. sinensis* and *D. dyeriana*. Finally, as *D. sinensis* is present in a wide range of natural habitats and is sometimes located in the same areas as *Acer* plants, it is possible that there might have been hybridization events between *D. sinensis* and *Acer* species during the evolutionary process, which may have significantly affected its phylogenetic position. Determination of whether *D. sinensis* is always in parallel with *Acer* will require analysis of more *Acer* chloroplast genomes in future. Overall, our analysis of chloroplast genomes has provided a valuable resource for future work on the phylogenetics of *Dipteronia* species.

## AUTHOR CONTRIBUTIONS

GZ and TZ conceived and designed the experiments. TZ, CC, YW, and YC performed the experiments and analyzed the data. GB prepared the samples. TZ wrote the paper. ZL and NK help to revise the paper. All authors read and approved the final manuscript.

## ACKNOWLEDGMENTS

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2016.01512

**Table S1 | Gene expression level of two *Dipteronia* species.**

**Table S2 | Ka, Ks values, and Ka/Ks ratio between orthologs in two *Dipteronia* species.**

**Table S3 | Annotation results of 30 orthologous pairs with Ka/Ks > 1 (*p* < 0.05).**

**Table S4 | The list of 55 candidate single copy nuclear genes homologous to APVO genes.**

**Table S5 | List of genes present in *Dipteronia* chloroplast genome.**

**Table S6 | The repeats distribution in *D. sinensis* chloroplast genome.**

**Table S7 | The repeats distribution in *D. dyeriana* chloroplast genome.**

**Table S8 | Ka/Ks ratio between pairwise of species protein coding sequences in two *Dipteronia* species.**

**Table S9 | List of primer pairs used in sequence verification and improvement of the *Dipteronia* chloroplast genome.**

**Figure S1 | Kyoto Encyclopedia of Genes and Genomes (KEGG) classification between *D. sinensis* and *D. dyeriana* transcriptome.**

**Figure S2 | Percentage of variable characters in aligned two *Dipteronia* chloroplast genomes. (A)** Coding region. **(B)** Noncoding region. These regions are oriented according to their locations in the chloroplast genome.

**Figure S3 | Frequency distribution of the SSRs identified in *Dipteronia* plastid genomes.**

**Figure S4 | Maximum likelihood phylogeny of the nine Sapindales species based on the large single copy (LSC) region sequences.** The numbers associated with the nodes are bootstrap support and posterior probability values.

**Figure S5 | Maximum likelihood phylogeny of the nine Sapindales species based on the inverted repeat A (IRa) region sequences.** The numbers associated with the nodes are bootstrap support and posterior probability values.

**Figure S6 | Maximum likelihood phylogeny of the nine Sapindales species based on the small single copy (SSC) region sequences.** The numbers associated with the nodes are bootstrap support and posterior probability values.

# REFERENCES

Bendich, A. J. (2004). Circular chloroplast chromosomes: the grand illusion. *Plant Cell* 16, 1661–1666. doi: 10.1105/tpc.160771

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27:573. doi: 10.1093/nar/27.2.573

Bewick, A. J., Chain, F. J. J., Heled, J., and Evans, B. J. (2012). The pipid root. *Syst. Biol.* 61, 913–926. doi: 10.1093/sysbio/sys039

Carbonell-Caballero, J., Alonso, R., Ibañez, V., Terol, J., Talon, M., and Dopazo, J. (2015). A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus *Citrus. Mol. Biol. Evol.* 32, 2015–2035. doi: 10.1093/molbev/msv082

Chen, C., Ren, B. B., Xu, X. H., Fu, C. X., and Qiu, Y. X. (2011). Isolation and characterization of microsatellite markers for *Dipteronia dyerana* (Sapindaceae), an endangered endemic species in China. *Am. J. Bot.* 98, e271–e273. doi: 10.3732/ajb.1100185

Chen, L.-Y., Zhao, S.-Y., Wang, Q.-F., and Moody, M. L. (2015). Transcriptome sequencing of three *Ranunculus* species (Ranunculaceae) reveals candidate genes in adaptation from terrestrial to aquatic habitats. *Sci. Rep.* 5:10098. doi: 10.1038/srep10098

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Müller, W. E., Wetter, T., et al. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* 14, 1147–1159. doi: 10.1101/gr.1917404

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

Curto, M. A., Puppo, P., Ferreira, D., Nogueira, M., and Meimberg, H. (2012). Development of phylogenetic markers from single-copy nuclear genes for multi locus, species level analyses in the mint family (Lamiaceae). *Mol. Phylogenet. Evol.* 63, 758–767. doi: 10.1016/j.ympev.2012.02.010

Doyle, J. J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Du, S., Wang, Z., Ingvarsson, P. K., Wang, D., Wang, J., Wu, Z., et al. (2015). Multilocus analysis of nucleotide variation and speciation in three closely related *Populus* (Salicaceae) species. *Mol. Ecol.* 24, 4994–5005. doi: 10.1111/mec.13368

Duarte, J. M., Wall, P. K., Edger, P. P., Landherr, L. L., Ma, H., Pires, J. C., et al. (2010). Identification of shared single copy nuclear genes in *Arabidopsis, Populus, Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evol. Biol.* 10:61. doi: 10.1186/1471-2148-10-61

Fay, J. C., and Wu, C.-I. (2003). Sequence divergence, functional constraint, and selection in protein evolution. *Annu. Rev. Genomics Hum. Genet.* 4, 213–235. doi: 10.1146/annurev.genom.4.020303.162528

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279. doi: 10.1093/nar/gkh458

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Guo, J., Liu, R., Huang, L., Zheng, X.-M., Liu, P.-L., Du, Y.-S., et al. (2016). Widespread and adaptive alterations in genome-wide gene expression associated with ecological divergence of two *Oryza* species. *Mol. Biol. Evol.* 33, 62–78. doi: 10.1093/molbev/msv196

Guo, Y.-Y., Luo, Y.-B., Liu, Z.-J., and Wang, X.-Q. (2015). Reticulate evolution and sea-level fluctuations together drove species diversification of slipper orchids (*Paphiopedilum*) in South-East Asia. *Mol. Ecol.* 24, 2838–2855. doi: 10.1111/mec.13189

Hahn, C., Bachmann, L., and Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* 41:e129. doi: 10.1093/nar/gkt371

Huang, L. K., Yan, H. D., Zhao, X. X., Zhang, X. Q., Wang, J., Frazier, T., et al. (2015). Identifying differentially expressed genes under heat stress and developing molecular markers in orchardgrass (*Dactylis glomerata* L.) through transcriptome analysis. *Mol. Ecol. Resour.* 15, 1497–1509. doi: 10.1111/1755-0998.12418

Hurst, L. D. (2002). The *Ka/Ks* ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18, 486–487. doi: 10.1016/S0168-9525(02)02722-1

Jalal, A., Schwarz, C., Schmitz-Linneweber, C., Vallon, O., Nickelsen, J., and Bohne, A.-V. (2015). A small multifunctional pentatricopeptide repeat protein in the chloroplast of *Chlamydomonas reinhardtii. Mol. Plant* 8, 412–426. doi: 10.1016/j.molp.2014.11.019

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi: 10.1093/nar/29.22.4633

Larkin, M. A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

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

Li, S. S., Li, Q. Z., Rong, L. P., Tang, L., Wang, J. J., and Zhang, B. (2015). Analysis of the transcriptome of green and mutant golden-yellow leaves of *Acer palmatum* Thunb. using high-throughput RNA sequencing. *J. Hortic. Sci. Biotechnol.* 90, 388–394. doi: 10.1080/14620316.2015.11513199

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., and Chen, S. (2015). Plant DNA barcoding: from gene to genome. *Biol. Rev.* 90, 157–166. doi: 10.1111/brv.12104

Li, Z. H., Xie, Y. S., Zhou, T., Jia, Y., He, Y. L., and Yang, J. (2015). The complete chloroplast genome sequence of *Acer morrisonense* (Aceraceae). *Mitochondrial DNA*. doi: 10.3109/19401736.2015.1118091. [Epub ahead of print].

Logacheva, M., Kasianov, A., Vinogradov, D., Samigullin, T., Gelfand, M., Makeev, V., et al. (2011). *De novo* sequencing and characterization of floral

transcriptome in two species of buckwheat (*Fagopyrum*). *BMC Genomics* 12:30. doi: 10.1186/1471-2164-12-30

Lohse, M., Drechsel, O., Kahlau, S., and Bock, R. (2013). OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* 41, W575–W581. doi: 10.1093/nar/gkt289

Ma, P. F., Zhang, Y. X., Zeng, C. X., Guo, Z. H., and Li, D. Z. (2014). Chloroplast phylogenomic analyses resolve deep-level relationships of an intractable bamboo tribe Arundinarieae (Poaceae). *Syst. Biol.* 63, 933–950. doi: 10.1093/sysbio/syu054

Mao, Y., Zhang, Y., Xu, C., and Qiu, Y. (2016). Comparative transcriptome resources of two *Dysosma* species (Berberidaceae) and molecular evolution of the CYP719A gene in Podophylloideae. *Mol. Ecol. Resour.* 16, 228–241. doi: 10.1111/1755-0998.12415

McClain, A. M., and Manchester, S. R. (2001). *Dipteronia* (Sapindaceae) from the Tertiary of North America and implications for the phytogeographic history of the Aceroideae. *Am. J. Bot.* 88, 1316–1325. doi: 10.2307/3558343

Mu, X., Hou, G., Song, H., Xu, P., Luo, D., Gu, D., et al. (2015). Transcriptome analysis between invasive *Pomacea canaliculata* and indigenous *Cipangopaludina cahayensis* reveals genomic divergence and diagnostic microsatellite/SSR markers. *BMC Genet.* 16:12. doi: 10.1186/s12863-015-0175-2

Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D. N., Roopra, S., et al. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38, D196–D203. doi: 10.1093/nar/gkp931

Patel, R. K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE* 7:e30619. doi: 10.1371/journal.pone.0030619

Peng, H., and Thomas, W. (2008). *Flora of China*, Vol. 11. Missouri: Missouri Botanical Garden Press.

Posada, D., and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* 53, 793–808. doi: 10.1080/10635150490522304

Posada, D., and Crandall, K. A. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics* 14, 817–818. doi: 10.1093/bioinformatics/14.9.817

Rautengarten, C., Usadel, B., Neumetzler, L., Hartmann, J., Büssis, D., and Altmann, T. (2008). A subtilisin-like serine protease essential for mucilage release from *Arabidopsis* seed coats. *Plant J.* 54, 466–480. doi: 10.1111/j.1365-313X.2008.03437.x

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/S0168-9525(00)02024-2

Rong, L., Li, Q., Li, S., Tang, L., and Wen, J. (2016). *De novo* transcriptome sequencing of *Acer palmatum* and comprehensive analysis of differentially expressed genes under salt stress in two contrasting genotypes. *Mol. Genet. Genomics* 291, 575–586. doi: 10.1007/s00438-015-1127-2

Ronquist, F., and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. doi: 10.1093/bioinformatics/btg180

Rousseau-Gueutin, M., Bellot, S., Martin, G. E., Boutte, J., Chelaifa, H., Lima, O., et al. (2015). The chloroplast genome of the hexaploid *Spartina maritima* (Poaceae, Chloridoideae): comparative analyses and molecular dating. *Mol. Phylogenet. Evol.* 93, 5–16. doi: 10.1016/j.ympev.2015.06.013

Salas-Leiva, D. E., Meerow, A. W., Francisco-Ortega, J., Calonje, M., Griffith, M. P., Stevenson, D. W., et al. (2014). Conserved genetic regions across angiosperms as tools to develop single-copy nuclear markers in gymnosperms: an example using cycads. *Mol. Ecol. Resour.* 14, 831–845. doi: 10.1111/1755-0998.12228

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446

Su, H. L., Li, Z. H., Zhao, P., Bai, G.-Q., Zhou, T. H., Liu, Z. L., et al. (2012). Isolation and characterization of polymorphic microsatellite loci in the endangered plant *Dipteronia sinensis* (Sapindaceae). *Am. J. Bot.* 99, e425–e427. doi: 10.3732/ajb.1200151

Swanson, W. J., Wong, A., Wolfner, M. F., and Aquadro, C. F. (2004). Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. *Genetics* 168, 1457–1465. doi: 10.1534/genetics.104.030478

Tiffin, P., and Hahn, M. W. (2002). Coding sequence divergence between two closely related plant species: *Arabidopsis thaliana* and *Brassica rapa* ssp. pekinensis. *J. Mol. Evol.* 54, 746–753. doi: 10.1007/s0023901-0074-1

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Weng, M.-L., Blazier, J. C., Govindu, M., and Jansen, R. K. (2013). Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats and nucleotide substitution rates. *Mol. Biol. Evol.* 31, 645–659. doi: 10.1093/molbev/mst257

Wyman, S. K., Jansen, R. K., and Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20, 3252–3255. doi: 10.1093/bioinformatics/bth352

Xu, J. H., Liu, Q., Hu, W., Wang, T., Xue, Q., and Messing, J. (2015). Dynamics of chloroplast genomes in green plants. *Genomics* 106, 221–231. doi: 10.1016/j.ygeno.2015.07.004

Yang, J., Li, S., Sun, G., Yuan, Y., and Zhao, G. (2008). Population structure and genetic variation in the genus *Dipteronia* Oliv.(Aceraceae) endemic to China as revealed by cpSSR analysis. *Plant Syst. Evol.* 272, 97–106. doi: 10.1007/s00606-007-0641-z

Yang, J., Qian, Z.-Q., Liu, Z.-L., Li, S., Sun, G.-L., and Zhao, G.-F. (2007). Genetic diversity and geographical differentiation of *Dipteronia* Oliv.(Aceraceae) endemic to China as revealed by AFLP analysis. *Biochem. Syst. Ecol.* 35, 593–599. doi: 10.1016/j.bse.2007.03.022

Yang, J., Wang, X. M., Li, S., and Zhao, G. F. (2010). What is the phylogenetic placement of *Dipteronia dyerana* Henry? An example of plant species placement based on nucleotide sequences. *Plant Biosyst.* 144, 634–643. doi: 10.1080/11263504.2010.490032

Yang, J. B., Li, D. Z., and Li, H. T. (2014). Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. *Mol. Ecol. Resour.* 14, 1024–1031. doi: 10.1111/1755-0998.12251

Yang, X., Cheng, Y. F., Deng, C., Ma, Y., Wang, Z. W., Chen, X. H., et al. (2014). Comparative transcriptome analysis of eggplant (*Solanum melongena* L.) and turkey berry (*Solanum torvum* Sw.): phylogenomics and disease resistance analysis. *BMC Genomics* 15:412. doi: 10.1186/1471-2164-15-412

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., et al. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* 34, W293–W297. doi: 10.1093/nar/gkl031

Zhang, J., Xie, P., Lascoux, M., Meagher, T. R., and Liu, J. (2013). Rapidly evolving genes and stress adaptation of two desert poplars, *Populus euphratica* and *P. pruinosa*. *PLoS ONE* 8:e66370. doi: 10.1371/journal.pone.0066370

Zhang, L., Yan, H. F., Wu, W., Yu, H., and Ge, X. J. (2013). Comparative transcriptome analysis and marker development of two closely related Primrose species (*Primula poissonii* and *Primula wilsonii*). *BMC Genomics* 14:329. doi: 10.1186/1471-2164-14-329

Zhang, N., Zeng, L., Shan, H., and Ma, H. (2012). Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. *New Phytol.* 195, 923–937. doi: 10.1111/j.1469-8137.2012.04212.x

Zhang, Y. J., Ma, P. F., and Li, D. Z. (2011). High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). *PLoS ONE* 6:e20596. doi: 10.1371/journal.pone.0020596

Zhang, Z., Li, J., Zhao, X. Q., Wang, J., Wong, G. K. S., and Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* 4, 259–263. doi: 10.1016/S1672-0229(07)60007-2

Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X., et al. (2012). ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* 419, 779–781. doi: 10.1016/j.bbrc.2012.02.101

Zhao, L., Zhang, N., Ma, P. F., Liu, Q., Li, D. Z., and Guo, Z. H. (2013). Phylogenomic analyses of nuclear genes reveal the evolutionary relationships within the BEP clade and the evidence of positive selection in Poaceae. *PLoS ONE* 8:e64642. doi: 10.1371/journal.pone.0064642

Zhou, T., Li, Z. H., Bai, G. Q., Feng, L., Chen, C., Wei, Y., et al. (2016). Transcriptome sequencing and development of genic SSR markers of an endangered Chinese endemic genus *Dipteronia* Oliver (Aceraceae). *Molecules* 21, 166. doi: 10.3390/molecules21030166

Zhou, T., Zhao, J. X., Yang, Y.-C., Bai, G. Q., Chen, C., and Zhao, G. F. (2015). The complete chloroplast genome of *Dipteronia sinensis* (Aceraceae), an endangered endemic species to China. *Mitochondrial DNA.* doi: 10.3109/19401736.2015.1111352. [Epub ahead of print].

Zhou, X., Xu, S., Xu, J., Chen, B., Zhou, K., and Yang, G. (2012). Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the Laurasiatherian mammals. *Syst. Biol.* 61, 150–164. doi: 10.1093/sysbio/syr089

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.