De Novo Assembly of Transcriptome Sequencing and EST-SSR development in Aesculus chinensis Bunge

Yuling Li
Shandong Academy of Forestry
Xin Mao
Shandong Agricultural University
Tianxu Sun
Shandong Institute of Territorial and Spatial Planning
Cuilan Liu
Shandong Academy of Forestry
Yuanshuai Zhang
DaQingShan forest farm of Feixian
Yuanfu Dong
DaQingShan forest farm of Feixian
Xiuhong Mao (xiuhongmao@shandong.cn)
Shandong Agricultural University

Research Article

Keywords: Aesculus chinensis Bunge, transcriptome sequencing, Gene annotation, EST-SSR markers

Posted Date: June 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1717796/v1

Abstract

Aesculus chinensis Bunge (*A. chinensis* Bunge), a deciduous tree in the genus *Aesculus* of the family Hippocastanaceae. and has ornamental, economic and medicinal values. Lack of comprehensive genetic information hinders the development and further utilization of this species and EST-SSR markers of *A. chinensis* Bunge have not been previously reported. For this reason we obtained *A. chinensis* Bunge transcriptomes using Illumina HiSeq TM2000 and de novo assembly, developed the EST-SSR markers. In this study, 74,805,734 clean reads were generated using Illumina HiSeq TM2000 and obtained 33,365 unigenes with a mean length of 1348bp after assembly. A total of 25450 (76.27%) unigenes showed similarities in at least one database including Nr (71.39%), Swiss Prot (57.03%), GO (53.31%), KO (27.13%) and KOG (19.27%). Based on the transcriptome data collected for *A. chinensis* Bunge, 16,103 SSR loci were detected with the distribution density of one SSR per 2.79 kb. The most dinucleotide and trinucleotide repeats were AG/CT and AAG/CTT. Then 58 primer pairs were randomly selected for polymorphism validation and 31 primer pairs were polymorphic among eight *Aesculus* species. A total of 167 alleles were identified, varying from 2 to 9 with a mean of 5.387 alleles per locus.

Key Message

We performed de novo transcriptome sequencing in *A. chinensis* Bunge via RNA-seq and developed 31 polymorphic EST-SSR markers with moderate cross-transferability in genus *Aesculus*. The valuable dataset will be useful in molecular marker-assisted selection breeding of *Aesculus.*

Introduction

A. chinensis Bunge, also known as Sala tree, belongs to the genus *Aesculus* of the family Hippocastanaceae. As deciduous trees, *A. chinensis* Bunge has tall trunk, dark brown or grayish brown bark, oblong-lanceolate or oblong-oblanceolate, leaflets range from five to seven. inflorescence rachis with puberulent, florets with 5–10 flowers and subglobose seeds. *A. chinensis* Bunge grows wild only in the Qin Ling Mountains while is cultivated in southern Shaanxi, southern Hebei and northern Henan of China (Chinese Flora.1982). It was one of the four famous street trees around the world, which has many functions, such as ornamental trees, medicinal plants, food and so on. Seeds of genus *Aesculus* contain aescin, which has a wide range of pharmacological activities, including anti-inflammation, inhibition of gastric acid secretion, anti-edema, scavenging of reactive oxygen species, anti-liver injury and so on (Liu and Zhou 2010; Kim et al.2017; Chen et al.2019). Studies about the molecular biology and the preservation of genetic resources in *A. chinensis* Bunge are very limited. Only the chloroplast genome of *A. chinensis* has been recently sequenced (Zhang et al.2019; Liu et al.2020).

Transcriptome Sequencing based on the Illumina HiSeq TM2000 has been widely used for transcriptome profiling, gene expression analysis, and the detection of functional genes due to its high accuracy, high speed and low cost (Severin et al,2010). Transcriptome sequencing can obtain the law of growth and metabolism of species under the condition of lack of gene resources, and reveal the internal relationship between their biological characteristics and genes. It can also analyze and explore the gene structure and function in a deeper level, which is a cost-effective means of gene sequence research (Simon et al,2009; Reddy et al,2015). As de novo transcriptome assembly functions independently from existing genomic sequences, it can be particularly useful for the analysis of non-model species containing large nuclear genomes (Wang et al,2009). To date, transcriptome sequencing technology has been widely used in woody plants such as Liriodendron chinense (Yang et al.2014), Hazelnut (Cheng et al.2015), Tea (Guo et al.2018), Blueberry (Qi et al.2019), Date palm (Naganeeswaran et al.2020), ect.

For non-model plants without reference genomes, transcriptome sequencing is an effective means to discover new genes and collect ESTs, which will also facilitate the development of molecular markers (Wang et al.2009; Grabherr et al.2011). Molecular markers development via transcriptome sequencing have the advantages of cheap, simple and rapid, high polymorphism and transferable among plant species, which can greatly improve working efficiency. One of the most common molecular marker is simple sequence repeat (SSR), which is a subclass of tandem repeats consisting of 1–6 nucleotides, found in the genomes of all prokaryotes and eukaryotes (Taheri et al,2014). SSRs are generally classified into EST-SSR and genomic-SSR. EST-SSR markers are present in more conserved regions in comparison to markers that are generated from genomic sequences, and therefore they show more transferability between species. EST-SSR markers have been used in genetic diversity assessment, evolution, linkage map, variety identification and marker-assisted selection breeding because of their high polymorphism, repeatability, sensitivity and easy detection (Liu et al.2013, Zhou et al.2016, Taheri et al.2018; He et al,2020).

The breeding programs in *A. chinensis* Bunge have been hindered due to the limited availability of genomic resources. In the present study, transcriptome sequencing and assemble were carried out on the leaves and inflorescences of *A. chinensis* Bunge and a set of EST-SSR makers were developed from the transcriptome data. The purpose of the study is to provide a theoretical basis for screening genes, developing SSR molecular markers and further functional genome research.

Materials And Methods

Materials and RNA extraction

The materials used in this experiment were the leaves and inflorescences of three *A. chinensis* Bunge plants, which were collected on the campus of Shandong University (117.060105E, 36.673287N) in April 10 of 2019. After sampling, the samples were quickly placed in liquid nitrogen and then transferred to -80 ℃ refrigerator. RNA extraction and purification follow the instructions of the “Plant RNA Kit” (Omega Bio-tek Int, China). The 1% agarose gels were used to monitor RNA degradation and contamination. NanoPhotometer® spectrophotometer (IMPLEN, CA, USA) was used to check RNA purity. The RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA) was used to assess RNA integrity.

Library preparation for Transcriptome sequencing

The amount of 1.5 μg RNA samples were used for the preparations of RNA samples. According to the manufacturer’s recommendations, Sequencing libraries were generated using NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB, USA). The mRNA + poly (A) tail was enriched by Oligo (dT) magnetic beads and fragmentation buffer was added to cut the mRNA into short fragments. Using fragment mRNA as template and random oligonucleotides as primers, the first cDNA strand was synthesized by M-MuLV reverse transcriptase system, and second-strand cDNA was synthesized by buffer, dNTPs, RNaseH and DNA polymerase I. On the basis of amplified products, cDNAs were purified (AMPure XP system) and library quality was evaluated on the Agilent Bioanalyzer 2100system.

Clustering and sequencing

The clustering of the index-coded samples were performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v3-cBot-HS (Illumia) according to the manufacturer’s instructions. After cluster generation, the library preparations were sequenced on an Illumina Hiseq TM2000 platform (Reuter et al. 2015) and paired-end reads were generated.

Data analysis

Quality control: Raw data of fast q format were firstly processed through in-house perl scripts. In this step, clean data were obtained by removing reads containing adapter, reads containing ploy-N and low qualities reads from raw data. At the same time, Q20, Q30, GC-content and sequence duplication level of the clean data were calculated (Grabherr et al. 2011). All the downstream analyses were based on clean data with high quality.

Transcriptome assembly: The left files (read1 files) from all libraries/samples were pooled into one big left.fq file, and right files (read2 files) into one big right.fq file. Transcriptome assembly was accomplished based on the left.fq and right.fq using Trinity v2.4.0 (https://github.com/trini tyrnaseq/trinityrnaseq/issues/270) with min_kmer_cov set to 2 by default and all other parameters set default.

Gene functional annotation: Gene functions were annotated to the following databases: Nr (NCBI non-redundant protein sequences; e-value = 1e-5) (https://www.ncbi.nlm.nih.gov/protein/); Pfam (Protein family; e-value = 0.01) (http://pfam.xfam.org); Nt (NCBI non-redundant nucleotide sequences; e-value = 1e-5) (https://www.ncbi.nlm.nih.gov/nucleotide/); KOG (Clusters of Orthologous Groups of proteins; e-value = 1e-5) (http://www.ncbi.nlm.nih.gov/COG); Swiss-Prot (e-value = 1e-5) (http://www.expas y.ch/sprot/); KO (KEGG Ortholog database; e-value = 1e-10) (http://www.genom e.jp/kegg); GO (Gene Ontology; e-value = 1e-6) (http://www.geneontology.org/).

CDS prediction

Unigenes were compared according to the priority order of Nr and Swiss-Prot protein databases. Secondly, unigenes failing to hit any in the above databases were then subjected to the software ESTS can (version 3.0.3) (https://sourceforge.net/projects/estscan/) for CDS prediction.

SSR detection and Primer Design

SSR loci in transcriptome were identified by MISA (Haas et al. 2013) (http://pgrc.ipkgatersleben.de/misa/misa.html). The repeat units were set as follows: 10 repeat units for mono-nucleotides, 6 for di-nucleotides, and 5 for tri-, tetra-, penta-, and hexa-nucleotides. SSR primers were designed using the Primer 3 (Untergasser et al.2012) version 2.3.5 with default parameter values (http://probes.pw.usda.gov/cgi-bin/batchprimer3/batchprimer3.cgi) and the primers were synthesized by Beijing Rui Bo Xing Ke Biotechnology Co., LTD (Beijing, China).

EST-SSR Amplification and Validation

Based on the obtained EST-SSRs, eight individuals of *Aesculus* (Table S8) were selected for polymorphism analyses. The genomic DNA was extracted with FastPure Plant DNA Isolation Mini Kit (Vazyme Biotech Co., Ltd.). The PCR reaction was carried out in volume of 20 μL, which contained 2 × Mix (10 μL), M13 forward primer (0.2 μL), reverse primer (0.3μL), forward primer (0.1μL), DNA (1μL) and add H_2O up to 20 μL. PCR amplification was carried out at 95 ℃ for 5 min, followed by 20 denaturation cycles (95 ℃ for 30 s, 55 ℃ for 30 s, and 72 ℃ for 30 s. then 95 ℃ for 30 s, 52 ℃ for 30 s, 72 ℃ for 30 s)and a final extension for 10 min at 72 ℃. The PCR products were verified through capillary electrophoresis (ABI manufacturer). Electrophoretograms of capillary electrophoresis were read and analyzed through Gene Marker V2.2.0 (https://genem arker. softw are.infor mer.com/2.2/) software. The number of alleles (*Na*), the number of effective alleles (*Ne*), Observed

heterozygosity (*Ho*), expected heterozygosity (*He*), and Shannon's information index (*I*) were calculated using the GenALEX6.5 software (Peakall and Smouse. 2012).

Results

De novo assembly

A total 758,298,28 raw reads were generated by the Illumina HiSeq TM2000 sequencing of *A. chinensis* Bunge. To perform the de novo assembly, 748,057,34 clean reads were used to generate 888,69 transcripts. The transcripts lengths ranged from 301 to 18,146 bp, with a mean length of 1550 bp and an N50 of 2109 bp (Table 1). The length of 64,576 (72.66%) transcripts ranged from 301 to 2000 bp (Fig. 1A). The sequencing result was accurate and can be used for subsequent transcriptome assembly. The transcripts were assembled into 333,65 unigenes, with N50 length of 1955bp and average length of 1348bp, which were longer than the *Lycium barbarum* with N50 and average length was 1579bp and 1049bp (Chen et al.2017). Among the 333,65 unigenes, 26,118 (78.27%) ranged in size from 301bp to 2 kb and 7247 (21.72%) were longer than 2 kb (Fig. 1A). The number of unigenes splicing length of 301-400bp was the largest (4141), followed by 401-500bp (3855), and the number of genes decreased with the increase of gene length (Fig. 1B).

Table 1
Summary of sequence assembly using Illumina sequencing

category	Items	Value
Reads	Total raw reads	758,298,28
	Total clean reads	748,057,34
	Clean base	11.22Gb
	Error rate (%)	0.03
	Q20 percentage (%)	98.05
	Q30 percentage (%)	94.18
	GC percentage (%)	45.41
Transcripts	Min length	301
	Max length	18146
	Mean length	1550
	N50	2109
	Total nucleotides (nt)	137,750,895
Unigenes	Min length	301
	Max length	18146
	Mean length	1348
	N50	1959
	Total nucleotides (nt)	449,654,05
note: Q20/Q30: represents the correct base identification rate of 99%/99.9%		
N50: Arrange all the sequences from long to short, and add the length of the sequence in that order. when the added length reaches 50% of the total length of the sequence, the length of the last sequence.		

Gene functional annotation

The 23820, 19031, 17789, 17789, 9054 and 6430 unigenes were annotated according to Nr, Swiss-Prot, PFAM, GO, KO and KOG databases, respectively (Table 2). Among them, unigenes had the most functional annotations in the Nr database (71.39%). In total, 25450 (76.27%) unigenes exhibited homologous matches in at least one database and 3989 (11.95%) unigenes were annotated in all seven public databases.

By comparing and annotating with the Nr database, we can obtain the similarity between the gene sequence of *A. chinensis* Bunge and gene sequences of related species. 23820 (71.39%) unigenes were assigned to Nr database in this study. The e-value distributions of high homologous sequences in Nr database had 48.20% (0-1e-100) with a high homology (Fig. 2A). while the similarity distribution shows that 91.6% of the

sequences with more than 60% similarity and 52.7% of the sequences with more than 80% comparability (Fig. 2B). Furthermore, the highest matches in the Nr database were *Citrus clementina* (18.8%), *Citrus sinensis* (17.3%) and *Citrus unshiu* (14.8%) (Fig. 2C). In addition, 42.1% of the unigenes belonged to other sequences, which may be different from other tree species or belonged to a sequence with a non-coding region.

Table 2
Unigene annotations in the public database

Database	Number of unigenes	Percentage(%)
Nr	23820	71.39
NT	19747	59.18
Swiss-Prot	19031	57.03
PFAM	17789	53.31
GO	17789	53.31
KO	9054	27.13
KOG	6430	19.27
Annotated in all Databases	3989	11.95
Annotated in at least one Databases	25450	76.27
Without annotation genes	7915	23.73
Total unigenes	33365	100.00

Functional Classification by KOG and GO

6430 unigenes were assigned to KOG database, 7164 functional annotations were obtained, which were divided into 25 categories. Posttranslational modification, protein turnover and chaperones (876, 12.22%) was the largest group. Followed by General function prediction only (833,11.62%), Translation, ribosomal structure and biosynthesis (673, 9.39%) (Fig. 3; Table S1). The smallest group was Cell motility (3), followed by extracellular structure (9). Through the functional annotation classification of KOG database, It was found that these unigenes were abundant in protein posttranslational modification, protein conversion, translation, ribosomal structure and biosynthesis. In addition, General function prediction and unknown function accounted for 17.92% of the total, indicating that some genes whose function are unclear and need to be further studied.

A total of 97,978 annotations of 17,789 sequenced genes were obtained in GO database, which were divided into 3 categories and 56 subcategories. The unigenes were classified into biological processes categories (46,486, 47.45%), Followed by cellular components (29,402, 30.00%) and molecular functions (22,090, 22.55%). Under the category of biological process, the largest group were cellular process (10,409, 22.39%) and metabolic process (9518, 20.47%), Followed by single organism process (7596, 16.34%), biological regulation (3760, 8.09%), regulation of biological process (3465, 7.45%) and localization (3071, 6.61%) (Fig. 4; Table S2). In the cellular component, cells and cell parts accounted for 39.02%, the organelle and organelle part accounted for 20.34%, and the membrane and membrane part accounted for 20.53%. In the cluster of molecular functions, binding (10,375, 46.97%) and catalytic activities (8149, 36.89%) were the largest categories.

Functional Classification by KEGG

The KEGG pathway database is a knowledge base for the systematic analysis of gene functions in specific organisms. In the study, 6719 with important matches in the database were divided into 5 main categories including 129 KEGG pathways (Table S3; Fig. 5). The largest category was Metabolism (3598;53.55%), followed by Genetic information processing (1926; 28.66%). It indicated that active metabolic processes were arising in the early growth of leaves and flowers. A total of 11 sub-categories were included in the Metabolism, and the largest categories was Carbohydrate metabolism (19.57%), followed by Overview (13.42%) and Amino acid metabolism (12.20%). Genetic information processing was divided into four categories, the largest categories was Translation (44.44%), followed by Folding, classification and degradation (31.41%). Among the 129 KEGG pathways, the top ten pathways were listed in Table 3. Ribosome represented the largest number of unigenes (428), followed by Plant-pathogen interaction (358) and Carbon metabolism (277). These data afford a solid foundation for the identification of genes involved in specific metabolic pathways.

Table 3
The top 10 KEGG pathway for the assembled unigenes

KEGG pathway	Pathway ID	Gene number	Function
Ribosome	ko03010	428	Genetic Information Processing
Plant-pathogen interaction	ko04626	358	Organismal Systems
Carbon metabolism	ko01200	277	Metabolism
Plant hormone signal transduction	ko04075	261	Environmental Information Processing
Biosynthesis of amino acids	ko01230	242	Metabolism
Protein processing in endoplasmic reticulum	ko04141	213	Genetic Information Processing
Starch and sucrose metabolism	ko00500	195	Metabolism
Spliceosome	ko03040	194	Genetic Information Processing
Endocytosis	ko04144	191	Cellular Processes
RNA transport	ko03013	187	Genetic Information Processing

CDS prediction

Unigenes were aligned in the Nr and Swiss-Prot databases for CDS prediction. The non-hit unigenes were then subjected to ESTS can (version 3.0.3). In total, the CDS of 22,638 unigenes were predicted in Nr and Swiss-Prot databases. The majority of predicted CDS were in the ranges of 200–600 bp (7321, 32.34%), followed by CDS of 601–1000 bp (4968, 21.95%), CDS of 1001–1500 bp (4406, 19.46%), CDS of 1501-2000bp (2260, 9.98%) and CDS of > 2000 bp (2331, 10.42%) (Fig. 6A; Table S4). The majority of CDS of unigenes (5073, 68.80%) predicted by the ESTS can software were less than 600bp (Fig. 6B; Table S5).

Frequency and Distribution of SSRs in the unigenes

By searching the unigenes SSR loci of 33,365 (44,965,405 bp) -length *A. chinensis* Bunge, A total of 16,103 SSR loci were found. The SSR frequency was 48%, and the distribution density of one SSR per 2.79 kb. SSR repeat types were distributed from single nucleoside to hexanucleotide, of which the largest repeat type was single nucleotide, followed by dinucleotide and trinucleotide, and their numbers were 7217, 4707 and 3715, respectively, The sum of three accounting for 97.12% of the total SSR type (Table 4). Further analysis showed that the main type of single nucleotide was A/T, accounting for 99.78%. The most dinucleotide repeats were AG/CT (58.49%), followed by AT/AT (34.65%). There are 10 repeat types in trinucleotides, among which the number of AAG/CTT (29.37%) was the largest. The main tetranucleotides including ACAT/ATGT, AAAT/ATTT, AAAG/CTTT, AAAC/GTTT (Fig. 7A).

In addition, Among the single nucleotides, the most repeats were 9–12 (5038, 69.81%), followed by 13–16 (1411,19.55%). Dinucleotides were mainly 5–8 repeats (2884, 61.27%), followed by 9–12 repeats (1114, 23.67%). Trinucleotides were mainly composed of 5–8 repeat units (3529, 94.99%) (Table 5; Fig. 7B; Table S6). The others (tetra-/ Penta-/ Hexa-) were also mainly 5–8 repeats. Overall, the main repeat units of SSR motif (from single nucleoside to hexanucleotide) were 5–8 and 9–12 repeats, and the sum of them accounts for 81.93% of the total (Table 5). According to the location of SSR locus in the gene, SSR can be divided into three categories, located in 3'UTR, 5'UTR and CDS. 1290 SSRs are located in CDS, and the trinucleotides were 1088, accounting for 84.34%. 7138 SSRs were located in non-coding regions, of which 4203 are 5'UTR and 2935 are 3'UTR (Fig. 7C). The number of dinucleotides and trinucleotides in 5'UTR were significantly higher than that in 3'UTR.

Table 4
Summary of simple sequence repeat (SSR)

Statistical item	Number	Percentage (%)
Total number of sequences examined	33365	
Total size of examined sequences (bp)	44965405	
Total number of identified SSRs	16103	
Number of sequences containing more than 1 SSR	3452	
Number of SSRs present in compound formation	1465	
Mono-nucleotide	7217	44.82
Di-nucleotide	4707	29.23
Tri-nucleotide	3715	23.07
Tetra-nucleotide	213	1.32
Penta-nucleotide	85	0.53
Hexa-nucleotide	166	1.03

Table 5
Distribution of SSR motif in the transcriptome of *A. chinensis* Bunge

Repeat number	Motif type						Total	Ratio (%)
	Mono-	Di-	Tri-	Tetra-	Penta-	Hexa-		
5–8	-	2884	3529	205	84	148	6850	42.54
9–12	5038	1114	167	7	-	17	6343	39.39
13–16	1411	453	11	1	1	1	1878	11.66
17–20	497	183	6	-	-	-	686	4.26
21–24	165	47	2	-	-	-	214	1.33
25–28	59	13	-	-	-	-	72	0.45
29–32	18	6	-	-	-	-	24	0.15
> 32	29	7	-	-	-	-	36	0.22
total	7217	4707	3715	213	85	166	16103	
Proportion%	44.82	29.23	23.07	1.32	0.53	1.03		100.00

Validation of EST-SSR Markers

Based on the designed 5842 EST-SSR primers, 58 primer pairs (Table S7) were randomly selected and were successfully used for validation via PCR amplification. Among the 58 primer pairs, 48 were highly successful amplified by *A. chinensis* Bunge genomic DNA, with an amplification efficiency rate of 82.7%. The rest of 10 primers pairs failed to amplify PCR products at different annealing temperatures. Eight *Aesculus* samples were used as PCR templates, 31of the 48 primer pairs were polymorphic, and 1 primer was recognized as monomorphic, and the polymorphism frequency was 64.6%.

The 31 polymorphic SSR markers were used to assess the eight *Aesculus* samples, The polymorphisms of loci were detected by capillary electrophoresis (Fig. 8). As presented in Table 6, 167 alleles were detected and the number of alleles varied from 2 to 9, with a mean of 5.387 alleles per locus. The Number of effective allele (*Ne*) ranged from 1.438 to 6.095 with a mean of 3.628, The Shannon's information index (*I*) ranged from 0.483 to 1.960, with a mean of 1.380. The Observed heterozygosity (*Ho*) ranged from 0.000 to 0.875, with a mean of 0.476. and expected heterozygosity (*He*) ranged from 0.305 to 0.844 with a mean of 0.680. Except for P6 and P28 primers, the values of expected heterozygosity (*He*) were all higher than 0.5, suggesting that genetic diversity were abundant of these EST-SSR markers.

Table 6
Characteristics of the 31 EST-SSR markers in A. chinensis Bunge

primer	Forward primer (5′-3′)	Reverse primer (5′-3′)	Tm (°C)	SSRs	*Na*	*Ne*	*I*	*Ho*	*He*
P1	AGGCTCTTTCTGAGGTTGGC	AAGAGCCTGTTCACCACTGG	59.9	(AC)6	5	3.200	1.332	0.625	0.688
P2	AGAGTTGGGGAGGATGCTCT	ACGTATGCACCCGAGACTTC	59.9	(TA)8	5	2.783	1.230	0.625	0.641
P3	CCAACTCGGCCATCTTGACT	GGAAGAAATGGCGCATGCTT	59.9	(TA)6	7	4.923	1.750	0.750	0.797
P4	CACCACAAGCCTTCCATCCT	CCATCCATGTGCCTGACTGT	60.0	(CCT)5	6	3.765	1.511	0.875	0.734
P5	TCTGCTTTGACCTCGCCATT	GCTGTCTCCTTTGACGGTGA	60.0	(TA)14	7	6.095	1.862	0.375	0.836
P6	AAATGCTCACTCTCTCCGGC	AGACGGCACAGATGGTTTGT	60.0	(GC)6	2	1.438	0.483	0.375	0.305
P7	CCTCCTCCTCCATCTGCTCT	GCACCCACCTGTCAACCTTA	60.0	(TC)7	5	2.000	1.037	0.250	0.500
P8	GGCAAGAGGGTGCTTGAGAT	AAGGATGGCCACAGCTCTTC	60.0	(TG)6	5	3.556	1.423	0.375	0.719
P9	TGGTTGCTCTTTCCAGCCAT	TTTGGCCCTTCATTGGAGCA	60.0	(GAAGT)5	6	3.765	1.511	0.625	0.734
P10	ACCTCGTGGATTTGGCAGAG	CGCTGTCAACGACTCCATCT	60.1	(TC)11	6	3.556	1.488	0.500	0.719
P11	ATTGGGCTCGGTCTTCTTGG	AGCCACCCGAGTCTATGACT	60.0	(AT)7	5	3.657	1.424	0.250	0.727
P12	CCCTGGAGGCACAAACTGAT	AGTCCACATCTCACTGCTGC	60.0	(AG)14	8	4.923	1.841	0.750	0.797
P13	CTTCCGCTTGCAGCATTTGT	CCGGGATTACATGGACGGTT	59.9	(TCT)6	3	2.133	0.900	0.000	0.531
P14	TGGTGGCATGCCTTCTTCTT	AGTTGCTGCTGCTGTTGTTG	59.9	(CTT)6	3	2.462	0.974	0.250	0.594
P15	CTTCCGATGTGCCTTCCAGT	ATTCATCGCTACCGCTACCG	60.0	(GGAGAG)5	4	2.844	1.163	0.125	0.648
P16	CCAGGCGCATAGTCATGACT	TGGCATGCACTTGACAGGAT	59.9	(AT)6	4	3.122	1.223	0.375	0.680
P17	ATTCCGCACAACTCACCAGT	GCAGGAAACAGAGCACTTGC	60.0	(CT)9	8	6.095	1.927	0.625	0.836
P18	TCTGTCTACCAAGCGGCAAG	ATCGTCCCCATCAGAGGTCA	60.0	(CT)8	6	4.000	1.548	0.625	0.750
P19	CACCGCACCCACTTACTCTT	AGCAACTGCCCAAGATCCTC	60.0	(TTTA)5	5	3.368	1.369	0.375	0.703
P20	TACTGGTCTGAGGCAGCTCT	ACCTTGCCCAGATTACCTGC	60.0	(TGC)5	2	2.000	0.693	0.000	0.500
P21	ACATCCAGACAGAAGCCAGC	TGCTGCAGGAATAACAGGCA	60.0	(CT)7	6	4.129	1.581	0.500	0.758
P22	GTGGCCCATACCAGTAGAGC	GCTTTTGGGTGCTGTTGAGG	59.9	(AG)13	8	6.400	1.960	0.625	0.844
P23	GTCTGTCCGTCCGCGATTAT	CGCAAACTTGTCGACGGATC	59.9	(GA)8	7	5.333	1.787	0.625	0.813
P24	CGGTCCCATTCAACTCACGA	TTTCACGATCTCCGGCGTAG	60.0	(TC)8	8	5.818	1.906	0.625	0.828
P25	AGCTTGCTTTCTCGAGCCAT	ATGAGAGTGCGATCGAGCTG	60.0	(AGG)6	5	2.977	1.300	0.375	0.664
P26	TGGTGGCCTTCAACTGGATC	CAATCCAATGCCGAGGGAGT	60.0	(GGC)5	4	2.246	1.041	0.625	0.555
P27	GCAGCATCTCCAGCCTTAGT	AATTTGGTTGCAAGCGCCTT	59.8	(GA)13	9	5.333	1.923	0.875	0.813
P28	TCCCACCCTCCTCTTCTCTG	AGAGCCACTGTACAACCTGC	60.0	(CT)6	3	1.684	0.736	0.375	0.406
P29	AGGCAAGTACAGCAAGCCAT	TTTCTCCAGTCAGTCGTGGC	60.0	(TA)10	5	3.122	1.354	0.125	0.680
P30	AGCCAGAGAGAGAGGAGGTG	CTCCCACACTCCTTGTTGCT	60.0	(TC)10	3	2.462	0.974	0.500	0.594
P31	TACAACGCACCTACACCACC	ATGAGCAATTCCAGCAGCCT	60.0	(TC)13	7	3.282	1.527	0.750	0.695
mean					5.387	3.628	1.380	0.476	0.680

Note. Na: Number of alleles; Ne: Number of effective alleles; I: Shannon's information index; Ho: Observed heterozygosity; He: Expected heterozygosity

Discussion

Transcriptome characteristics of A. chinensis **Bunge**

Transcriptome assembly quality was assessed by several criteria, including N50, average length and examination of the RNA-Seq read (Al-Qurainy et al. 2019). The N50 was used to assess transcriptome assemblies, in which a high value corresponds to a high quality. In this research, a total of 33,365 unigenes with N50 (1955bp) and average length of 1348bp were obtained, which was longer than previously reported such as *A. chinensis* (N50 = 1807bp, mean length = 999.86bp) (Wei et al. 2019). *Platanus acerifolia* (N50 = 1579bp, mean length = 1015.15bp) (Li et al. 2019). *Pigeonpea* (N50 = 1393bp, mean length = 396bp) (Nigam et al. 2017). *Sodom apple* (N50 = 1733bp, mean length = 858.83 bp) (Muriira et al. 2015). The Q20 of 98.05% with GC of 45.41% reflects a high quality sequencing run too. This indicated sequencing data and transcriptome assembly for *A. chinensis* Bunge had high quality in the present study. The number of unigenes splicing length of 301-400bp was the largest, that was common in de novo transcriptomes assembly in non-model plants, just like the length of 200-400bp was the largest in *A. chinensis* (Wei et al. 2019), A half of the unigenes were short (200–500 bp) in *Caladium* (Cherukupalli et al.2016; Wang et al.2016; Ma et al.2015). The main reason for the occurrence of short unigenes including lack of reference genome, the use of stringent parameters and highly complex transcriptomes (Cao and Deng 2017). By comparing the homology with Nr protein database, *A. chinensis* Bunge has higher homology with *Citrus clementina* and *Citrus sinensis*, which was consistent with Wei 's conclusion (Wei et al.2019). Maybe there are fewer protein sequences of *Aesculus* in the public database while there are more protein sequences related to *Citrus*. The transcriptome data will offer a solid basis for further gene mining and marker-assisted selection breeding for *A. chinensis* Bunge.

Transcriptome annotation of A. chinensis Bunge

The assembled unigenes were successfully annotated to 7 public databases, 25,450 unigenes (76.27%) were annotated to at least one database. But 23.73% unigenes were not annotated to any database, it may be that these genes of the genus *Aesculus* have only short non-coding regions or fewer gene annotations in the public database. By comparing with KOG database, 6430 unigenes were divided into 25 categories, of which most categories were connected to Posttranslational modification, protein conversion and molecular chaperone (12.22%), followed by General functional prediction only (11.62%) and Translation, ribosomal structure and biosynthesis (9.39%), which were greatly similar to *Apocynum venetum L* (Yuan et al.2020). A total of 97,978 annotations were obtained from the GO database, which were assigned to three categories: biological process, cellular components and molecular function. In the biological process, the largest group was cellular process (22.39%), followed by metabolic process (20.47%) and single organism process (16.34%). Coincidentally, the main group were cellular process (22.04%), metabolic process (20.99%) and single organism process of the biological process in *Apocynum venetum L* (Yuan et al.2020) too. The most abundant categories were binding and catalytic activity in molecular function, which was similar to *Apocynum venetum L* and *Platanus acerifolia* (Yuan et al.2020; Li et al.2019). Metabolic pathway was the largest category in KEGG database, among which the largest three groups were carbohydrate metabolism, overview and amino acid metabolism with a proportion of 19.57%, 13.42% and 12.20%, respectively, which were similar to *Apocynum venetum L* (Yuan et al.2020), *Platanus acerifolia* (Li et al.2019) and *guar* (Tanwar et al.2017). These transcriptome annotations will offer important resources for the study of specific physiological processes, gene structure and function, and metabolic pathways in *A. chinensis* Bunge.

SSR characterization of A. chinensis Bunge

SSR molecular markers are codominant, specific, multi-allelic, highly polymorphic and can be transferred between species within genera (Varshney et al. 2005). To date, there has been no report of EST-SSR identification in *A. chinensis* Bunge. In the study, A total of 16,103 SSR loci were found and the largest number of SSR repeat types was single nucleotides, followed by dinucleotides and trinucleotides. The distribution density of SSR in the *A. chinensis* Bunge unigenes was 1/2.79kb. which was much higher than those obtained in *R.rex* (1/5.65kb) (Zhang et al. 2017) and *tree peony* (1/9.24kb) (Wu et al. 2014), but lower than the frequency of *Stephanandra incisa* (1/1.60kb) (Zhang et al. 2021). Among the SSR repeat types, The main dinucleotide types were AG/CT (58.49%), AT/AT (34.65%), and trinucleotides were mainly AAG/CTT, consistent with *Apocynum venetum L* (Yuan et al. 2020). Meanwhile, the most types of SSR are AG/CT (15.4%), AAG/CTT (9.8%) and AT/AT (6.5%) in *Pistacia vera L*. (Karcı et al. 2020), and AG/CT is also dominant in *Jerusalem artichoke* (32.9%) (Yang et al. 2019), AAG/CTT was also the most (12.89%) of trinucleotide in *Chelidonium majus L*. (Pourmazaheri et al. 2019). The above results indicate that dinucleotide (AG/CT) is the most common and trinucleotide (AAG/CTT) is the most common in plants (Wei et al. 2011; Liu et al. 2015). Further analysis, 1290 SSRs (trinucleotide was1088, 84.34%) were located in CDS. Similarly, trinucleotide (444, 68.7%) and dinucleotide (242, 49.4%) repeats were preferentially present in the CDS of *Blackgram* (Raizada and Souframanien. 2019). 7138 SSRs were located in the non-coding region, and the number of di- and tri-nucleotides in 5'UTR is significantly larger than that in 3'UTR. Interestingly, in *Blackgram*, the 5'UTRs contained more trinucleotides than 3'UTRs too (Raizada and Souframanien. 2019; Souframanien and Reddy. 2015). The large number of SSRs can greatly facilited largescale genotyping studies such as genetic diversity assessment and association mapping for important traits in *A. chinensis* Bunge.

Validation of EST-SSR Makers

58 pairs of primers were selected for PCR validation, 48 (82.7%) primers produced clear bands. The PCR amplification efficiency was higher than *Boea clarkeana* (81.08%) (Wang et al. 2017), *Hevea brasiliensis* (75%) (Triwitayakorn et al. 2011) and *R.rex* (50%) (Zhang et al. 2017). But lower than the rate reported for *Haloxylon ammodendron* (85%) (Long et al.2014), *Caragana korshinskii Kom* (93.7%) (Long et al.2015) and *Lycium barbarum* (88%) (Chen et al.2017), which imply the EST-SSR makers have moderate cross-transferability in *Aesculus* genus. The polymorphism

frequency was 64.6% among the eight *Aesculus* samples, and this was much higher than that in *R.rex* (36%) (Zhang et al. 2017) and *Stephanandra incisa* (29%) (Zhang et al. 2021). Suggesting these primers can be used in molecular-assisted breeding.

167 alleles were identified among the eight *Aesculus* samples using the 31 EST-SSR polymorphic makers. The alleles ranged from 2 to 9 per locus with average 5.387 alleles. The observed heterozygosity (*Ho*) averaging 0.476, expected heterozygosity (*He*) averaging 0.680, which were higher than that of EST-SSRs in *Apocynum venetum L* (*Ho* = 0.336, *He* = 0.342) (Yuan et al. 2020). *Stephanandra incisa* (*He* = 0.434) (Zhang et al. 2021), *Pistachiovera* L (*Ho* = 0.39, *He* = 0.65) (Karcı et al. 2020). *Torreya grandis* (*He* = 0.432) (Zeng et al. 2018). The Shannon's information index (*I*) can also reflect the polymorphism of markers, ranged from 0.483 to 1.960, with average 1.380. This finding indicated a high level of informativeness within these EST-SSR markers and can be used for a wide range of genetic analysis among species and evolutionary adaptation at the species level.

Conclusion

To our knowledge, this is the first study in which *de novo* transcriptomes sequencing of *A. chinensis* Bunge were used to develop EST-SSR markers. These results not only provide a significant information resource for genetic improvement in *A. chinensis* Bunge, but also offer a powerful molecular tool for genetic relationship analysis in other *Aesculus* species.

Abbreviations

GO: Gene ontology; KEGG: Kyoto encyclopedia for genes and genomes; Nr: Non-redundant;

KOG: Clusters of orthologous groups for eukaryotic complete genomes; Pfam: Protein family;

CDS: Coding DNA sequence; EST-SSR: Expression sequence tag-Simple sequence repeat

Declarations

Ethics approval and consent to participate Not applicable

Consent for publication Not applicable

Availability of data and materials Data and materials are available from the corresponding author on reasonable request

Competing interests The authors declare that they have no competing interests.

Funding This study was financially supported by grant 2020LZGC011 from Shandong Agricultural Improved seed Project, China.

Authors' contributions Xin Mao, Tianxu Sun, Cuilan Liu, Yuanshuai Zhang and Yuanfu Dong collected the plant material; Xiuhong Mao and Yuling Li conceived and designed the experiments; Yuling Li, Xin Mao and Xiuhong Mao analyzed the data; Yuling Li drafted the paper; Yuling Li and Xin Mao prepared figures 1-8; Xiuhong Mao edited and provided critical review of the manuscript. All authors read and approved the final manuscript.

Acknowledgements We are very grateful to Professor Fang-fang Ma of the Forestry College of Shandong Agricultural University for her revision of the language in the paper

References

1. Chinese Flora. Editorial Committee of Chinese Flora, Chinese Academy of Sciences. Beijing: Science Publishing House. 1982,46:276-277. (in Chinese)
2. Liu L J, Zhou H. Pharmacological effects of aescin and its clinical application. Progress in Modern Biomedicine.2010,(5): 957-960.(in Chinese)
3. Kim SE, Kim TH, Park SA, Kim WT, Park YW, Ahn JS, Jeong M, Kim MY, Seo K.Efficacy of horse chestnut leaf extract ALH-L1005 as a matrix metalloproteinase inhibitor in ligature-induced periodontitis in canine model. J Vet Sci.2017,18(2): 245-251. DOI 10.4142/jvs.2017.18.2.245.
4. Chen C, Wang S, Chen J, Liu X, Zhang M, Wang X, Xu W, Zhang Y, Li H, Pan X, Si M. Escin suppresses HMGB1-induced overexpression of aquaporin-1 and increased permeability in endothelial cells. FEBS Open Bio.2019,9(5): 891-900. DOI 10.1002/2211-5463.12622.
5. Zhang Z Y, Chen Y,Jiang X B , Zhu P, Li L, Zeng Y L & Tang T M. The complete chloroplast genome of Aesculus chinensis. Mitochondrial DNA Part B.2019, 4(1): 1955-1956. DOI 10.1080/23802359.2019.1617056

6. Liu Z, Zhang J, Zhou Y, Liu Y, Hu Z, Zheng G, Shi Z. The complete chloroplast genome of Aesculus chinensis var. wilsonii. Mitochondrial DNA B Resour.2020, 5(3): 2547-2549. DOI 10.1080/23802359.2020.1780972.
7. Severin, A.J., Woody, J.L., Bolon, Y.-T., Joseph, B., Diers, B.W., Farmer, A.D., Muehlbauer, G.J., Nelson, R.T. Grant, D., Specht, J.E. RNA-Seq Atlas of Glycine max: A guide to the soybean transcriptome. BMC Plant Biol.2010, 10, 160. DOI 0.1186/1471-2229-10-160.
8. Simon S A, Zhai J X, Nandety R S, et al. Short-read sequencing technologies for transcriptional analyses. Annual review of Plant Biology, 2009,60: 305-333. https://doi.org/10.1146/annurev.arplant.043008.092032.
9. Reddy B,Patel A K,Singh K M,et al. De novo assembly and transcriptome characterization of canine retina using high-throughput sequencing. Genetics Research International,2015: 638-679. https://doi.org/10.1155/2015/638679.
10. Wang Z, Gerstein M, Snyder M.RNA-SEQ: A revolutionary tool for transcriptomics. Nat Rev Gene.2009, 10(1): 57-63. DOI 10.1038/nrg2484.
11. Yang Y, Xu M, Luo Q, Wang J, Li H. De novo transcriptome analysis of Liriodendron chinense petals and leaves by illumina sequencing. Gene.2014,534(2): 155-162. DOI 10.1016/j.gene.2013.10.073.
12. Cheng Y, Liu J, Zhang H, Wang J, Zhao Y, Geng W. Transcriptome analysis and gene expression profiling of abortive and developing ovules during fruit development in hazelnut. PLoS One.2015,10(4): e0122072. DOI 10.1371/journal.pone.0122072.
13. Guo YQ, ChangX J, Zhu C, ZhangS T, Li X Z, Fu H F, Chen C S, Lin Y L, Lai Z X. De novo transcriptome combined with spectrophotometry and gas chromatography-mass spectrometer (GC-MS) reveals differentially expressed genes during accumulation of secondary metabolites in purple-leaf tea (Camellia sinensis cv Hongyafoshou). J Hortic Sci Biotechnol.2018,94(3): 349-367. DOI 10.1080/14620316.2018.1521708
14. Qi X, Ogden EL, Ehlenfeldt MK, Rowland LJ. Dataset of de novo assembly and functional annotation of the transcriptome of blueberry (Vaccinium spp.). Data Brief.2019,25: 104390. DOI 10.1016/j.dib.2019.104390
15. Naganeeswaran S, Fayas TP, Rajesh MK..Dataset of transcriptome assembly of date palm embryogenic calli and functional annotation. Data Brief.2020,31: 105760. DOI 10.1016/j.dib.2020.105760.
16. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology.2011,29(7):644-652. DOI 10.1038/nbt.1883.
17. Taheri S, Abdullah, T.L.,Ahmad, Z.,Abdullah, N.A.P. Effect of acute gamma irradiation on Curcuma alismatifolia varieties and detection of DNA polymorphism through SSR Marker. BioMed Res. Int. 2014, 2014. https://doi.org/10.1155/2014/631813.
18. Liu Z, Chen T, Ma L, Zhao Z, Zhao PX, Nan Z, Wang Y. Global. Transcriptome Sequencing Using the Illumina Platform and the Development of EST SSR Markers in Autotetraploid Alfalfa. PLoS One.2013,8(12): e83549. DOI 10.1371/journal.pone.0083549.
19. Zhou Q, Luo D, Ma L, Xie W, Wang Y, Wang Y, Liu Z. Development and cross-species transferability of EST-SSR markers in Siberian wildrye (Elymus sibiricus L.) using Illumina sequencing. Sci.Rep.2016,6: 20549. DOI 10.1038/srep20549.
20. Taheri S, Lee Abdullah T, Yusop MR, et al. Mining and development of novel SSR markers using next generation sequencing (NGS) data in plants. Molecules. 2018, 23:399. DOI 10.3390/molecules23020399.
21. He D, Zhang J, Zhang X, He S, Xie D, Liu Y, Li C, Wang Z, Liu Y. Development of SSR markers in Paeonia based on De Novo transcriptomic assemblies. PLoS One. 2020, 15(1): e0227794. DOI 10.1371/journal.pone.022779424.
22. Chen C, Xu M, Wang C, Qiao G, Wang W, Tan Z, Wu T, Zhang Z. Characterization of the Lycium barbarum fruit transcriptome and development of EST-SSR markers. PLoS One.2017,12(11): e0187738. DOI 10.1371/journal.pone.0187738.
23. Al-Qurainy F, Alshameri A, Gaafar AR, Khan S, Nadeem M, Alameri AA, Tarroum M, Ashraf M. Comprehensive Stress-Based. De Novo Transcriptome Assembly and Annotation of Guar (Cyamopsis tetragonoloba (L.) Taub.): An Important Industrial and Forage Crop. Int J Genomics. 2019, 7295859. DOI 10.1155/2019/7295859.
24. Wei Y D, Xiong C, Zhang T Y, Gao H, Yin Q G, Yao H, Sun W, Hu Z G, Chen S L.Synthesis of triterpenoid saponins from Aesculus chinensis based on transcriptome data. China journal of chinese Materia Medica.2019,44(6): 1135-1144.[in chinese]
25. Li F, Wu C, Gao M, Jiao M, Qu C, Gonzalez-Uriarte A, Luo C. Transcriptome sequencing, molecular markers, and transcription factor discovery of Platanus acerifolia in the presence of Corythucha ciliata. Sci Data.2019,6(1): 128. DOI 10.1038/s41597-019-0111-9.
26. Nigam D, Saxena S, Ramakrishna G, Singh A, Singh NK, Gaikwad K.*De novo* Assembly and Characterization of *Cajanus scarabaeoides* (L.) Thouars Transcriptome by Paired-End Sequencing. Front Mol Biosci.2017, 4:48. DOI 10.3389/fmolb.2017.00048.
27. Muriira N G, Xu W, Muchugi A, Xu J, Liu A. De novo sequencing and assembly analysis of transcriptome in the Sodom apple (Calotropis gigantea). BMC Genomics.2015,16(1):723. DOI 10.1186/s12864-015-1908-3.
28. Cherukupalli N, Divate M, Mittapelli SR, Khareedu VR, Vudem DR. De novo Assembly of Leaf Transcriptome in the Medicinal Plant Andrographis paniculata. Front Plant Sci.2016,7:1203. DOI 10.3389/fpls.2016.01203.
29. Wang Y, Li X, Zhou W, Li T, Tian C.De novo assembly and transcriptome characterization of spruce dwarf mistletoe Arceuthobium sichuanense uncovers gene expression profiling associated with plant development. BMC Genomics.2016,17(1):771. DOI 10.1186/s12864-016-3127-y.

30. Ma X, Wang P, Zhou S, Sun Y, Liu N, Li X, Hou Y.De novo transcriptome sequencing and comprehensive analysis of the drought-responsive genes in the desert plant Cynanchum komarovii. BMC Genomics.2015,16:753. DOI 10.1186/s12864-015-1873-x.
31. Cao Z, Deng Z. De Novo Assembly, Annotation, and Characterization of Root Transcriptomes of Three Caladium Cultivars with a Focus on Necrotrophic Pathogen Resistance/Defense-Related Genes. Int J Mol Sci. 2017,18(4):712. DOI 10.3390/ijms18040712.
32. Yuan N, Li M, Jia C. De novo transcriptome assembly and population genetic analyses of an important coastal shrub, Apocynum venetum L. BMC Plant Biol.2020,20(1):408. DOI 10.1186/s12870-020-02626-7.
33. Li F, Wu C, Gao M, Jiao M, Qu C, Gonzalez-Uriarte A, Luo C. Transcriptome sequencing, molecular markers, and transcription factor discovery of Platanus acerifolia in the presence of Corythucha ciliata. Sci Data.2019,6(1):128. DOI 10.1038/s41597-019-0111-9.
34. Tanwar U K, Pruthi V, Randhawa G S. RNA-Seq of Guar (*Cyamopsis tetragonoloba*, L. Taub.) Leaves: *De novo* Transcriptome Assembly, Functional Annotation and Development of Genomic Resources. Front Plant Sci.2017,8:91. DOI 10.3389/fpls.2017.00091.
35. Varshney, R. K.Sigmund, R.,Borner, A.,Korzun, V.Stein, N.Sorrells, M. E.,Langridge, P.Graner,A. Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. Plant Sci.2005,168: 195-202. DOI: 10.1016/j.plantsci.2004.08.001
36. Zhang Y, Zhang X, Wang YH, Shen SK. *De Novo* Assembly of Transcriptome and Development of Novel EST-SSR Markers in *Rhododendron rex* Lévl. through Illumina Sequencing. Front Plant Sci.2017,8:1664. DOI 10.3389/fpls.2017.01664
37. Wu J., Cai C F, Cheng F, Cui H L, Zhou H. Characterisation and development of EST-SSR markers in tree peony using transcriptome sequences. Mol. Breed.2014,34(4): 1853-1866. DOI 10.1007/s11032-014-0144-x
38. Zhang C, Wu Z, Jiang X, Li W, Lu Y, Wang K. De novo transcriptomic analysis and identification of EST-SSR markers in Stephanandra incisa. Sci Rep.2021,11(1):1059. DOI 10.1038/s41598-020-80329-7.
39. Karcı H, Paizila A, Topçu H, Ilikçioğlu E, Kafkas S. Transcriptome Sequencing and Development of Novel Genic SSR Markers From *Pistacia vera* L. Front Genet.2020,11:1021. DOI 10.3389/fgene.2020.01021.
40. Yang S, Sun X, Jiang X, Wang L, Tian J, Li L, Zhao M, Zhong Q. Characterization of the Tibet plateau Jerusalem artichoke (*Helianthus tuberosus* L.) transcriptome by de novo assembly to discover genes associated with fructan synthesis and SSR analysis. Hereditas.2019,156:9. DOI 10.1186/s41065-019-0086-8.
41. Pourmazaheri H, Soorni A, Kohnerouz BB, Dehaghi NK, Kalantar E, Omidi M, Naghavi MR. Comparative analysis of the root and leaf transcriptomes in Chelidonium majus L. PLoS One.2019,14(4): e0215165. DOI 10.1371/journal.pone.0215165.
42. Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D, Lv H, Zhang X. Characterization of the sesame (Sesamum indicum L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. BMC Genomics.2011,12:451. DOI 10.1186/1471-2164-12-451.
43. Liu Y, Zhang P, Song M, Hou J, Qing M, Wang W, Liu C.Transcriptome Analysis and Development of SSR Molecular Markers in Glycyrrhiza uralensis Fisch. PLoS One.2015,10(11): e0143017. DOI 10.1371/journal.pone.0143017.
44. Raizada A, Souframanien J. Transcriptome sequencing, de novo assembly, characterisation of wild accession of blackgram (Vigna mungo var. silvestris) as a rich resource for development of molecular markers and validation of SNPs by high resolution melting (HRM) analysis. BMC Plant Biol.2019,19(1):358. DOI 10.1186/s12870-019-1954-0.
45. Souframanien J and Reddy. De novo assembly, characterization of immature seed transcriptome and development of genic-SSR markers in blackgram[Vigna mungo (L.) Hepper]. PLoS One.2015,10: e0128748. DOI 10.1186/s12870-019-1954-0.
46. Wang Y, Liu K, Bi D, Zhou S, Shao J. Characterization of the transcriptome and EST-SSR development in *Boea clarkeana*, a desiccation-tolerant plant endemic to China. PeerJ.2017, 5: e3422. DOI 10.7717/peerj.3422.
47. Triwitayakorn K, Chatkulkawin P, Kanjanawattanawong S, Sraphet S, Yoocha T, Sangsrakru D, Chanprasert J, Ngamphiw C, Jomchai N, Therawattanasuk K, Tangphatsornruang S. Transcriptome sequencing of Hevea brasiliensis for development of microsatellite markers and construction of a genetic linkage map. DNA Res.2011,18(6):471-82. DOI 10.1093/dnares/dsr034.
48. Long Y, Zhang J, Tian X, Wu S, Zhang Q, Zhang J, Dang Z, Pei XW. De novo assembly of the desert tree Haloxylon ammodendron (C. A. Mey.) based on RNA-Seq data provides insight into drought response, gene discovery and marker identification. BMC Genomics.2014,15(1):1111. DOI 10.1186/1471-2164-15-1111.
49. Long Y, Wang Y, Wu S, Wang J, Tian X, Pei X. De novo assembly of transcriptome sequencing in Caragana korshinskii Kom. and characterization of EST-SSR markers. PLoS One.2015,10(1): e0115805. DOI 10.1371/journal.pone.0115805.
50. Chen C, Xu M, Wang C, Qiao G, Wang W, Tan Z, Wu T, Zhang Z. Characterization of the Lycium barbarum fruit transcriptome and development of EST-SSR markers. PLoS One.2017,12(11): e0187738. DOI 10.1371/journal.pone.0187738.
51. Zeng J, Chen J, Kou Y, Wang Y. Application of EST-SSR markers developed from the transcriptome of *Torreya grandis* (Taxaceae), a threatened nut-yielding conifer tree. PeerJ.2018, 6: e5606. DOI 10.7717/peerj.5606.
52. Reuter JA, Spacek D, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015, 58(4): 586–597. https://doi.org/10.1016/j.molcel.2015.05.004

53. Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc.2013, 8(8): 1494-1512. DOI: 10.1038/nprot.2013.084
54. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3–new capabilities and interfaces. Nucleic Acids Research.2012, 40(15): e115. https://doi.org/10.1093/nar/gks596
55. Peakall R, Smouse PE. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research–an update. Bioinformatics. 2012, 28(1):2537-2539. DOI 10.1111/j.1471-8286.2005. 01155. x

Figures

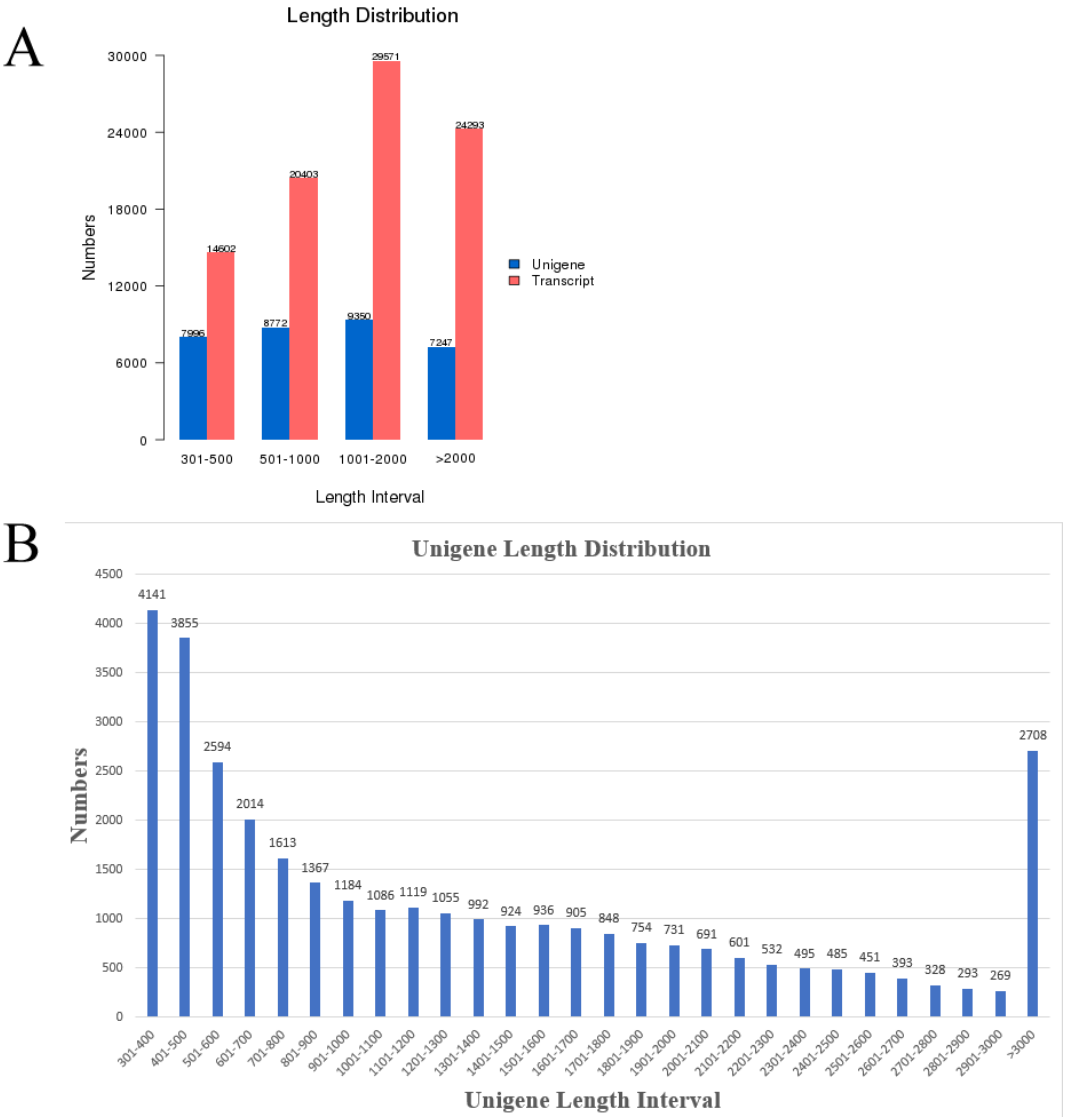

Figure 1

Length distribution of Unigene and Transcripts. (A): Frequency distribution of length of Unigene and Transcript splicing. (B): Unigene splicing length frequency distribution

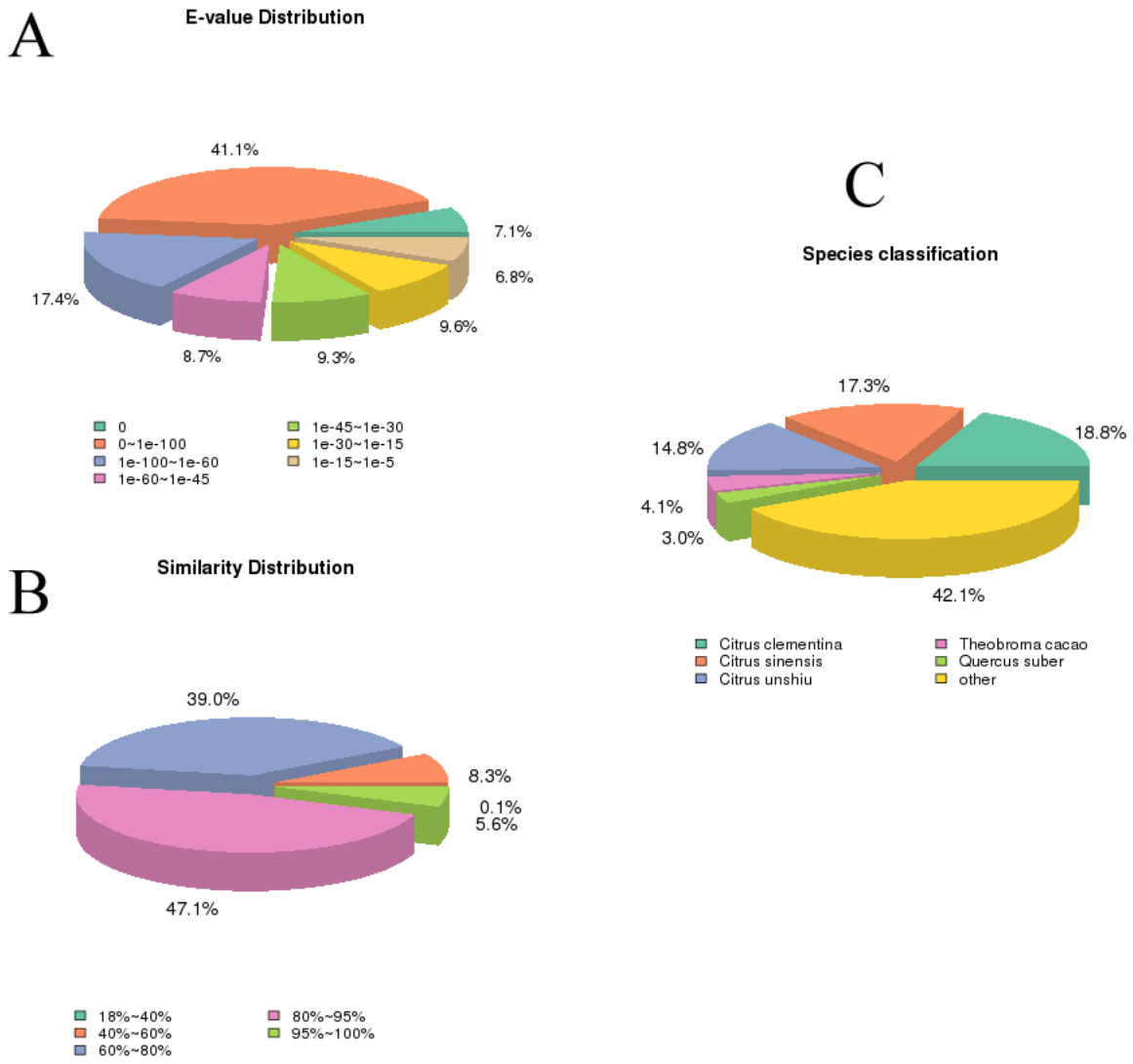

Figure 2

Unigene annotations in the Nr database

(A). E-value distribution of the BLAST x hits against the Nr database for Unigene. (B). Similarity distribution of the BLAST x hits against the Nr database for Unigene. (C). Species distribution of the top BLAST x hits against the Nr database for Unigene.

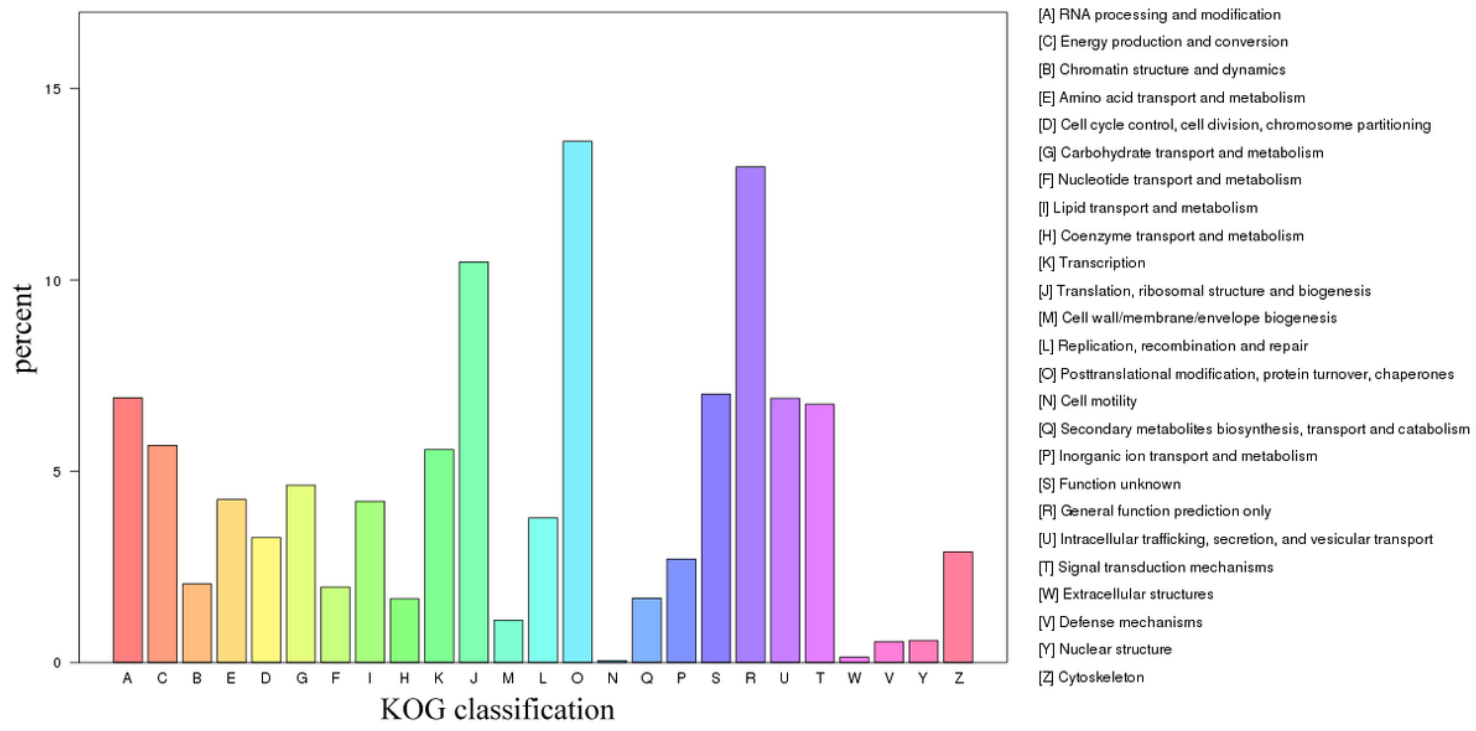

Figure 3

Functional classification of Unigene in KOG

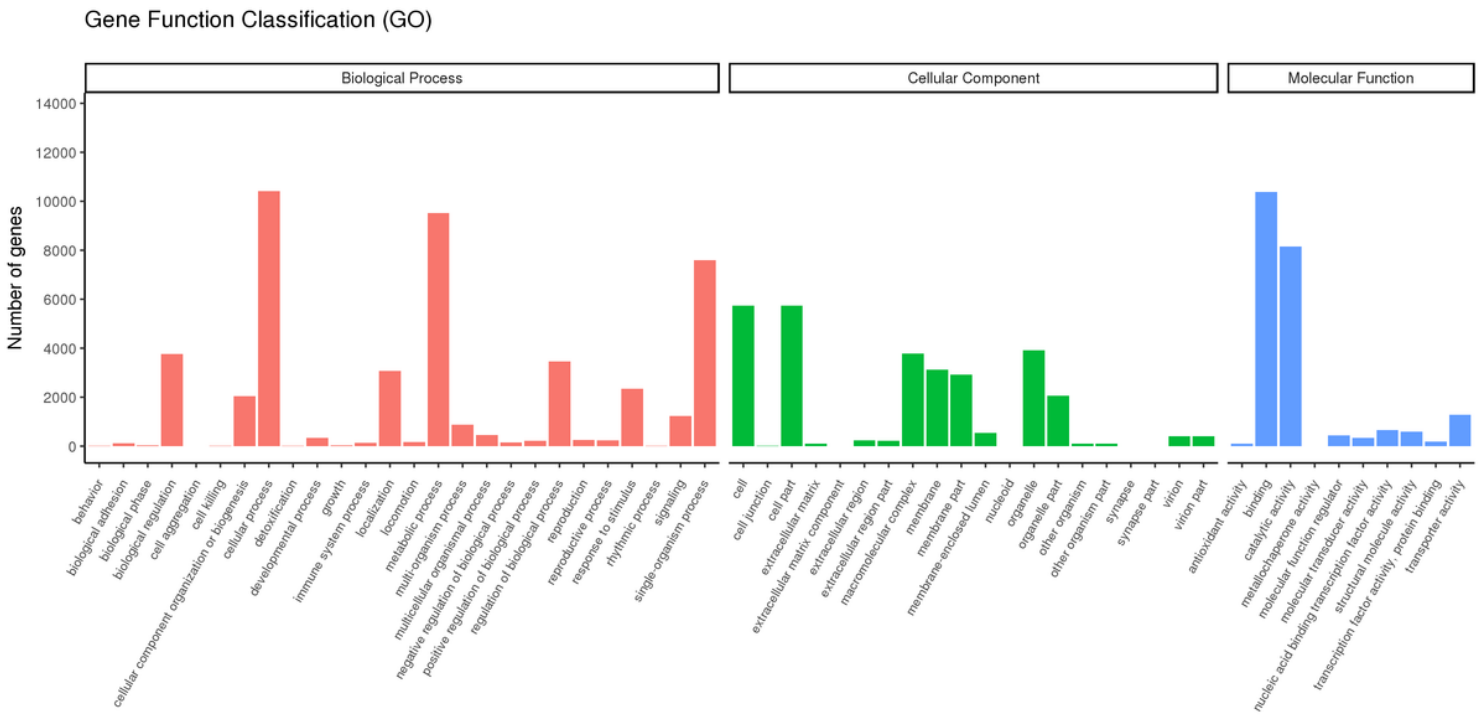

Figure 4

Functional classification of Unigene in GO database

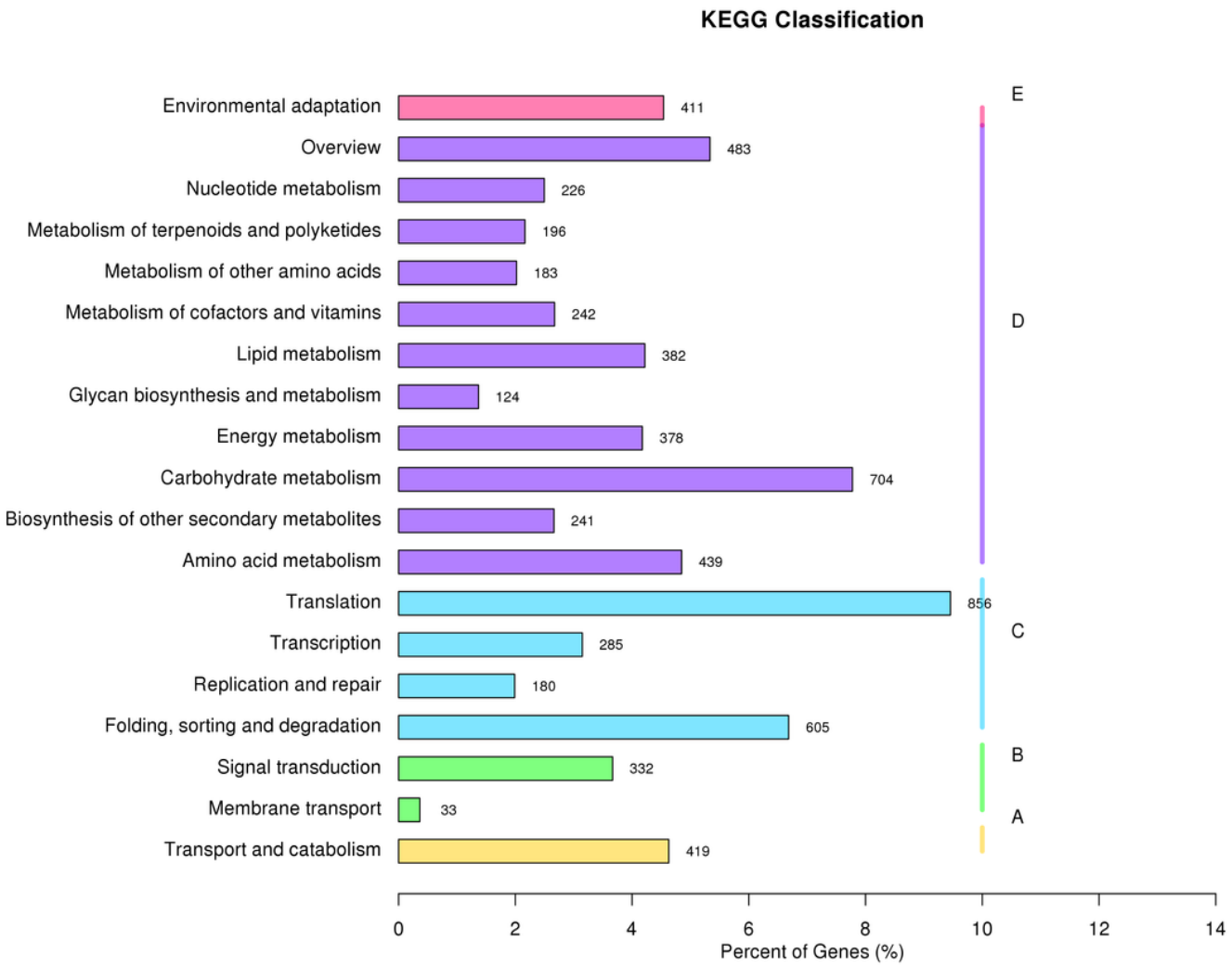

Figure 5

Functional classification of Unigene in KEGG database.

A: Cellular Processes B: Environmental Information Processing C: Genetic Information Processing D: Metabolism E: Organismal Systems

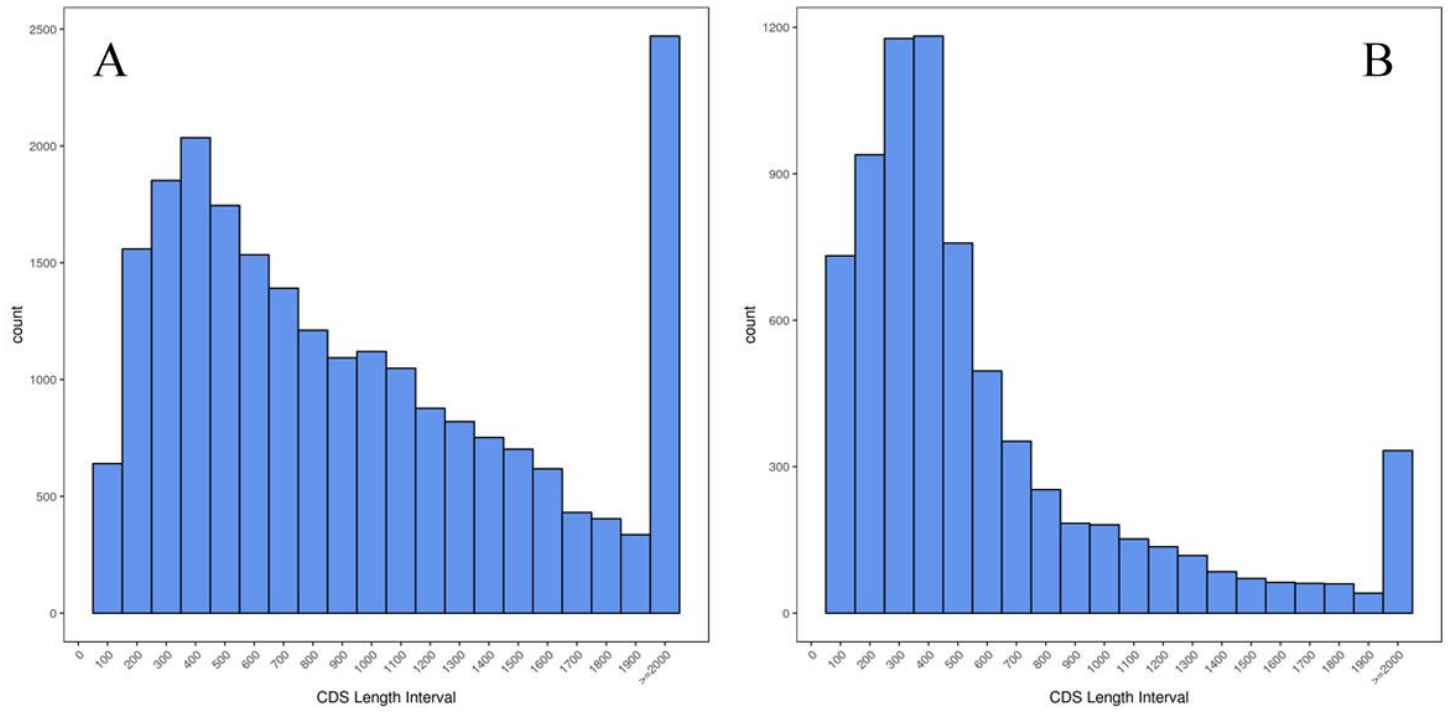

Figure 6

Length distribution of CDS predicted in the Nr and Swiss-Prot (A) or by the ESTS can (B)

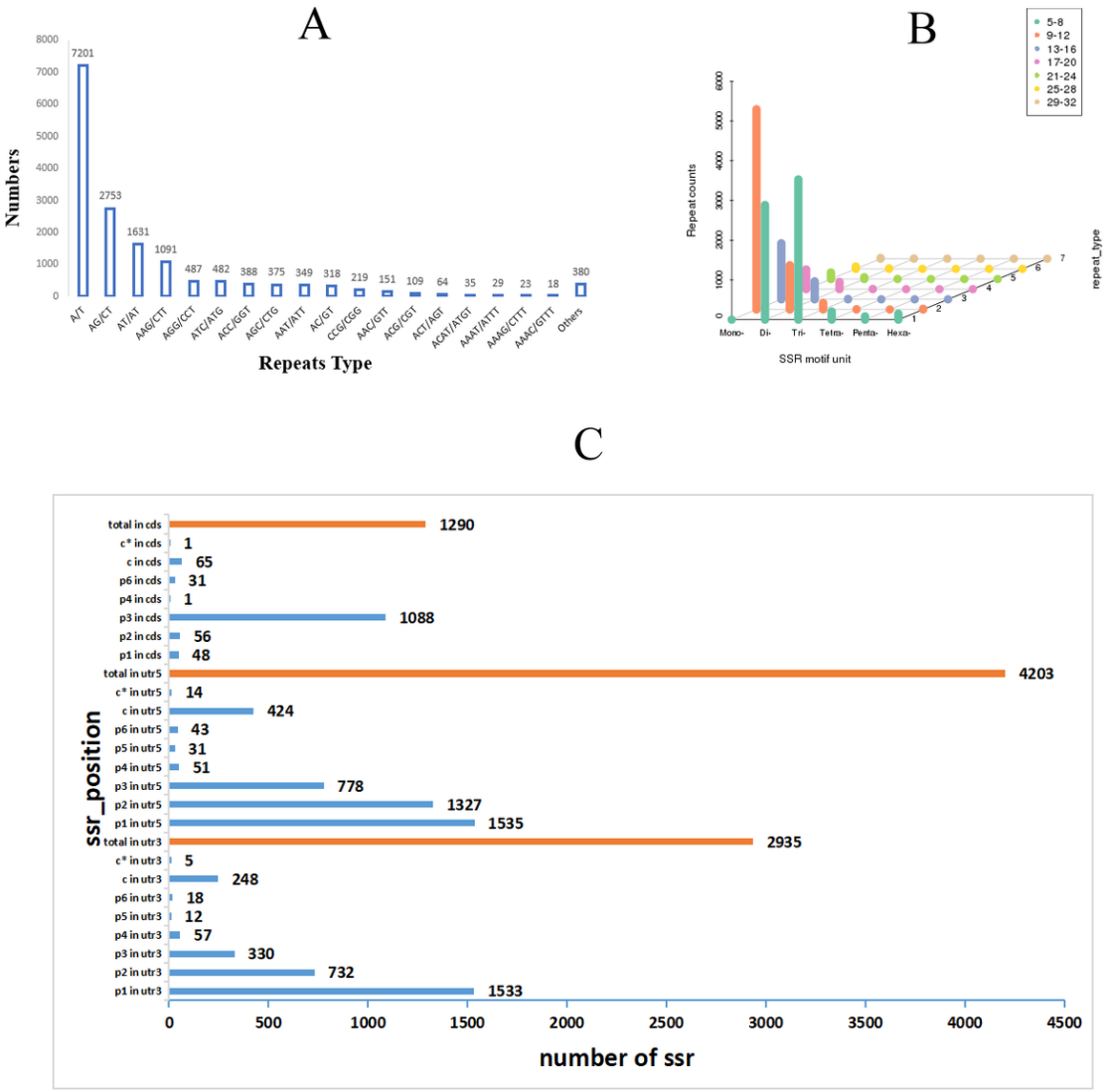

Figure 7

(A): Number distribution of SSR repeat types (B): Distribution of SSR motif in the transcriptome (C): Quantity distribution of SSR locations in *A. chinensis* Bunge. Note: P1-P6: Mono-, Di-, Tri-, Tetra-, Penta-, Hexa- nucleotide; c and c*: Complex and repetition type

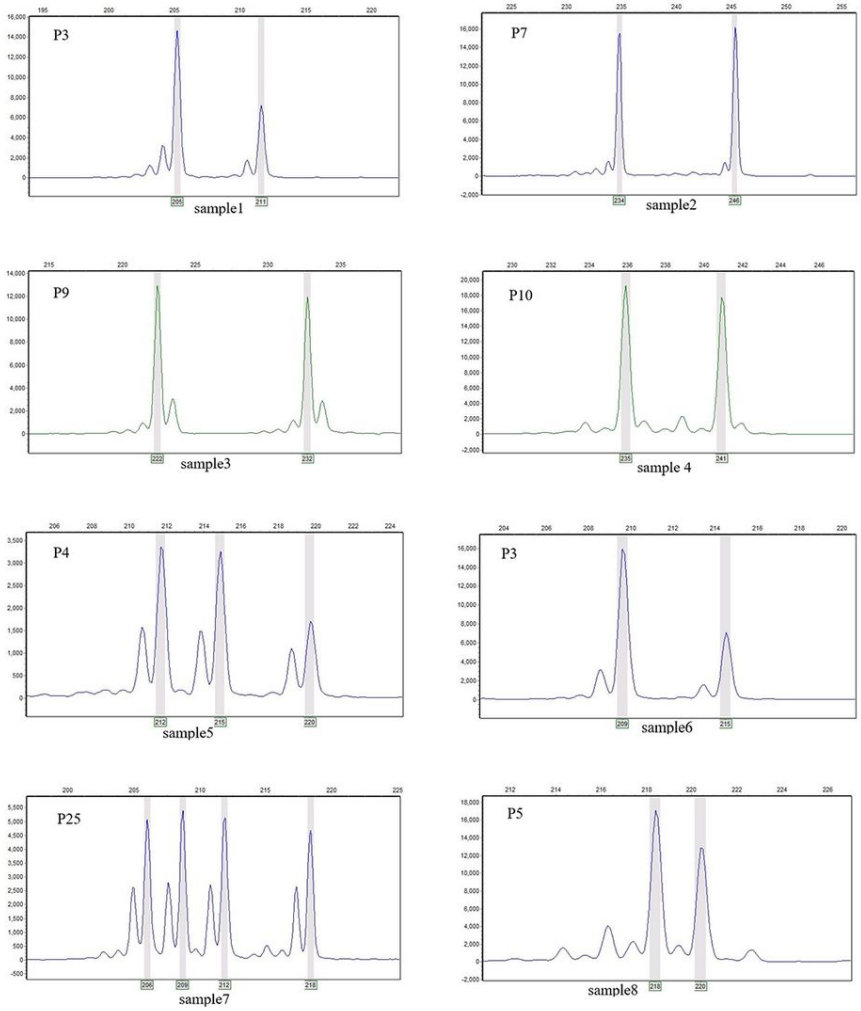

Figure 8

Partial capillary electrophoresis diagram of eight *Aesculus* samples

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfiles.zip