Characterization of Global Transcriptome Using Illumina Paired-End Sequencing and Development of EST-SSR Markers in Two Species of *Gynostemma* (*Cucurbitaceae*)

Yue-Mei Zhao [1,2], Tao Zhou [1], Zhong-Hu Li [1] and Gui-Fang Zhao [1,*]

[1] Key Laboratory of Resource Biology and Biotechnology in Western China (Ministry of Education), College of Life Sciences, Northwest University, 229 Taibai Bei Road, Xi'an 710069, China; yezi19820320@163.com (Y.-M.Z.); woody196@163.com (T.Z.); lizhonghu@nwu.edu.cn (Z.-H.L.)

[2] College of Biopharmaceutical and Food Engineering, Shangluo University, Beixin Street, Shangluo 726000, China

* Correspondence: gfzhao@nwu.edu.cn; Tel.: +86-29-8830-5207

Abstract: *Gynostemma pentaphyllum* is an important medicinal herb of the *Cucurbitaceae* family, but limited genomic data have hindered genetic studies. In this study, transcriptomes of two closely-related *Gynostemma* species, *Gynostemma cardiospermum* and *G. pentaphyllum*, were sequenced using Illumina paired-end sequencing technology. A total of 71,607 nonredundant unigenes were assembled. Of these unigenes, 60.45% (43,288) were annotated based on sequence similarity search with known proteins. A total of 11,059 unigenes were identified in the Kyoto Encyclopedia of Genes and Genomes Pathway (KEGG) database. A total of 3891 simple sequence repeats (SSRs) were detected in 3526 nonredundant unigenes, 2596 primer pairs were designed and 360 of them were randomly selected for validation. Of these, 268 primer pairs yielded clear products among six *G. pentaphyllum* samples. Thirty polymorphic SSR markers were used to test polymorphism and transferability in *Gynostemma*. Finally, 15 SSR makers that amplified in all 12 *Gynostemma* species were used to assess genetic diversity. Our results generated a comprehensive sequence resource for *Gynostemma* research.

Keywords: *Gynostemma*; transcriptome; illumina paired-end sequencing; *de novo* assembly; EST-SSR

1. Introduction

Gynostemma (*Cucurbitaceae*), is a genus of perennial creeping herbs with both sexual reproduction and clonal growth by rhizomes or bulbils [1]. This genus has around 16 species and two varieties, distributed in forests, scrubs and bush habitats at 60–3200 m elevations throughout China, India, Myanmar, Korea and Japan [2]. Drainage areas of the Yangtze River in Southwest China's Yunnan Province are thought to be the modern distribution center of this genus. *Gynostemma* can be divided into two subgenera (*Gynostemma* and *Triostellum*) according to different fruit morphology [2]. In recent years, *Gynostemma* has been attracting attention since they contain saponins, amino acids, and reducing sugars, which could be commercially useful. For example, *Gynostemma pentaphyllum*, widespread as a traditional Chinese medicinal herb, is thought to inhibit tumor cell growth, anti-ulceration and to enhance immunity [3,4]. Approximately 84 dammarane-type saponin glycosides were found in *G. pentaphyllum* [5], some of them have structural similarities to glycosides found in *Panax ginseng* C.A. Mey [6], but different saponin contents were observed in various other species of *Gynostemma* [7]. Natural populations of *Gynostemma* were destroyed in recent years due to

excessive harvesting, especially *G. pentaphyllum,* which has been listed as a Grade II Key Protected Wild Plant Species by the Chinese Government [8]. It is necessary to preserve natural stocks of *Gynostemma* spp., and assess their genetic diversity and differentiation. Molecular genetic research of *Gynostemma* is limited [9] because most studies have mainly focused on extracting bioactive components [5,10–13]. Subramaniyam *et al.* (2011) [14] reported *de novo* transcriptome assembly of *G. pentaphyllum* with Roche platforms using the materials of leaves and roots, but the research focused on the identification of secondary metabolite genes. So far, only 14 genomic simple sequence repeats (SSRs) and 14 inter-simple sequence repeats (ISSRs) have been exploited in *Gynostemma* [15,16]. Thus, more markers are needed to better understand the genetic diversity and to develop conservation strategies for *Gynostemma*.

SSR markers have turned out to be an effective tool for germplasm characterization and genetic diversity studies. SSRs can be divided into two categories based on the original sequences used for development of SSRs: genomic SSRs and expressed sequence tag (EST)-SSRs. Developing genomic SSR markers from random genomic sequences is labor, money and time intensive [17,18]. On the contrary, EST-SSRs identified from transcribed RNA sequences are more conserved than noncoding sequences. EST-SSRs are becoming more and more widespread, not only because they are potentially linked with particular transcriptional regions that contribute to agronomic phenotypes [19,20], but also because they have high transferability among closely-related species [21–26]. With the development of next-generation sequencing (NGS), it has become possible to generate large numbers of transcriptomic datasets for nonmodel organisms [27] using various platforms such as Roche 454, Illumina HiSeq, and Applied Biosystems SOLiD. Obtaining large numbers of valuable EST sequences via NGS is important for gene annotation and discovery [28,29], comparative genomics [30], development of molecular markers [31,32], and population genomics studies of genetic variation linked to adaptive traits [33]. Recently, an increasing number of EST datasets have become available for model and non-model organisms, but only a limited number of *Gynostemma* EST sequences are available in the public database.

In this study, we describe the generation, *de novo* assembly, and annotation of a transcriptome-derived EST dataset using Illumina paired-end sequencing technology from two *Gynostemma* species, *G. pentaphyllum* and *G. cardiospermum*. In addition, we mined and validated a large set of EST-SSR markers and investigated the genetic relationship among 12 selected species. This EST datasets will serve as a valuable genomic resource for further studies in *Gynostemma*, e.g., novel gene discovery and marker-assisted selective breeding.

2. Results

2.1. Assembly of Gynostemma Transcriptome Data from Illumina Sequencing

After stringent quality assessment and data filtering, Illumina HiSeq™ 2000 sequencing generated 43,175,448 high-quality reads for *Gynostemma pentaphyllum* and 52,782,146 high-quality reads for *Gynostemma cardiospermum*, respectively. The raw data were deposited in the NCBI Sequence Read Archive (SRA) under the accession number SRA305674. Using the Trinity assembler software [34], short-read sequences from *G. pentaphyllum* and *G. cardiospermum* were assembled *de novo* into 1,488,035 contigs and 1,911,378 contigs, respectively. Transcriptome reads and assembled contigs information for two *Gynostemma* species are shown in Table 1. The frequency distribution of contigs length from these two *Gynostemma* species showed little difference, except for 100–200 bp, which showed more contigs in *G. cardiospermum* (Figure 1). Using paired end-joining, gap-filling, and Trinity, these contigs were assembled into scaffolds which were further assembled into unigenes. Finally, we obtained 40,257 and 44,000 unigenes from *G. pentaphyllum* and *G. cardiospermum*, respectively (Table 2). In addition, we obtained a nonredundant set of unigene sequences by pooling contigs from the two species and assembling them together into 71,607 unigenes. As shown in Table 2, the 71,607 nonredundant unigenes were used for *in silico* mining and validation of genic-SSR markers.

Among the combined 71,607 nonredundant sequences, there were 51,250 (71.57%) that ranged from 200 to 1000 bp, 12,852 (17.95%) from 1000 to 2000 bp, and 7505 (10.48%) greater than 2000 bp.

Table 1. Transcriptome reads and assembled contigs information for two *Gynostemma* species.

Species	Total Reads	Total Clean Nucleotides (Nt)	Q30 Percentage	GC Percentage	Total Number of Contigs	Total Length of Contigs (Nt)	N50 of Contigs	Mean (Nt)
G. pentaphyllum	43,175,448	4,360,277,191	80.16%	43.55%	1,488,035	110,745,998	71	74
G. cardiospermum	52,782,146	5,330,256,148	82.69%	44.00%	1,911,378	136,726,651	66	71

Table 2. Summary of the unigenes from two *Gynostemma* species.

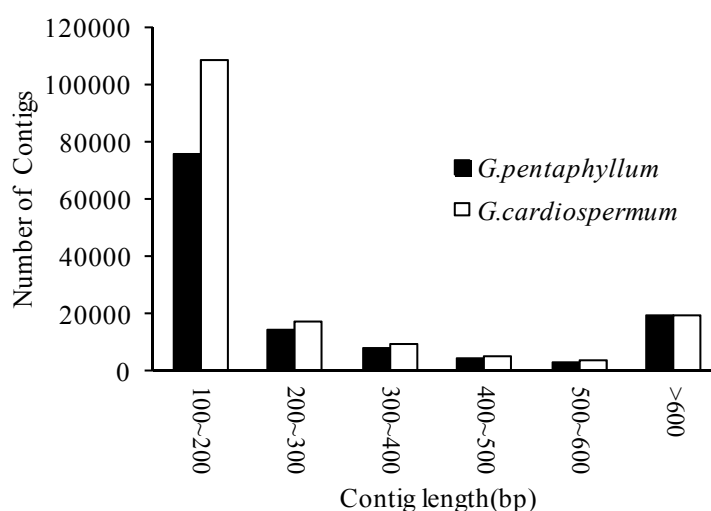| Unigene Source	Total Number of Unigenes	Total Length of Unigenes (Nt)	Mean Length of Unigenes (Nt)	N50 of Unigenes
G. pentaphyllum	40,257	35,161,843	873.43	1516
G. cardiospermum	44,000	37,234,004	846.23	1504
All	71,607	61,367,129	857.00	1535

Figure 1. Frequency distribution of the contig sizes from two *Gynostemma* species. The frequency distribution of contig sizes resulting from Illumina HiSeq™ 2000 sequencing, as assembled using Trinity.

2.2. Functional Annotation and Classification

A homology-based approach was conducted for validation and annotation of assembled unigenes. Among 71,607 unigenes, 60.28% (43,167) showed homology in the nonredundant (nr) database, and 46.04% (32,970) unigenes had BLAST hits in Swiss-Prot database. A total of 60.45% (43,288) unigenes were successfully annotated in the nr and/or Swiss-Prot databases. Additionally, 97.73% (19,895 of 20,357) of the unigenes over 1000 bp in length showed homologous matches, whereas only 32.33% (6619 of 20,474) of the unigenes shorter than 300 bp showed matches (Figure 2). The unigenes homologous to known protein sequences in nr database were further assigned to gene ontology (GO) terms using Blast2GO. A total of 35,968 unigenes were assigned to 549,570 GO term annotations, which belonged to biological processes, molecular functions and cellular components clusters (Figure 3). Among biological processes, "cellular process" was the most dominant group, followed by "metabolic process", "response to stimulus", and "biological regulation". Regarding

the molecular functions category, the major GO terms were "binding" and "catalytic activity". Under the cellular components category, "cell part" and "cell" represented the most abundant classification, followed by "organelle" and "organelle part". All unigenes were searched against the COG database to predict possible functions and phylogenetically classify orthologous gene products. Out of 43,167 nr hits, 20,585 sequences were assigned to one or more COG classifications (Figure 4). Among the 25 COG categories, the cluster for "general function prediction" was the largest group, followed by "translation ribosomal structure and biogenesis", "replication, recombination and repair", and "posttranslational modification, protein turnover, chaperones". In contrast, only a few unigenes were assigned to "nuclear structure and extracellular structure". According to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, 11,059 unigenes were identified with pathway annotation and were assigned to 117 KEGG pathways (Table S1). The top 20 pathways, including 5443 unigenes, are listed in Figure 5. The most highly represented pathways were "Ribosome", followed by "Protein processing in endoplasmic reticulum" and "Plant hormone signal transduction". Being important medicinal plants used in China, previous research on *Gynostemma* spp. has mostly focused on saponins biosynthesis pathways. As expected, some key genes encoding enzymes related to the synthesis of triterpene compounds, which are important components of saponins, were revealed in metabolism pathway. For example, the genes involved in mevalonate (MVA) and 2-*C*-methyl-D-erythritol 4-phosphate (MEP) pathways were identified in our study (Table 3) and these genes may provide valuable resources for research on gynosaponin biosynthesis.
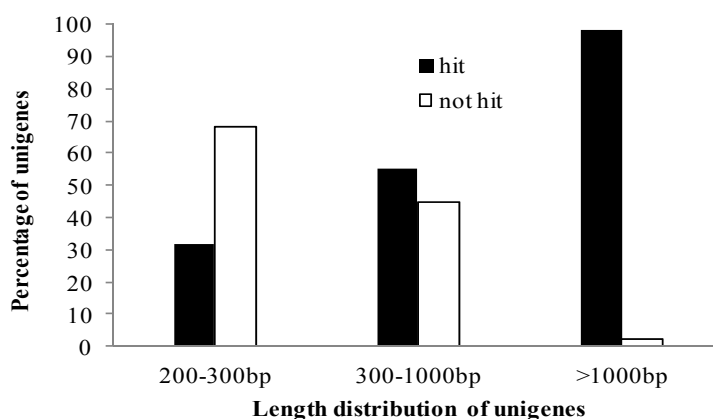

Figure 2. Comparison of unigene length between hit and not-hit unigenes. Longer unigenes were more likely to have BLAST matches in protein databases.
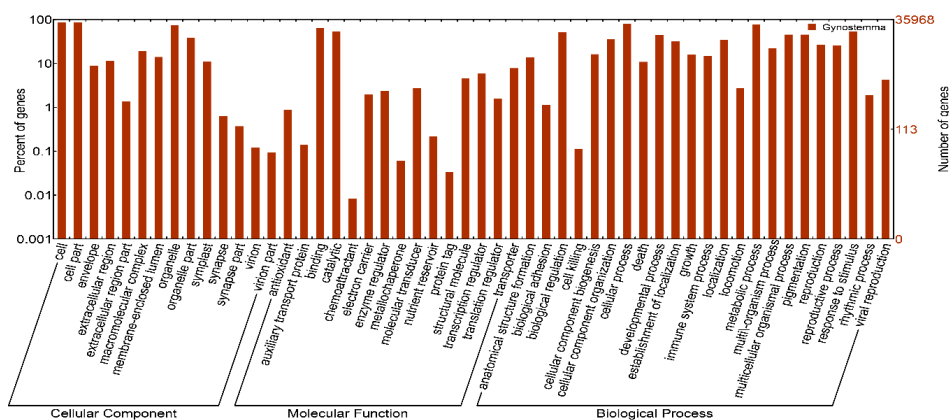

Figure 3. Gene Ontology classification of assembled unigenes. The results are summarized in three main categories: Biological Process, Cellular Component and Molecular Function. In total, 35,968 unigenes with BLAST matches to known proteins were assigned to gene ontology terms.
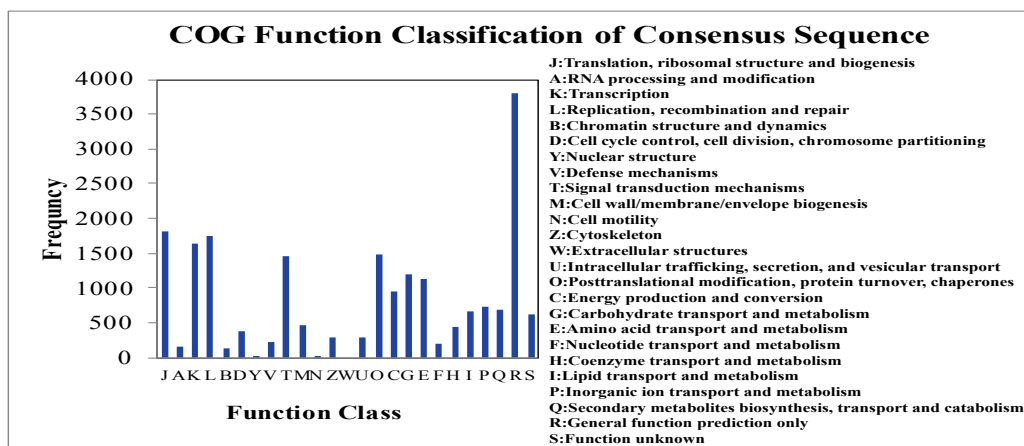
Figure 4. Histogram presentation of clusters of orthologous groups (COG). In total, 20,585 sequences were assigned to 25 COG classifications.
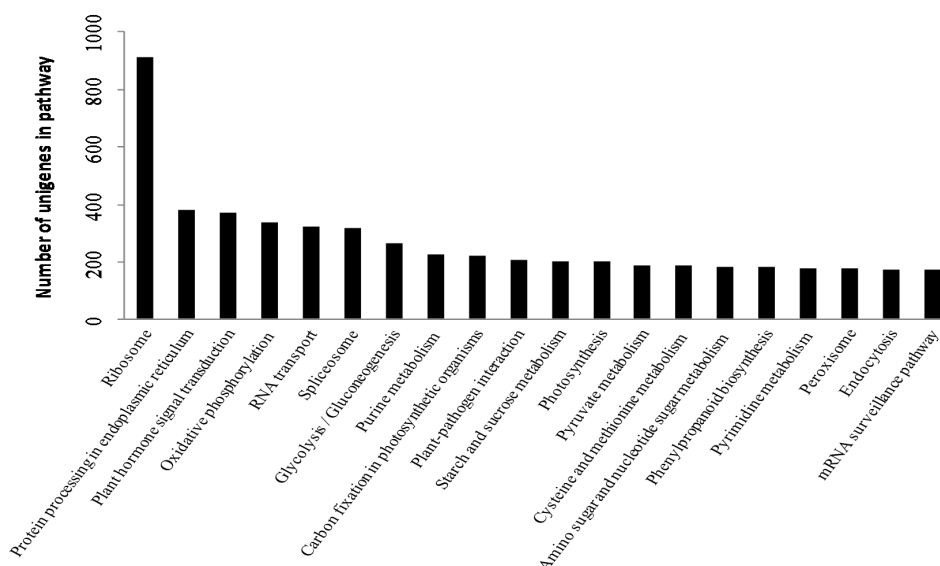

Figure 5. Kyoto Encyclopedia of Genes and Genomes (KEGG) classification of unigenes. 5,443 unigenes were assigned to the top 20 pathways in KEGG.

Table 3. List of triterpene saponin biosynthesis-related genes in the *Gynostemma* transcriptome.

Gene ID	Length	KO ID	Annotation
T3_Unigene_BMK.23540	2373	K01662	1-deoxyxylulose-5-phosphate synthase, (EC:2.2.1.7)
T3_Unigene_BMK.25990	2720	K01662	1-deoxyxylulose-5-phosphate synthase, (EC:2.2.1.7)
T3_Unigene_BMK.37882	301	K01662	1-deoxyxylulose-5-phosphate synthase, (EC:2.2.1.7)
T4_Unigene_BMK.23719	2489	K01662	1-deoxyxylulose-5-phosphate synthase, (EC:2.2.1.7)
T4_Unigene_BMK.30793	3057	K01662	1-deoxyxylulose-5-phosphate synthase, (EC:2.2.1.7)
T4_Unigene_BMK.33182	2483	K01662	1-deoxyxylulose-5-phosphate synthase, (EC:2.2.1.7)
CL10430Contig1	2754	K01662	1-deoxyxylulose-5-phosphate synthase, (EC:2.2.1.7)
T3_Unigene_BMK.15665	2671	K01662	1-deoxyxylulose-5-phosphate synthase, (EC:2.2.1.7)
CL9143Contig1	1586	K00919	4-diphosphocytidyl-2-*C*-methyl-D-erythritol kinase, (EC:2.7.1.148)

Table 3. *Cont.*

Gene ID	Length	KO ID	Annotation
T3_Unigene_BMK.22099	359	K03526	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase, (EC:1.17.7.1)
T3_Unigene_BMK.28355	906	K03527	4-hydroxy-3-methylbut-2-enyl diphosphate reductase, (EC:1.17.1.2)
CL12440Contig1	616	K03527	1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase, (EC:1.17.1.2)
T3_Unigene_BMK.9388	411	K03527	1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase, (EC:1.17.1.2)
CL12699Contig1	1973	K01823	isopentenyl-diphosphate delta-isomerase, (EC:5.3.3.2)
T4_Unigene_BMK.25472	575	K01823	isopentenyl-diphosphate delta-isomerase, (EC:5.3.3.2)
CL7433Contig1	750	K01823	isopentenyl-diphosphate delta-isomerase, (EC:5.3.3.2)
T3_Unigene_BMK.21000	1177	K13789	geranyl geranyl pyrophosphate synthase, (EC:2.5.1.29)
T4_Unigene_BMK.20525	1569	K14066	geranylgeranyl pyrophosphate synthase, (EC:2.5.1.30)
T3_Unigene_BMK.18996	1754	K14066	geranylgeranyl pyrophosphate synthase, (EC:2.5.1.30)
T3_Unigene_BMK.35842	305	K00626	acetyl-CoA acetyltransferase, mitochondrial, (EC:2.3.1.9)
T3_Unigene_BMK.35857	278	K00626	acetyl-CoA acetyltransferase, mitochondrial, (EC:2.3.1.9)
T4_Unigene_BMK.23349	1785	K00626	acetyl-CoA C-acetyltransferase, (EC:2.3.1.9)
T3_Unigene_BMK.18569	1657	K00626	acetyl-CoA C-acetyltransferase, (EC:2.3.1.9)
CL11048Contig1	2014	K01641	hydroxymethylglutaryl-CoA synthase, (EC:2.3.3.10)
T3_Unigene_BMK.2712	475	K00021	hmg-CoA reductase, (EC:1.1.1.34)
CL14352Contig1	2368	K00021	hmg-CoA reductase, (EC:1.1.1.34)
T4_Unigene_BMK.28606	1468	K00021	hmg-CoA reductase, (EC:1.1.1.34)
T4_Unigene_BMK.28692	322	K00021	hmg-CoA reductase, (EC:1.1.1.34)
T3_Unigene_BMK.22340	296	K00021	hmg-CoA reductase, (EC:1.1.1.34)
T3_Unigene_BMK.3271	467	K00869	mevalonate kinase, (EC:2.7.1.36)
CL13115Contig1	526	K00869	mevalonate kinase, (EC:2.7.1.36)
T4_Unigene_BMK.23148	1576	K00869	mevalonate kinase, (EC:2.7.1.36)
T4_Unigene_BMK.26974	522	K00869	mevalonate kinase, (EC:2.7.1.36)
T3_Unigene_BMK.14474	1691	K00869	mevalonate kinase, (EC:2.7.1.36)
T3_Unigene_BMK.22024	1903	K00938	Phosphomevalonate kinase, (EC:2.7.4.2)
T4_Unigene_BMK.29327	2282	K01597	diphosphomevalonate decarboxylase, (EC:4.1.1.33)
T3_Unigene_BMK.15864	1831	K01597	diphosphomevalonate decarboxylase, (EC:4.1.1.33)
T3_Unigene_BMK.5322	215	K11778	undecaprenyl pyrophosphate synthetase, (EC:2.5.1.31)
T4_Unigene_BMK.33183	1237	K11778	undecaprenyl pyrophosphate synthetase, (EC:2.5.1.31)
T4_Unigene_BMK.21428	1774	K00801	farnesyl-diphosphate farnesyltransferase, (EC:2.5.1.21)
T3_Unigene_BMK.3061	271	K00511	Squalene monooxygenase, (EC:1.14.99.7)
CL1336Contig1	2048	K00511	Squalene monooxygenase, (EC:1.14.99.7)

KO: KEGG Orthology.

2.3. Frequency and Distribution of Different Types of SSR Markers

A total of 3891 SSR loci were identified from 3526 nonredundant unigenes, representing 4.92% of the total 71,607 unigenes. The distribution density was one SSR locus per 15.78 kb. This study excluded mononucleotide repeats and complex SSRs. There were 329 unigenes with more than one SSR locus. The most frequent repeat unit in nonredundant unigenes were trinucleotides followed by dinucleotides (Figure 6A); and di- and tri-nucleotide repeats constituted 3778 (97.10%) of the identified SSR loci. The number of reiterations of a given repeat unit ranged from five to 12 and SSRs with five reiterations were the most abundant (Figure 6B). The majority of the SSR sequences were from 12 to 21 bp in length, accounting for 96.53% (3756) of the total identified SSR loci; SSR loci with a length of 15 bp were the most frequent. The longest SSR locus was 30 bp (Figure 6C). More details about different repeat motif of di- and trinucleotide repeats in EST-SSRs are listed in Table 4. The dominant di- and tri-nucleotide repeat motif in SSRs were AG/CT and AAG/CTT respectively. There was only one CG/CG repeat motif and very few ACT/AGT repeats in our results (Table 4).
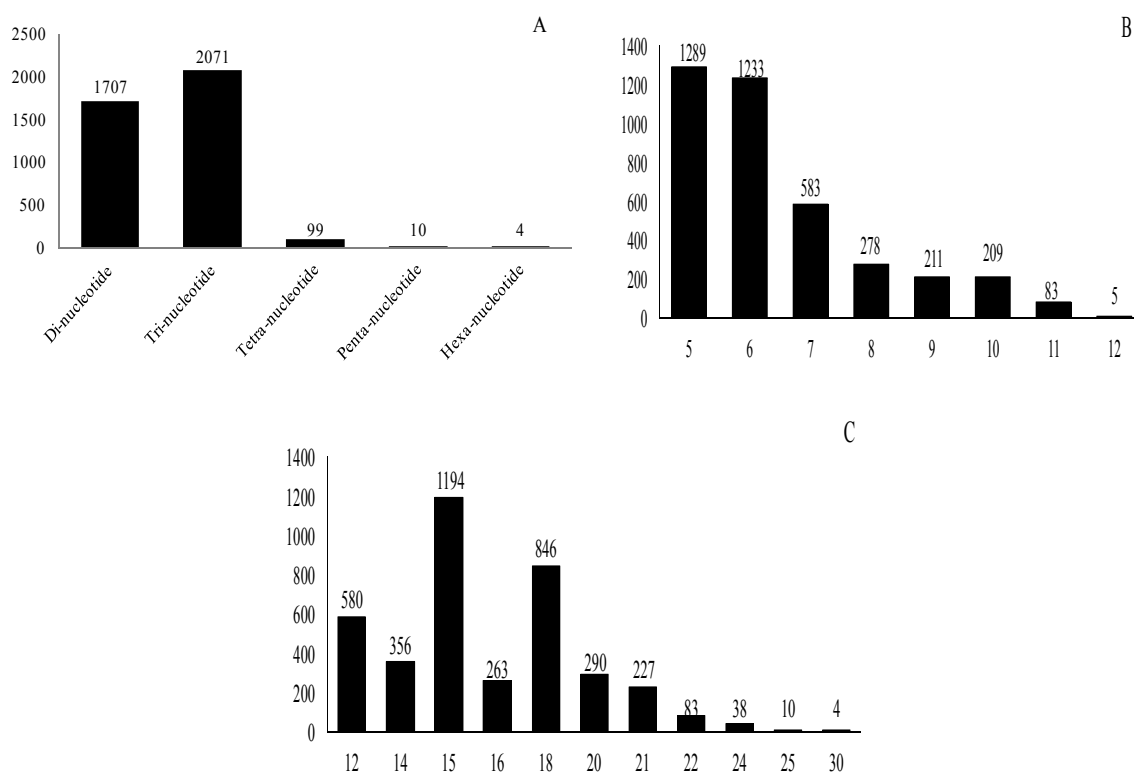
Figure 6. Frequency distribution of the *Gynostemma* expressed sequence tag (EST)-SSRs of different sizes. (**A**) Unit size; (**B**) Number of repeats; (**C**) SSR locus length.

Table 4. Frequency distribution of the di- and tri-nucleotide repeat motifs in the *Gynostemma*.

Serial No.	Repeat Motif	Number of Reiterations of the Motif								Total
		5	6	7	8	9	10	11	12	
1	AG/CT	#	456	285	217	189	192	81	3	1423
2	AAG/CTT	576	358	137	6	0	0	0	0	1077
3	ATC/ATG	159	74	25	0	0	0	0	0	258
4	AT/AT	#	95	54	36	18	14	2	1	220
5	AAT/ATT	97	50	19	2	0	0	0	0	168
6	AGG/CCT	90	41	12	1	0	0	0	0	144
7	AGC/CTG	83	24	9	1	0	0	0	0	117
8	ACC/GGT	52	32	6	1	0	0	0	0	91
9	AAC/GTT	49	24	14	2	0	0	0	0	89
13	CCG/CGG	49	17	3	0	0	0	0	0	69
10	AC/GT	#	28	17	10	4	3	0	1	63
11	ACG/CGT	24	10	0	0	0	0	0	0	34
12	ACT/AGT	15	5	2	2	0	0	0	0	24
14	CG/CG	0	1	0	0	0	0	0	0	1
15	Other motifs *	95	18	0	0	0	0	0	0	113
	Total	1289	1233	583	278	211	209	83	5	3891

Note: * indicated tetra-, penta-, and hexa-nucleotide motifs in our study; # means that this item was not considered when detecting EST-SSRs.

2.4. Development, Validation and Transferability of SSR Markers

To further evaluate the assembly quality, 2586 primer pairs (Table S2) were designed using Primer 3.0 based on 3891 SSR loci generated from MISA. Primer pairs for the remaining 1305 SSR loci could not be designed successfully because their flanking sequences were either too short or the nature of sequences did not satisfy the criteria of BatchPrimer3 v 1.0 software. All 2586 unigene

sequences were subjected to BLAST analysis to predict the likely function of these EST-SSRs. There were 2559 transcriptome sequences that showed homology to functional loci of other plants (Table S2). From the 2586 primer pairs, 360 (Table S3) were randomly selected for validation using DNA from the six samples of *G. pentaphyllum* of three different populations. Among the 360 primer pairs, 268 were successfully amplified via PCR. The remaining 92 primer pairs failed to generate PCR products, even when the annealing temperature was reduced by 8 °C. Of the 268 working primer pairs, 239 amplified products of the expected size including 23 monomorphic loci and 216 polymorphic loci among six genotypes of *G. pentaphyllum*. The other 29 generated larger products than the expected size, suggesting that there may exist introns in the amplifying regions. To test interspecies transferability across 12 related species (26 individuals) in the genus *Gynostemma*, 30 SSR pairs were selected from the 216 microsatellites that produce polymorphic size fragments. Of the 30 polymorphic SSRs, 15 primer pairs could amplify PCR products and show polymorphic fragments from all 12 *Gynostemma* species (Tables 5 and 6). Four pairs of primers (G-EST-SSR93, G-EST-SSR29, G-EST-SSR55, and G-EST-SSR31) failed to produce PCR fragments in *Gynostemma laxiflorum*; five pairs (G-EST-SSR85, G-EST-SSR42, G-EST-SSR54, G-EST-SSR59, and G-EST-SSR51) failed to produce PCR fragments in *Gynostemma caulopterum*; one pair of primers (G-EST-SSR14) failed to produce PCR fragments in *Gynostemma laxum* and *G. caulopterum*; one pair of primers (G-EST-SSR7) failed to produce PCR fragments in *Gynostemma pubescens*, *G. laxum*, *G. caulopterum*, and *G. laxiflorum*; one pair of primers (G-EST-SSR62) failed to produce PCR fragments in *Gynostemma microspermum*, *G. pubescens*, *G. laxum*, *G. laxiflorum*, and *G. caulopterum*; and three pairs of primers (G-EST-SSR158, G-EST-SSR92, and G-EST-SSR3) failed to produce PCR fragments in *G. laxum*. The cross-species amplification of these 30 EST-SSRs developed from *G. pentaphyllum* and *G. cardiospermum* in 10 additional *Gynostemma* species (*G. pentaphyllum* and *G. cardiospermum* were not considered when calculating the transferability of the EST-SSRs) was 92.33% in 300 combinations tested (30 SSRs × 10 species).

Table 5. Details of 15 genic-SSR loci showing polymorphism among 12 *Gynostemma* species.

Primer Name	Base on Sequence ID	SSRs	EPS (bp)	OPS (bp)	Alleles	PIC
G-EST-SSR19	CL10682Contig1	(AT)6	149	149–161	5	0.67
G-EST-SSR20	CL10765Contig1	(GAA)5	151	148–163	8	0.82
G-EST-SSR40	CL11560Contig1	(TCT)5	151	151–160	4	0.68
G-EST-SSR44	CL13255Contig1	(AG)6	164	162–196	11	0.87
G-EST-SSR47	CL13343Contig1	(TCT)6	148	151–172	8	0.79
G-EST-SSR57	CL14153Contig1	(TGA)5	149	152–170	6	0.74
G-EST-SSR75	CL400Contig1	(TCA)6	172	172–202	7	0.80
G-EST-SSR76	CL435Contig1	(TCT)5	131	122–137	6	0.76
G-EST-SSR89	CL1983Contig1	(GAAA)5	146	146–166	6	0.79
G-EST-SSR100	CL2782Contig1	(GAA)6	145	136–163	9	0.79
G-EST-SSR118	CL8580Contig1	(CGG)5	141	138–153	6	0.62
G-EST-SSR131	CL9957Contig1	(TCT)5	149	149–158	5	0.55
G-EST-SSR140	CL10320Contig1	(CTA)6	158	158–170	6	0.73
G-EST-SSR306	CL11525Contig1	(CT)7	169	169–189	6	0.73
G-EST-SSR316	CL12699Contig1	(GAA)5	135	135–159	8	0.66
Mean	-	-	-	-	6.73	0.73

EPS: expected product size; OPS: observed product size.

Table 6. The 26 individual plants (belonging to 12 species) used for validation and the genetic diversity study.

No.	Species	Locality, Province	Latitude (N), longitude (E)	Characteristics
1	*G. pentaphyllum*	Ankang, Shaanxi	32°25′N,109°04′E	wild
2	*G. pentaphyllum*	Ankang, Shaanxi	32°25′N,109°04′E	Cultivar
3	*G. pentaphyllum*	Kunming, Yunnan	25°14′N,102°49′E	Wild
4	*G. pentaphyllum*	Kunming, Yunnan	25°14′N,102°49′E	Cultivar
5	*G. pentaphyllum*	Panzhihua, Sichuan	26°36′N,101°43′E	Wild
6	*G. pentaphyllum*	Panzhihua, Sichuan	26°36′N,101°43′E	Cultivar
7	*G. burmanicum var. molle*	Dehong, Yunnan	24°48′N, 98°17′E	Wild
8	*G. burmanicum*	Menghai, Yunnan	22°02′N, 100°22′E	Wild
9	*G. burmanicum*	Dehong, Yunnan	24°36′N, 97°39′E	Wild
10	*G.pentaphyllum var. dasycarpum*	Mengla, Yunnan	22°14′N, 101°15′E	Wild
11	*G. pentaphyllum var. dasycarpum*	Jinghong, Yunnan,	22°01′N, 100°45′E	Wild
12	*G. longipes*	Xuancheng, Anhui	30°55′N,118°46′E	Wild
13	*G. longipes*	Ankang, Shaanxi	32°25′N,109°22′E	Wild
14	*G. longipes*	Enshi, Hubei	30°03′N, 109°49′E	Wild
15	*G. longipes*	Zhaotong, Yunnan	27°43′N,103°54′E	Wild
16	*G. pubescens*	Menghai, Yunnan	21°56′N, 100°36′E	Wild
17	*G. pubescens*	Menglun, Yunnan	21°56′N, 101°14′E	Wild
18	*G. pubescens*	Yingjiang, Yunnan	24°42′N, 97°55′E	Wild
19	*G. pubescens*	Enshi, Hubei	30°18′N, 109°31′E	Wild
20	*G. laxum*	Jiujiang, Jiangxi,	29°17′N, 115°07′E	Wild
21	*G. microspermum*	Pu'er, Yunnan	23°07′N, 100°22′E	Wild
22	*G. laxiflorum*	Xuancheng, Anhui	30°41′N,118°39′E	Wild
23	*G. yixingense*	Tongling, Anhui	30°57′N,117°48′E	Wild
24	*G. caulopterum*	Renhuai, Guizhou	27°48′N, 106°26′E	Wild
25	*G. cardiospermum*	Ankang, Shaanxi	32°13′N,109°01′E	Wild
26	*G. cardiospermum*	Ankang, Shaanxi	32°13′N,109°01′E	Wild

2.5. Genetic Diversity and Relatedness in the Genus Gynostemma

The 15 primer pairs that yielded clear, highly polymorphic bands from all *Gynostemma* species were used to assess the genetic diversity in a set of 26 individual plants representing 12 species of *Gynostemma* (Tables 5 and 6). A total of 101 alleles were identified, the number of alleles per locus ranged from four to 11 with an average of 6.73 alleles. Polymorphism information content (PIC) ranged from 0.55 to 0.87 with an average of 0.73, suggesting that the developed EST-SSRs were highly polymorphic. A phenogram based on Jaccard's similarity coefficients was constructed to resolve the relationship of 26 individuals from 12 species (Figure 7), which showed two distinct clusters at a cut-off similarity index of 0.71. Cluster I contained seven species, which corresponded to *subgen. Gynostemma*, and was divided into five sub-clusters: Ia, Ib, Ic, Id and Ie, at a cut-off similarity index of 0.78; Sub-cluster Ia comprised six *G. pentaphyllum* genotypes (three wild and three cultivar genotypes); Sub-cluster Ib comprised four *G. pubescens* from four populations and one *G. laxum*; Sub-cluster Ic comprised four *G. longipes* from four populations and one *G. burmanicum var. molle*; Sub-cluster Id comprised two *G. burmanicum* from two populations; and Sub-cluster Ie comprised two *G. pentaphyllum var. dasycarpum* from Jinghong and Mengla locations, respectively. Cluster II included five species that corresponded to *subgen. Triostellum*, and was divided into two sub-clusters, IIa and IIb, at a cut-off similarity index of 0.79. Sub-cluster IIa comprised one *G. microspermum*, one *G. laxiflorum* and one *G. caulopterum*, and Sub-cluster IIb comprised one *G. yixingense* and two *G. cardiospermum* individuals.
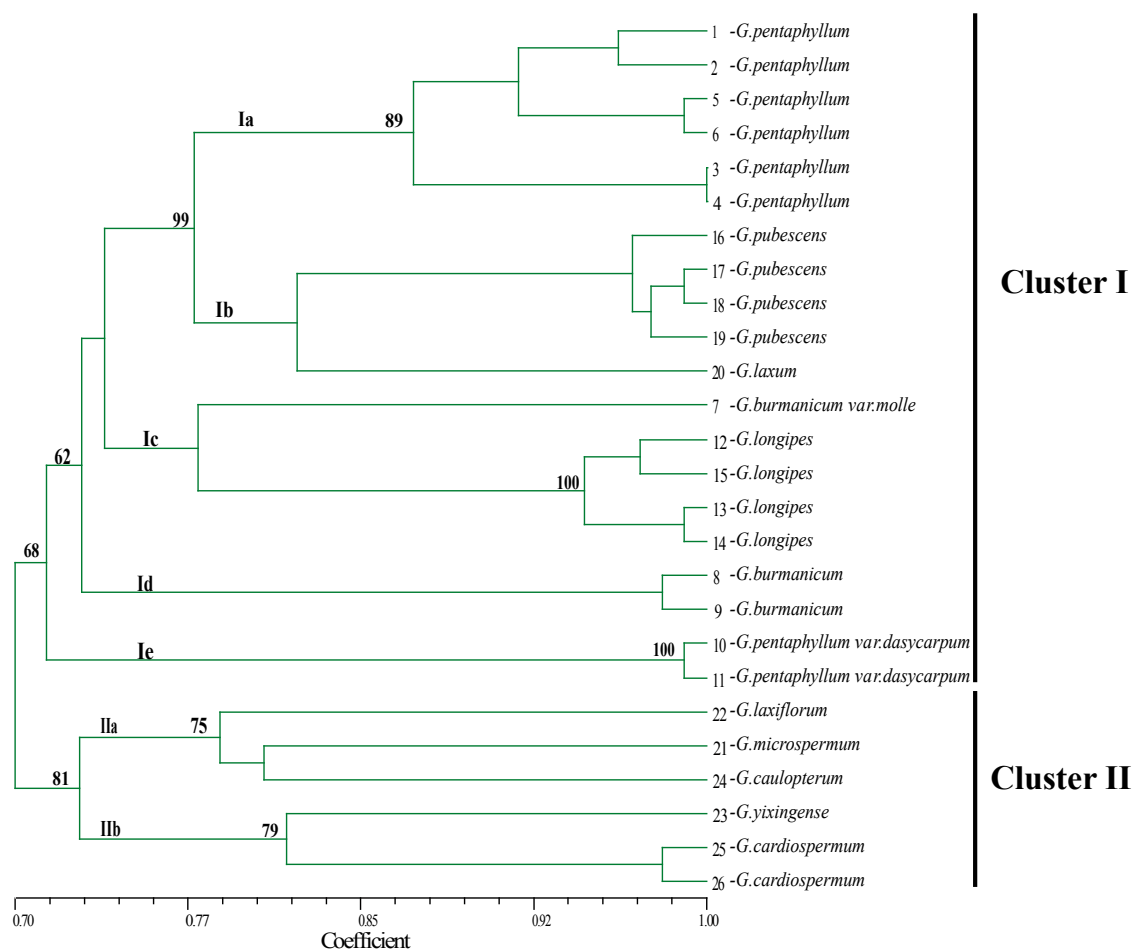
Figure 7. Genetic relationships among *Gynostemma* species based on EST-SSR markers. Genetic relationships among 26 individual plants. The scale at the bottom of the dendrogram indicates the level of similarity between the genotypes. Bootstrap values (>50) were labeled on the branches from 1000 re-samplings.

3. Discussion

3.1. Functional Annotation of Unigenes

Presently, most research concentrate on isolating bioactive components from *Gynostemma* spp., but the potential molecular mechanisms producing such compounds is still unclear. Transcriptome sequencing is an effective method for novel gene discovery and SSR marker development. In this study, 71,607 nonredundant unigenes were obtained after assembly. In total, 60.45% (43,288) of all unigenes had homologs in the NCBI nr or Swiss-Prot protein databases, which was lower than that reported by Subramaniyam *et al.* [14] with leaves and roots as sequencing materials in *Gynostemma pentaphyllum*. Compared with other plants used in Chinese medicine, this was higher than *Epimedium sagittatum* [35], but lower than *Panax quinquefolius* [36] and *Panax notoginseng* [37]. Among the 43,288 unigenes with BLAST matches in the NCBI nr or Swiss-Prot protein database, 97.73% were over 1000 bp, whereas unigenes shorter than 300 bp in length only accounted for 32.33% of the total. Therefore, we infer that the large proportion of BLAST matches in *Gynostemma* was probably due to the large number of long sequences in our unigene database, which was also validated in other plants [38–40]. Perhaps the lack of a characterized protein domain, a common feature of the shorter unigene sequences, was the cause of the small number of shorter sequences showing BLAST hits in the protein databases. Further research with GO analysis revealed that most genes are involved

in many biological processes in *Gynostemma*. Many genes were assigned to "metabolic process" and "catalytic activity" classes, which suggest a great deal of enzymes involved in primary and secondary metabolism. Among the KEGG pathways, the well-represented pathways discovered in our study were "Ribosome", "Protein processing in endoplasmic reticulum" and "Plant hormone signal transduction". Furthermore, some key genes involved in the biosynthesis of terpenoids were identified, several of which were found in other species [36,37]. Compared with the transcriptome of leaves and roots, genes related to biosynthesis of terpenoids showed little difference. For instance, 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (EC: 1.17.7.1), farnesyl-diphosphate farnesyltransferase (EC: 2.5.1.21) and Squalene monooxygenase (EC: 1.14.99.7), which were found in our study, were not presented in the results of Subramaniyam *et al.* [14]. Likewise, the genes present in results of Subramaniyam *et al.* [14], Hydroxymethylglutaryl-CoA reductase (EC: 1.1.1.88), Geranyl transtransferase (EC: 2.5.1.10), Dimethylallyltranstransferase (EC: 2.5.1.1), *etc.*, did not appear in our study either. These results reflect that there might exist different transcriptomic signatures in different tissues. Some unigenes without BLASTx hits may be potential *Gynostemma*-specific genes. Both of these classes can provide valuable information for the further study of *Gynostemma* spp., such as novel gene discovery and cloning, functional studies, and metabolic engineering of enzymes.

3.2. SSR Marker Frequency and Distribution in Gynostemma Transcriptome

Polymorphic SSRs play an important role in genetic diversity research, genetic mapping studies, comparative genomics, and marker-assisted selection breeding [41]. Transcriptomics provides a rich source for SSR discovery because it generates plenty of sequences. A total of 3891 SSRs greater than 12 bp in length were identified from 3526 nonredundant unigenes, 4.92% of the total 71,607 unigenes possessed SSRs. It is obvious that the SSR frequency detected in *Gynostemma* is in accordance with the range of frequencies (2.65%–16.82%) reported before for other dicotyledonous species [42]. Several factors affect the EST-SSR frequency. First, the criteria for calling microsatellites is the most important factor of EST-SSR frequency, e.g., the repeat length threshold and the number of repeat motifs. Most studies have excluded the mononucleotide repeat motifs because they may result from sequencing errors. Some studies take three-repeat units into account when calculating the number of dinucleotide repeat units [40], while others do not [43–45]. In addition, we identified SSRs primarily from unigenes over 1000 bp, which may reduce the frequency to a certain degree. Secondly, genome structure or composition could also influence SSR frequency [46]. For example, it is reported that the small genome size of rice was the cause of the high frequency of EST-SSR sequences [47]. Finally, the different software used to detect SSR loci can also affect the SSR frequency.

Theoretically, the frequency of di-, tri-, tetra-, penta-, and hexanucleotide repeats should be in turn decreased according to the relative probability of replication slippage events [48]. The most abundant type of repeat motif among the *Gynostemma* unigenes analyzed was trinucleotide (Figure 6A). This finding is consistent with the earlier results reported before [22,35,45,48–54], which showed the trinucleotide motif is the most frequent repeat type. Some studies point out the reason for the high frequency of tri-nucleotide SSRs is that the selection against frameshift mutations might limit the expansion of other SSR types [43,55–57]. Meanwhile, other studies show that the most abundant class of SSRs was dinucleotide [38,39,42,58,59]. There are also some plant species showing approximately equal proportions of dinucleotide and tri-nucleotide repeats in their transcriptome sequences, e.g., *Aspidistra saxicola* [44], sweet potato [39], and oak [60]. The most frequent repeats of di- and tri-nucleotide were AG/CT and AAG/CTT, respectively, which was in accordance with the reports in sesame [38], oil palm [61], sweet potato [39], *Primula* [62], and *Nothofagus nervosa* [63].

3.3. Transferability of SSR Markers and Genetic Relationships among Different Species of Gynostemma

In this study, 3891 SSR markers were developed and 360 primer pairs were randomly selected to evaluate the assembly quality of reads and validity of markers in *Gynostemma*. In total, 268 primer pairs (74.44%) yielded clear fragments among six *G. pentaphyllum* genotypes. This result

matches the 60%–90% success rate reported before. In total, 216 Polymorphic EST-SSR markers were obtained with a polymorphic proportion of 90.38%, which was similar to *Amorphophallus* [26], but was higher than other plants [20,21,52]. Our results suggest that the transcriptome assembly was reliable, and that the EST-SSR markers are usable across 12 species in the genus *Gynostemma*. The observed number of alleles ranged from four to 11 with an average of 6.73, indicating the potential application of these primer pairs. In the present study, EST-SSRs derived from *G. pentaphyllum* and *G. cardiospermum* had a higher transferability rate, which was also observed in other plant taxa [22,25,26,35,64]. It has been proposed that the high transferability rate of EST-SSRs might be due to several factors: (1) the EST-SSRs derived from transcriptome database are conservative when compared with genomic SSRs [65,66]; (2) the more consistent efficiency of amplification of EST-SSRs enhances cross-species transferability [67,68]; and (3) closely-related species benefit from a high SSR transferability rate [26]. However, at the same time, [49] explained that the limitation on the interspecific transfer of SSR markers is caused by homoplasy of band sizes and complex mutational events. The genetic relationship among 26 individuals representing 12 species of *Gynostemma* based on 15 polymorphic SSR loci was clearly shown in dendrogram graph. Two major groups representing *subgen. Gynostemma* and *subgen. Triostellum* respectively were identified at a cut-off similarity index of 0.71, the level of genetic similarity was 0.70–1.00, indicating relatively high resolution power and potential utility of polymorphic SSR markers in phylogenetics of *Gynostemma*. As expected, the six *G. pentaphyllum* individuals were classified into three groups, and wild individuals were clustered with cultivated individuals from the same population. The variation between populations was higher than the other *Gynostemma* species, implying that *G. pentaphyllum*, as a widespread species, has a high level of genetic diversity. These results are concordant with previous reports [69]. Therefore, the potential EST-SSRs identified in this study will be an effective tool for germplasm polymorphism assessment or quantitative trait loci mapping in *Gynostemma*.

4. Materials and Methods

4.1. Plant Materials

Young leaves, flowers and immature seeds from two species in the genus *Gynostemma* (*G. pentaphyllum* and *G. cardiospermum*) were used for RNA extraction and transcriptome sequencing. DNA from 26 individual plants collected from southeast China was used to validate SSR markers and diversity analysis. Detailed information for the plant materials is listed in Table 6.

4.2. RNA Extraction, Reverse Transcription and Sequencing

G. pentaphyllum and *G. cardiospermum* were collected from two locations of Ankang in Shaanxi province during July 2013 (*G. pentaphyllum*: 32°25′N, 109°04′E; *G. cardiospermum*: 32°13′N, 109°01′E), the multiple individual plants mixture including leaves, stems, flowers, shoot tips and developing seeds for each species were frozen immediately in liquid nitrogen, and stored at −70 °C. After mixing an approximately equal weight of mixture for each species, total RNA was extracted using the TRIzol reagent (Invitrogen, Carlsbad, CA, USA) according to manufacturer instructions, then poly-A mRNA was isolated from total RNA using poly-T oligo-attached magnetic beads (Illumina Inc., San Diego, CA, USA). The quantity and quality of RNA were assessed by gel electrophoresis and spectrophotometry. Purified RNA was used to construct a directional cDNA library using the cDNA Synthesis Kit (Illumina), and then the cDNA library was sequenced using a HiSeq 2000 (Illumina) to obtain short sequences.

4.3. Transcript Assembly and Analysis

All raw reads from the two *Gynostemma* species were prescreened to remove adapter sequences, reads with greater than 10% unknown bases, and reads with an average base quality less than 30. High-quality filtered transcriptome reads were assembled into contigs by *de novo* assembly using

Trinity tools [34]. A nonredundant set of unigene sequences was then created using paired-end reads by further alignments of the contigs from each species. To annotate them, all unigenes were searched against NCBI's nonredundant protein (nr) database and Swiss-Prot protein databases using BLASTx with an E-value <10^{-5}. The Blast2GO program [70] was used to get Gene Ontology (GO) terms to describe gene products according to three ontologies: molecular function, biological process and cellular component [71]. The unigene sequences were also aligned to the COG database to predict and classify functions. To further understand the biological functions and interactions of genes, pathway assignments were performed based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [72] using BLASTx with an E-value threshold of 10^{-5}.

4.4. EST-SSR Detection and Pprimer Design

Nonredundant unigene sequences longer than 1000 bp were used for mining SSR loci using the MISA tool [49], and primers were designed using BatchPrimer3 v1.0 software with default parameters [73]. Only cDNA-based SSR loci containing two to six nucleotide motifs were considered, the criteria for selection of SSRs were a minimum of six repeats for di-nucleotide motifs and five repeats for tri-, four repeats for tetra-, penta-, and hexa-nucleotide motifs. Mononucleotide repeats and complex SSR types were ignored. Frequency of SSR refers to the average number of kilobase pairs of cDNA sequence containing one SSR. The parameters for designing PCR primer pairs from sequences flanking SSRs were as follows: (1) primer length range from 18 to 25 bases (optimal 20 bases); (2) PCR product size range of 100 to 300 bp (optimal 200 bp); (3) annealing temperature of 50–60 °C (optimal 55 °C); and (4) a GC content of 40%–60% (optimal 50%). Other parameters were set at the default value of BatchPrimer3v1.0.

4.5. Plant DNA Extraction, PCR Conditions and Separation of SSR Markers

26 individuals, representing 12 *Gynostemma* species (Table 6), were selected for analysis of intraspecific genetic diversity, cross-species amplification with the EST-SSRs, and interspecific relationships. Plant DNA was extracted from leaf samples using the CTAB method [74], and DNA integrity was checked via electrophoresis on 1.5% agarose gel. PCR amplifications were carried out using a MyCycler™ Thermal Cycler (Bio-RAD, CA, USA) in a 10 μL final volume containing 1 × PCR buffer [10 mM Tris-HCl (pH 8.4), 1.5 mM MgCl$_2$], 0.2 mM dNTPs, 0.2 μM of each primer, 50 ng of genomic DNA, and 0.5 U *Taq* polymerase (Biostar, New Taipei, Taiwan). The PCR reaction program was: DNA denaturation at 95 °C for 5 min; followed by 32 cycles of 95 °C for 40 s, 50–60 °C (depending on optimized annealing temperature) for 30 s and 72 °C for 50 s. The final extension was performed at 72 °C for 10 min. PCR products were analyzed using 8% PAGE and silver stained [75] with a PBR322 DNA marker ladder (Tiangen Biotech, Beijing Co., Ltd., Beijing, China) for assessing the length of the DNA bands. A total of 360 genic SSR markers were selected randomly for genotyping six *G. pentaphyllum* samples from three populations, 30 highly polymorphic loci were selected for testing the transferability of EST-SSRs to the other ten species in the genus *Gynostemma.*

4.6. Genetic Analysis and Data Scoring

Of the 30 highly polymorphic loci, genic-SSR markers that amplified successfully in all 12 species were used to assess the genetic diversity in a set of 26 individual plants. Each allele was scored as present (1) or absent (0) for each of the SSR loci. The polymorphism information content (PIC) of each SSR primer was calculated to estimate the allelic variation of SSRs in the 26 individuals according to the formula: $PIC = 1 - \Sigma_{i=0}^{n} Pi^2$, where Pi is the frequency of the ith allele for a given SSR marker, and n is the total number of alleles detected for that SSR marker [76]. The genetic similarity between any two individuals was estimated based on Jaccard's similarity coefficient. All 26 individuals were clustered with the UPGMA algorithm and SAHN procedure of the NTSYS-PC v2.10t [77]. Bootstrapping analysis with 1000 replicates was carried out using the software FREETREE V.0.9.1.50 [78]. Bootstrap values over 50 were considered significant and provided on the dendrogram.

5. Conclusions

In this study, we used high-throughput sequencing to characterize the transcriptomes of two *Gynostemma* species. A large-scale EST dataset with 71,607 nonredundant unigenes from *G. pentaphyllum* and *G. cardiospermum* was established, which provided valuable sequences for the discovery of new genes and EST-SSR markers. These results support the view that NGS is a fast and cost-effective approach for gene discovery and molecular marker development in nonmodel species.

Author Contributions: G.-F.Z. and Y.-M.Z. conceived and designed the experiments. Y.-M.Z. performed the experiments and wrote the paper. T.Z. and Z.-H.L. analyzed the data. All authors read and approved the final manuscript.

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

References

1. Gao, X.F.; Chen, S.K.; Gu, Z.J.; Zhao, J.Z. A chromosomal study on the genus *Gynostemma* (*Cucurbitaceae*). *Acta Bot. Yunnanica* **1995**, *17*, 312–316.

2. Chen, S.K. A classificatory system and geographical distribution of the genus *Gynostemma* BL. (*Cucurbitaceae*). *Acta Phytotaxon. Sin.* **1995**, *33*, 403–410.

3. Zhou, Z.T.; Wang, Y.; Zhou, Y.M.; Zhang, S.L. Effect of *Gynostemma pentaphyllum* mak on carcinomatous conversions of golden hamster cheek pouches induced by dimethylbenzanthracene: A histological study. *Chin. Med. J.* **1998**, *111*, 847–850. [PubMed]

4. Lin, C.C.; Huang, P.C.; Lin, J.M. Antioxidant and hepatoprotective effects of *Anoectochilus formosanus* and *Gynostemma pentaphyllum*. *Am. J. Chin. Med.* **2000**, *28*, 87–96. [CrossRef] [PubMed]

5. Yin, F.; Hu, L.H.; Pan, R.X. Novel dammarane-type glycosides from *Gynostemma pentaphyllum*. *Chem. Pharm. Bull.* **2004**, *52*, 1440–1444. [CrossRef] [PubMed]

6. Liu, S.B.; Lin, R.; Hu, Z.H. Histochemical localization of Ginsenosides in *Gynostemma Pentaphyllum.* and the content changes of total gypenosides. *Acta Biol. Exp. Sin.* **2005**, *38*, 54–60.

7. Liu, S.B.; Lin, R.; Hu, Z.H. Comparison of stem and leaf structures and total gypenosides among 5 species of *Gynostemma*. *J. Fujian Agric. For. Univ.* **2006**, *35*, 495–499.

8. Yu, Y.F. A milestone of wild plant conservation in China. *Plants* **1999**, *5*, 3–11.

9. Jiang, L.Y.; Qian, Z.Q.; Guo, Z.G.; Wang, C.; Zhao, G.F. Polyploid origins in *Gynostemma pentaphyllum* (*Cucurbitaceae*) inferred from multiple gene sequences. *Mol. Phylogenet. Evol.* **2009**, *52*, 183–191. [CrossRef] [PubMed]

10. Xu, J.Q.; Shen, Q.; Li, J.; Hu, L.H. Dammaranes from *Gynostemma pentaphyllum* and synthesis of their derivatives as inhibitors of protein tyrosine phosphatase 1B. *Bioorg. Med. Chem.* **2010**, *18*, 3934–3939. [CrossRef] [PubMed]

11. Tsai, Y.C.; Lin, C.L.; Chen, B.H. Preparative chromatography of flavonoids and saponins in *Gynostemma pentaphyllum* and their antiproliferation effect on hepatoma cell. *Phytomedicine* **2010**, *18*, 2–10. [CrossRef] [PubMed]

12. Razmovski-Naumovski, V.; Huang, T.H.W.; Tran, V.H.; Li, G.Q.; Duke, C.C.; Roufogalis, B.D. Chemistry and pharmacology of *Gynostemma pentaphyllum*. *Phytochem. Rev.* **2005**, *4*, 197–219. [CrossRef]

13. Xie, Z.; Liu, W.; Huang, H.Q.; Slavin, M.; Zhao, Y.; Whent, M.; Blackford, J.; Lutterodt, H.; Zhou, H.P.; Chen, P.; *et al.* Chemical composition of five commercial *Gynostemma pentaphyllum* samples and their radical scavenging, antiproliferative, and anti-inflammatory properties. *J. Agric. Food Chem.* **2010**, *58*, 11243–11249. [CrossRef] [PubMed]

14. Subramaniyam, S.; Mathiyalagan, R.; In, J.G.; Lee, B.; Lee, S.; Y., D.C. Transcriptome profiling and *in silico* analysis of *Gynostemma pentaphyllum* using a next generation sequencer. *Plant. Cell. Rep.* **2011**, *30*, 2075–2083. [CrossRef] [PubMed]

15. Liao, H.; Zhao, Y.; Zhou, Y.; Wang, Y.G.; Wang, X.F.; Lu, F.; Song, Z.P. Microsatellite markers in the traditional Chinese medicinal herb *Gynostemma pentaphyllum* (*Cucurbitaceae*). *Am. J. Bot.* **2011**, *98*, e61–e63. [CrossRef] [PubMed]

16. Wang, C.; Zhang, H.; Qian, Z.Q.; Zhao, G.F. Genetic differentiation in endangered *Gynostemma pentaphyllum* (Thunb.) Makino based on ISSR polymorphism and its implications for conservation. *Biochem. Syst. Ecol.* **2008**, *36*, 699–705. [CrossRef]

17. Zane, L.; Bargelloni, L.; Patarnello, T. Strategies for microsatellite isolation: A review. *Mol. Ecol.* **2002**, *11*, 1–16. [CrossRef] [PubMed]

18. Squirrell, J.; Hollingsworth, P.M.; Woodhead, M.; Russell, J.; Lowe, A.J.; Gibby, M.; Powell, W. How much effort is required to isolate nuclear microsatellites from plants? *Mol. Ecol.* **2003**, *12*, 1339–1348. [CrossRef] [PubMed]

19. Bozhko, M.; Riegel, R.; Schubert, R.; Müller-Starck, G. A cyclophilin gene marker confirming geographical differentiation of Norway spruce populations and indicating viability response on excess soil-born salinity. *Mol. Ecol.* **2003**, *12*, 3147–3155. [CrossRef] [PubMed]

20. Varshney, R.K.; Graner, A.; Sorrells, M.E. Genic microsatellite markers in plants: Features and applications. *Trends Biotechnol.* **2005**, *23*, 48–55. [CrossRef] [PubMed]

21. Peakall, R.; Gilmore, S.; Keys, W.; Morgante, M.; Rafalski, A. Cross-species amplification of soybean (*Glycine max*) simple sequence repeats (SSRs) within the genus and other legume genera: Implications for the transferability of SSRs in plants. *Mol. Biol. Evol.* **1998**, *15*, 1275–1287. [CrossRef] [PubMed]

22. Eujayl, I.; Sledge, M.K.; Wang, L.; May, G.D.; Chekhovskiy, K.; Zwonitzer, J.C.; Mian, M.A. Medicago truncatula EST-SSRs reveal cross-species genetic markers for *Medicago* spp. *Theor. Appl. Genet.* **2004**, *108*, 414–422. [CrossRef] [PubMed]

23. Zhang, L.Y.; Bernard, M.; Leroy, P.; Feuillet, C.; Sourdille, P. High transferability of bread wheat EST-derived SSRs to other cereals. *Theor. Appl. Genet.* **2005**, *111*, 677–687. [CrossRef] [PubMed]

24. Poncet, V.; Rondeau, M.; Tranchant, C.; Cayrel, A.; Hamon, S.; de Kochko, A.; Hamon, P. SSR mining in coffee tree EST databases: Potential use of EST-SSRs as markers for the *Coffea* genus. *Mol. Genet. Genom.* **2006**, *276*, 436–449. [CrossRef] [PubMed]

25. Luro, F.L.; Costantino, G.; Terol, J.; Argout, X.; Allario, T.; Wincker, P.; Talon, M.; Ollitrault, P.; Morillon, R. Transferability of the EST-SSRs developed on *Nules clementine* (*Citrus clementina* Hort ex Tan) to other *Citrus* species and their effectiveness for genetic mapping. *BMC Genom.* **2008**, *9*, 287. [CrossRef] [PubMed]

26. Zheng, X.F.; Pan, C.; Diao, Y.; You, Y.N.; Yang, C.Z.; Hu, Z.L. Development of microsatellite markers by transcriptome sequencing in two species of *Amorphophallus* (*Araceae*). *BMC Genom.* **2013**, *14*, 490. [CrossRef] [PubMed]

27. Zalapa, J.E.; Cuevas, H.; Zhu, H.; Steffan, S.; Senalik, D.; Zeldin, E.; McCown, B.; Harbut, R.; Simon, P. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *Am. J. Bot.* **2012**, *99*, 193–208. [CrossRef] [PubMed]

28. Bouck, A.; Vision, T. The molecular ecologist's guide to expressed sequence tags. *Mol. Ecol.* **2007**, *16*, 907–924. [CrossRef] [PubMed]

29. Emrich, S.J.; Barbazuk, W.B.; Li, L.; Schnable, P.S. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* **2007**, *17*, 69–73. [CrossRef] [PubMed]

30. Vera, J.C.; Wheat, C.W.; Fescemyer, H.W.; Frilander, M.J.; Crawford, D.L.; Hanski, I.; Marden, J.H. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol.* **2008**, *17*, 1636–1647. [CrossRef] [PubMed]

31. Barbazuk, W.B.; Emrich, S.J.; Chen, H.D.; Li, L.; Schnable, P.S. SNP discovery via 454 transcriptome sequencing. *Plant. J.* **2007**, *51*, 910–918. [CrossRef] [PubMed]

32. Novaes, E.; Drost, D.R.; Farmerie, W.G.; Pappas, G.J., Jr.; Grattapaglia, D.; Sederoff, R.R.; Kirst, M. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genom.* **2008**, *9*, 312. [CrossRef] [PubMed]

33. Namroud, M.C.; Beaulieu, J.; Juge, N.; Laroche, J.; Bousquet, J. Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Mol. Ecol.* **2008**, *17*, 3599–3613. [CrossRef] [PubMed]

34. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef] [PubMed]

35. Zeng, S.H.; Xiao, G.; Guo, J.; Fei, Z.J.; Xu, Y.Q.; Roe, B.A.; Wang, Y. Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. *BMC Genom.* **2010**, *11*, 94. [CrossRef] [PubMed]

36. Sun, C.; Li, Y.; Wu, Q.; Luo, H.M.; Sun, Y.Z.; Song, J.Y.; Lui, E.M.; Chen, S.L. *De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. *BMC Genom.* **2010**, *11*, 262. [CrossRef] [PubMed]

37. Luo, H.M.; Sun, C.; Sun, Y.Z.; Wu, Q.; Li, Y.; Song, J.Y.; Niu, Y.Y.; Cheng, X.L.; Xu, H.X.; Li, C.Y.; *et al*. Analysis of the transcriptome of Panax notoginseng root uncovers putative triterpene saponin-biosynthetic genes and genetic markers. *BMC Genom.* **2011**, *12*, S5. [CrossRef] [PubMed]

38. Wei, W.L.; Qi, X.Q.; Wang, L.H.; Zhang, Y.X.; Hua, W.; Li, D.H.; Lv, H.X.; Zhang, X.R. Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC Genom.* **2011**, *12*, 451. [CrossRef] [PubMed]

39. Wang, Z.Y.; Fang, B.P.; Chen, J.Y.; Zhang, X.J.; Luo, Z.X.; Huang, L.F.; Chen, X.L.; Li, Y.J. *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (*Ipomoea batatas*). *BMC Genom.* **2010**, *11*, 726. [CrossRef] [PubMed]

40. Li, D.J.; Deng, Z.; Qin, B.; Liu, X.H.; Men, Z.H. *De novo* assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.). *BMC Genom.* **2012**, *13*, 192. [CrossRef] [PubMed]

41. Luikart, G.; England, P.R.; Tallmon, D.; Jordan, S.; Taberlet, P. The power and promise of population genomics: From genotyping to genome typing. *Nat. Rev. Genet.* **2003**, *4*, 981–994. [CrossRef] [PubMed]

42. Kumpatla, S.P.; Mukhopadhyay, S. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome* **2005**, *48*, 985–998. [CrossRef] [PubMed]

43. Jhanwar, S.; Priya, P.; Garg, R.; Parida, S.K.; Tyagi, A.K.; Jain, M. Transcriptome sequencing of wild chickpea as a rich resource for marker development. *Plant. Biotechnol. J.* **2012**, *10*, 690–702. [CrossRef] [PubMed]

44. Huang, D.N.; Zhang, Y.Q.; Jin, M.D.; Li, H.K.; Song, Z.P.; Wang, Y.G.; Chen, J.K. Characterization and high cross-species transferability of microsatellite markers from the floral transcriptome of *Aspidistra saxicola* (*Asparagaceae*). *Mol. Ecol. Resour.* **2014**, *14*, 569–577. [CrossRef] [PubMed]

45. Cloutier, S.; Niu, Z.X.; Datla, R.; Duguid, S. Development and analysis of EST-SSRs for flax (*Linum usitatissimum* L.). *Theor. Appl. Genet.* **2009**, *119*, 53–63. [CrossRef] [PubMed]

46. Tóth, G.; Gáspári, Z.; Jurka, J. Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res.* **2000**, *10*, 967–981. [CrossRef] [PubMed]

47. Varshney, R.K.; Thiel, T.; Stein, N.; Langridge, P.; Graner, A. *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell. Mol. Biol. Lett.* **2002**, *7*, 537–546. [PubMed]

48. Kaur, S.; Pembleton, L.W.; Cogan, N.O.; Savin, K.W.; Leonforte, T.; Paull, J.; Materne, M.; Forster, J.W. Transcriptome sequencing of field pea and faba bean for discovery and validation of SSR genetic markers. *BMC Genom.* **2012**, *13*, 104. [CrossRef] [PubMed]

49. Thiel, T.; Michalek, W.; Varshney, R.K.; Graner, A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* **2003**, *106*, 411–422. [PubMed]

50. Han, Z.G.; Guo, W.Z.; Song, X.L.; Zhang, T.Z. Genetic mapping of EST-derived microsatellites from the diploid *Gossypium arboreum* in allotetraploid cotton. *Mol. Genet. Genom.* **2004**, *272*, 308–327. [CrossRef] [PubMed]

51. Hisano, H.; Sato, S.; Isobe, S.; Sasamoto, S.; Wada, T.; Matsuno, A.; Fujishiro, T.; Yamada, M.; Nakayama, S.; Nakamura, Y.; Watanabe, S.; Harada, K.; Tabata, S. Characterization of the soybean genome using EST-derived microsatellite markers. *DNA Res.* **2007**, *14*, 271–281. [CrossRef] [PubMed]

52. Kaur, S.; Cogan, N.O.; Pembleton, L.W.; Shinozuka, M.; Savin, K.W.; Materne, M.; Forster, J.W. Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. *BMC Genom.* **2011**, *12*, 265. [CrossRef] [PubMed]

53. La Rota, M.; Kantety, R.V.; Yu, J.K.; Sorrells, M.E. Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genom.* **2005**, *6*, 23. [CrossRef] [PubMed]

54. Chen, C.X.; Zhou, P.; Choi, Y.A.; Huang, S.; Gmitter, F.G., Jr. Mining and characterizing microsatellites from citrus ESTs. *Theor. Appl. Genet.* **2006**, *112*, 1248–1257. [CrossRef] [PubMed]

55. Metzgar, D.; Bytof, J.; Wills, C. Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.* **2000**, *10*, 72–80. [PubMed]

56. Parida, S.K.; Kumar, K.A. R.; Dalal, V.; Singh, N.K.; Mohapatra, T. Unigene derived microsatellite markers for the cereal genomes. *Theor. Appl. Genet.* **2006**, *112*, 808–817. [CrossRef] [PubMed]

57. Parida, S.K.; Dalal, V.; Singh, A.K.; Singh, N.K.; Mohapatra, T. Genic non-coding microsatellites in the rice genome: Characterization, marker design and use in assessing genetic and evolutionary relationships among domesticated groups. *BMC Genom.* **2009**, *10*, 140. [CrossRef] [PubMed]

58. Dutta, S.; Kumawat, G.; Singh, B.P.; Gupta, D.K.; Singh, S.; Dogra, V.; Gaikwad, K.; Sharma, T.R.; Raje, R.S.; Bandhopadhya, T.K.; *et al*. Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan* (L.) Millspaugh]. *BMC Plant. Biol.* **2011**, *11*, 17. [CrossRef] [PubMed]

59. Rungis, D.; Berube, Y.; Zhang, J.; Ralph, S.; Ritland, C.E.; Ellis, B.E.; Douglas, C.; Bohlmann, J.; Ritland, K. Robust simple sequence repeat markers for spruce (*Picea* spp.) from expressed sequence tags. *Theor. Appl. Genet.* **2004**, *109*, 1283–1294. [CrossRef] [PubMed]

60. Ueno, S.; Le Provost, G.; Leger, V.; Klopp, C.; Noirot, C.; Frigerio, J.M.; Salin, F.; Salse, J.; Abrouk, M.; Murat, F.; *et al*. Bioinformatic analysis of ESTs collected by Sanger and pyrosequencing methods for a keystone forest tree species: Oak. *BMC Genom.* **2010**, *11*, 650. [CrossRef] [PubMed]

61. Singh, R.; Zaki, N.; Ting, N.-C.; Rosli, R.; Tan, S.-G.; Low, E.-T.; Ithnin, M.; Cheah, S.-C. Exploiting an oil palm EST database for the development of gene-derived SSR markers and their exploitation for assessment of genetic diversity. *Biologia* **2008**, *63*, 227–235. [CrossRef]

62. Zhang, L.; Yan, H.F.; Wu, W.; Yu, H.; Ge, X.J. Comparative transcriptome analysis and marker development of two closely related Primrose species (*Primula poissonii* and *Primula wilsonii*). *BMC Genom.* **2013**, *14*, 329. [CrossRef] [PubMed]

63. Torales, S.L.; Rivarola, M.; Pomponio, M.F.; Fernandez, P.; Acuna, C.V.; Marchelli, P.; Gonzalez, S.; Azpilicueta, M.M.; Hopp, H.E.; Gallo, L.A.; *et al*. Transcriptome survey of Patagonian southern beech *Nothofagus nervosa* (= *N. Alpina*.): Assembly, annotation and molecular marker discovery. *BMC Genom.* **2012**, *13*, 291. [CrossRef] [PubMed]

64. Saha, M.C.; Mian, M.R.; Eujayl, I.; Zwonitzer, J.C.; Wang, L.; May, G.D. Tall fescue EST-SSR markers with transferability across several grass species. *Theor. Appl. Genet.* **2004**, *109*, 783–791. [CrossRef] [PubMed]

65. Cho, Y.G.; Ishii, T.; Temnykh, S.; Chen, X.; Lipovich, L.; McCOUCH, S.R.; Park, W.D.; Ayres, N.; Cartinhour, S. Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* **2000**, *100*, 713–722. [CrossRef]

66. Eujayl, I.; Sorrells, M.; Baum, M.; Wolters, P.; Powell, W. Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. *Theor. Appl. Genet.* **2002**, *104*, 399–407. [CrossRef] [PubMed]

67. Ellis, J.R.; Burke, J.M. EST-SSRs as a resource for population genetic analyses. *Heredity* **2007**, *99*, 125–132. [CrossRef] [PubMed]

68. Barbara, T.; Palma-Silva, C.; Paggi, G.M.; Bered, F.; Fay, M.F.; Lexer, C. Cross-species transfer of nuclear microsatellite markers: Potential and limitations. *Mol. Ecol.* **2007**, *16*, 3759–3767. [CrossRef] [PubMed]

69. Zhang, X.; Zheng, Q.J.; Li, Z.H.; Zhao, G.F. Genetic diversity and population structure of *Gynostemma pentaphyllun*. *Chin. Tradit. Herb. Drugs* **2015**, *46*, 1958–1965.

70. Conesa, A.; Gotz, S.; Garcia-Gomez, J.M.; Terol, J.; Talon, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676. [CrossRef] [PubMed]

71. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; *et al*. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef] [PubMed]

72. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef] [PubMed]

73. You, F.M.; Huo, N.; Gu, Y.Q.; Luo, M.C.; Ma, Y.; Hane, D.; Lazo, G.R.; Dvorak, J.; Anderson, O.D. BatchPrimer3: A high throughput web application for PCR and sequencing primer design. *BMC Bioinform.* **2008**, *9*, 253. [CrossRef] [PubMed]

74. Porebski, S.; Bailey, L.G.; Baum, B.R. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Rep.* **1997**, *15*, 8–15. [CrossRef]

75. Creste, S.; Neto, A.T.; Figueira, A. Detection of single sequence repeat polymorphisms in denaturing polyacrylamide sequencing gels by silver staining. *Plant Mol. Biol. Rep.* **2001**, *19*, 299–306. [CrossRef]

76. Botstein, D.; White, R.L.; Skolnick, M.; Davis, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **1980**, *32*, 314. [PubMed]

77. Rolf, J. *Numerical Taxonomy and Multivariate Analysis System, Version 2.11*; T Exeter Software: Setauket, NY, USA, 2000.

78. Pavlicek, A.; Hrda, S.; Flegr, J. Free-Tree-freeware program for construction of phylogenetic trees on the basis of distance data and bootstrap/jackknife analysis of the tree robustness. Application in the RAPD analysis of genus *Frenkelia*. *Folia Biol.* **1999**, 97–99.

Sample Availability: All samples are available from the authors.