Mol Genet Genomics (2016) 291:849–862
DOI 10.1007/s00438-015-1147-y

ORIGINAL ARTICLE

De novo assembly and characterization of the leaf, bud, and fruit transcriptome from the vulnerable tree *Juglans mandshurica* for the development of 20 new microsatellite markers using Illumina sequencing

Zhuang Hu[1] · Tian Zhang[1] · Xiao-Xiao Gao[1] · Yang Wang[1] · Qiang Zhang[1] · Hui-Juan Zhou[1] · Gui-Fang Zhao[1] · Ma-Li Wang[1] · Keith E. Woeste[2] · Peng Zhao[1]

Received: 22 July 2015 / Accepted: 13 November 2015 / Published online: 27 November 2015

Abstract Manchurian walnut (*Juglans mandshurica* Maxim.) is a vulnerable, temperate deciduous tree valued for its wood and nut, but transcriptomic and genomic data for the species are very limited. Next generation sequencing (NGS) has made it possible to develop molecular markers for this species rapidly and efficiently. Our goal is to use transcriptome information from RNA-Seq to understand development in *J. mandshurica* and develop polymorphic simple sequence repeats (SSRs, microsatellites) to understand the species' population genetics. In this study, more than 47.7 million clean reads were generated using Illumina sequencing technology. De novo assembly yielded 99,869 unigenes with an average length of 747 bp. Based on sequence similarity search with known proteins, a total of 39,708 (42.32 %) genes were identified. Searching against the Kyoto Encyclopedia of Genes and Genomes Pathway database (KEGG) identified 15,903 (16.9 %) unigenes. Further, we identified and characterized 63 new transcriptome-derived microsatellite markers. By testing the markers on 4 to 14 individuals from four populations, we found that 20 were polymorphic and easily amplified. The number of alleles per locus ranged from 2 to 8. The observed and expected heterozygosity per locus ranged from 0.209 to 0.813 and 0.335 to 0.842, respectively. These twenty microsatellite markers will be useful for studies of population genetics, diversity, and genetic structure, and they will undoubtedly benefit future breeding studies of this walnut species. Moreover, the information uncovered in this research will also serve as a useful genetic resource for understanding the transcriptome and development of *J. mandshurica* and other *Juglans* species.

Keywords *Juglans mandshurica* · Microsatellites · Transcriptome · Next generation sequencing · Walnut

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-015-1147-y) contains supplementary material, which is available to authorized users.

✉ Peng Zhao
pengzhao@nwu.edu.cn

[1] College of Life Sciences, Northwest University, Xi'an 710069, Shaanxi, China

[2] Department of Forestry and Natural Resources, USDA Forest Service Hardwood Tree Improvement and Regeneration Center (HTIRC), Purdue University, 715 West State Street, West Lafayette 47907, IN, USA

Introduction

The Manchurian walnut (*Juglans mandshurica*), a member of Juglandaceae, grows in northern and northeastern China, Korea, Japan, and the far eastern section of Russia (Lu 1982; Bai et al. 2010). Manchurian walnut is a multi-use species. It is one of the most desirable and economically valuable hardwood timber tree species in northern and northeastern China, and its leaves, fruits, roots, bark, and seeds have been used in traditional medicine and are of interest for cancer treatment in Asia and Europe (Kim et al. 1998; Xu et al. 2010). Unfortunately, many populations of Manchurian walnut are endangered by over exploitation and deforestation (Ma et al. 2007; Lin et al. 2008). Despite this species' value, existing genomic resources for *J. mandshurica* include only 102 nucleotide sequences and 27 proteins in GenBank (as October 2015). There are a few microsatellites for *J. mandshurica* developed from genomic DNA (Chen et al. 2013) and related species (Woeste et al. 2002), but genetic studies of *J. mandshurica* (Bai et al.

2010) have been limited by the relatively small number of polymorphic markers available, and there are no ESTs (expressed sequence tags) available in GenBank for *J. mandshurica*.

Due to the large effort required to sequence eukaryotic genomes, only a few plant genomes were fully sequenced before 2014 (Wei et al. 2015). Next generation sequencing (NGS) technologies have enhanced our ability to obtain significant and robust sequencing depth and coverage in a rapid and low cost manner (Shendure and Ji 2008; Zhou et al. 2014), permitting the discovery of genomic information in previously uncharacterized species (Grabherr et al. 2011). Similarly, high-throughput RNA sequencing (RNA-Seq) has emerged as a powerful and cost-efficient tool. Compared with traditional methods for SSR marker development, high-throughput sequencing is more cost-efficient, timesaving, and powerful (Jiang et al. 2015). The transcriptome data generated by high-throughput sequencing has been an excellent resource for SSR marker development (Zhang et al. 2012; Fu et al. 2013; Jiang et al. 2013). EST-SSRs have a variety of uses; some sequences contain polymorphisms that are transferable to related taxa. These can facilitate population genetic studies in those species and comparative mapping (Ellis and Burke 2007; Bodénès et al. 2012). They can also be used for genetic analysis (Zhang et al. 2015; Wei et al. 2015;Yan et al. 2015; Jiang et al. 2015), novel gene discovery, molecular marker development, and comparative genomics (Fu et al. 2013; Wang et al. 2015; Jiang et al. 2015). Aside from the development of EST-SSRs, RNA-Seq also allows the potential recovery of rare transcripts, splicing variants, and the determination of gene expression levels in cells or tissues of interest (Kaur et al. 2011; Zhou et al. 2014; Poczai et al. 2013; Lesur et al. 2015).

The goals of this study were to (1) Sequence RNA from *J. mandshurica* using Illumina HiSeq 2000, then assemble de novo, annotate, and characterize a complete transcriptome for the species. (2) Use the resulting transcriptome data to identify a suite of polymorphic EST-SSRs for population genetic analysis. (3) Demonstrate the utility of the derived EST-SSRS by characterizing four populations of *J. mandshurica*.

Materials and methods

Sample collections, DNA extraction, and RNA extraction

For transcriptomic research, fresh leaves, buds, and immature fruits were collected on July 12, 2014 from a single, mature, healthy-appearing *J. mandshurica* tree growing in the Xiaolongmen National Forest Park and immediately frozen in liquid nitrogen prior to storage at −80 °C. Leaves, buds, and (entire) immature fruits were ground to a powder in liquid nitrogen and total RNA was extracted using a Plant RNA Kit (OMEGA Bio-Tek, Norcross, GA, USA). RNA degradation and contamination was monitored on 1 % agarose gels. RNA purity was assayed using the Nano Photometer spectrophotometer (IMPLEN, Westlake Village, CA, USA) and RNA concentration was measured using the Qubit RNA Assay Kit and a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). RNA integrity was assessed using the RNA Nano 6000 Assay Kit and Agilent Bioanalyzer 2100 system (Agilent Technologies, Santa Clara, CA, USA). To verify the polymorphism of EST-SSRs and for subsequent population genetic studies, we extracted DNA from 48 leaf samples collected in 2013 from *J. mandshurica* trees in natural populations in four locations in China (Table 1). The fresh leaves had been previously dried with silica gel. DNA was extracted following the methods described by Doyle and Doyle (1987) and Zhao and Woeste (2011) and stored at –20 °C.

RNA-seq library preparation for transcriptome sequencing

RNA-seq libraries were generated using NEBNext Ultra RNA Library Prep Kit from Illumina (NEB, Beverly, MA, USA) following manufacturer's recommendations, and index codes were added to attribute sequences to each sample (leaf, bud, and fruit). Briefly, mRNA was purified from 3 μg total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under elevated temperature in NEBNext

Table 1 Forty-eight individuals from four locations of *Juglans mandshurica* used in this study

Collection site	Population code	Species	Sample size	Longitude (E)	Latitude (N)	Elevation/m
Xiakoucun, Beijing	XK	*J. mandshurica*	14	115°91′	40°18′	311
Changbaishan, Jilin	CB	*J. mandshurica*	6	127°70′	41°56′	ND
Xinglongxian, Hebei	XL	*J. mandshurica*	4	118°08′	40°50′	ND
Xiaolongmen, Beijing	LM	*J. mandshurica*	24	115°85′	39°97′	1267
Total			48			

ND data was not available

first strand synthesis reaction buffer (5×). First strand cDNA was synthesized using random hexamer primer and M-MuLV reverse transcriptase (RNase H^-). Second strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. In the reaction buffer, dNTPs with dTTP were replaced by dUTP. Remaining overhangs were converted to blunt ends via the exonuclease/polymerase activity of the enzyme. After adenylation of 3′ ends of DNA fragments, an NEBNext Adaptor with hairpin loop structure were ligated on to prepare the cDNA for hybridization. To enrich for ~150–200 bp cDNA fragments, the library was first purified using the AMPure XP system (Beckman Coulter, Beverly, MA, USA). The size-selected, adaptor-ligated cDNA was treated with USER Enzyme (NEB, Beverly, MA, USA) and held at 37 °C for 15 min followed by 5 min at 95 °C. Next, PCR was performed using Phusion High-Fidelity DNA polymerase, universal PCR primers and index (X) primer. Finally, PCR products were purified (AMPure XP system) and library quality was assessed using the Agilent Bioanalyzer 2100 system.

Transcriptome assembly and gene annotation

Paired-end Illumina HiSeq 2000 sequencing was performed by Novogene Bioinformatics Technology Co., Ltd., Beijing, China (www.novogene.cn). De novo transcriptomes assembly was accomplished using Trinity (Grabherr et al. 2011) using the default setting. Unigenes of the transcriptome were annotated based on data from the NCBI non-redundant protein sequences (Nr) database, NCBI non-redundant nucleotide sequences (Nt) database, Clusters of Orthologous Group of proteins (KOG/COG) database, KEGG ortholog (KO) database, a manually annotated and reviewed protein sequence (Swiss-Prot protein) database, Gene Ontology (GO) database, and protein family (Pfam) database. The COGs protein database phylogenetically classifies the complete complement of proteins encoded in a genome. Each COG is a group of three or more proteins that are inferred to be orthologs or direct evolutionary counterparts. To further analyze the transcriptome of *J. mandshurica*, all of the unigenes were submitted to the KEGG pathway database. The KEGG pathway database is a knowledge base for the systematic analysis of gene functions (Long et al. 2014). KOG, Nr, Nt, and Swiss-Prot database used NCBI Blast 2.2.28+ (Boeckmann et al. 2003). All BLAST searches were performed with an e-value of 1e-5. Pfam protein domain prediction using software HMMER 3 package hmmscan (Farrar 2007). GO annotations using Blast2GO v2.5 were performed using the cutoff *e*-value was 1e-6 (Conesa et al. 2005; Götz et al. 2008).

Discovery of EST-SSRs, primer design

Microsatellites within the assembled unigenes were identified using MISA (http://pgrc.ipk-gatersleben.de/misa/misa.html). Sequences with ≥5 uninterrupted motifs (from 2 to 6 repeat motifs) were randomly selected for primer design by Primer3 (http://primer3.sourceforge.net/releases.php). Primer design parameters were set as follows: length range from 18 to 23 nucleotides with 21 bp as optimum, PCR product size range from 100 to 400 bp, optimum annealing temperature from 55 °C, and GC content 40–60 % with 50 % as optimum.

Amplification conditions and validation of primers

DNA was resuspended in 50 μL of water and dilutions were performed to obtain a final concentration of 10 ng/μL and stored at −20 °C until use. PCR amplification was carried out on an ABI Veriti96 with the following method: 3 min at 94 °C followed by 35 cycles of 15 s at 93 °C, 1 min at annealing temperature (Tm) (Table 2), 30 s at 72 °C, and extension of 10 min at 72 °C. PCR reactions contained 5 μL 2 × PCR Master Mix including 0.1 U Taq polymerase/μL; 500 μM each dNTP; 20 mM Tris–HCl (pH8.3); 100 mM KCl; 3.0 mM $MgCl_2$ (Tiangen, Beijing, China), 0.2 μM each primer (Shagon Biotech, Shanghai, China), 0.1 mg/mL bovine serum albumin, (Sigma, USA, 1 ng/μL DNA, and 2.6 μL ddH_2O to produce a total reaction volume of 10 μL. For primer design, 63 unigenes were randomly selected from 16,699 sequences containing microsatellites that were not single nucleotide repeats, contained greater than five motif repeats, and that would amplify a PCR product size of 150–280 bp from genomic DNA. Genomic DNA from 48 samples of *J. mandshurica* were used for PCR amplification and analysis of polymorphism. All 48 genotypes were tested at all 63 loci for polymorphisms. PCR products were resolved on 8 % polyacrylamide gels and visualized by silver staining. Fragment sizes of each locus were estimated using Quantity One Software (Bio-Rad Laboratories, Drive Hercules, CA, USA) and compared to a 50 bp ladders size standard.

Microsatellites data analysis

Genetic diversity per locus and population were evaluated based on the following descriptive summary statistics: number of alleles (N_A), observed (H_O), and expected (H_E) heterozygosity using the program GenAlEx6.5 (Peakall and Smouse 2012). GENEPOP version 4.2 (Rousset 2008) was used to test the Hardy–Weinberg equilibrium (HWE) and linkage disequilibrium (LD) for all loci. Their significance was assessed with 1500 permutations. CERVUS version 3.0 (Kalinowski et al. 2007) was used to calculate the polymorphic information content (PIC) and null alleles were found

using MICRO-CHECKER 2.2.3 (Van Oosterhout et al. 2004). An analysis to detect genetic structure was performed by using the software STRUCTURE (version 2.3.4) with a burn-in of 100,000 Markov Chain Monte Carlo iterations, a duration of 1000,000 iterations, ten replicates per run for *K* from 2 to 4 clusters, and the admixture model (Pritchard et al. 2000; Evanno et al. 2005). The program STRUCTURE HARVESTER was used to calculate the optimal value of *K* using the deltaK criterion (Earl and vonHoldt 2012). The inferred clusters were drawn as colored box-plots using program DISTRUCT (Rosenberg 2004).

Results

Illumina sequencing and De novo assembly

To obtain an overview of the *J. mandshurica* transcriptome, RNA was isolated from three plant organs: fresh leaves, vegetative buds, and immature fruits. RNA-Seq was performed using the Illumina HiSeq 2000 platform (100 bp paired end). In total, 5.04 GB sequence quality reads were generated, resulting in a total of 4.77 GB of clean data for assembly (Table 2). The maximum read counts were 268,844, and RPKM (Reads Per Kilobase of exon model per Million mapped reads) was 10,814 (Fig. S1). The mean Q20 percentage (sequencing error rate <0.05 %) was 97.22, mean Q30 percentage was 91.17, and GC percentage was 45.27 (Table 2). Using the software Trinity, de novo assemblies generated 205,305 transcripts including 99,869 unigenes. The length of the transcripts varied from 201 to 13, 402 bp, with an average of 1140 bp, and N50 value and N90 value of 1863 and 487 bp, respectively. The length of the unigenes varied from 201 to 13,402 bp, with an average of 747 bp, and N50 value and N90 value of 1309 and 288 bp, respectively (Fig. 1). The transcripts with the length over 500 bp accounted for about 62.03 % while the number of the length between 200 and 500 bp was 77, 964. The unigenes with length of over 500 bp accounted for about 41 % compared to lengths between 200 and 500 bp (Fig. 2; Fig. S2; Fig. S3).

Microsatellites (SSR) enrichment

In total, 22,019 sequences contained SSRs, that was 22 % of all unigenes, and the distribution density was 227 per MB. In these loci, mononucleotide repeats were the most frequent (7208; 43.2 %), followed by dinucleotide repeats (6180; 37.0 %) and trinucleotide repeats (2288; 13.7 %) (Fig. 3). SSRs with ten tandem repeats 5023 (22.81 %) were the most common, followed by six tandem repeats 2896 (13.15 %), 11 tandem repeats (9.7 %), seven tandem repeats (9.08 %), five tandem repeats (8.07 %), nine tandem

Table 2 Summary of sequencing output statistics for *Juglans mandshurica* compared to other recent studies

Species	Raw reads	Clean reads	Number of transcripts	Mean length of Transcripts (bp)	Number of unigenes	Mean length of unigenes (bp)	Error (%)	Q20 (%)	Q30 (%)	GC (%)	References
J. mandshurica	50,409,752	47,689,934	205, 305	1140	99, 869	747	0.03	97.22	91.17	45.27	Present study
Lilium	36,926,680	–	75,474	–	39,636	837	0.02	97.59	–	50.33	Zhang et al. (2015)
Ipomoea nil	–	53,644,440	220,117	920	–	–	0.01	99.0	–	45.00	Wei et al. (2015)
Rosa roxburghii	–	53,535,304	263,892	208	106,590	343	0.04	95.75	–	–	Yan et al. (2015)
Oenanthe javanica	–	–	58,072	–	40,208	539.2	0.06	94.26	–	–	Jiang et al. (2015)

– data not reported

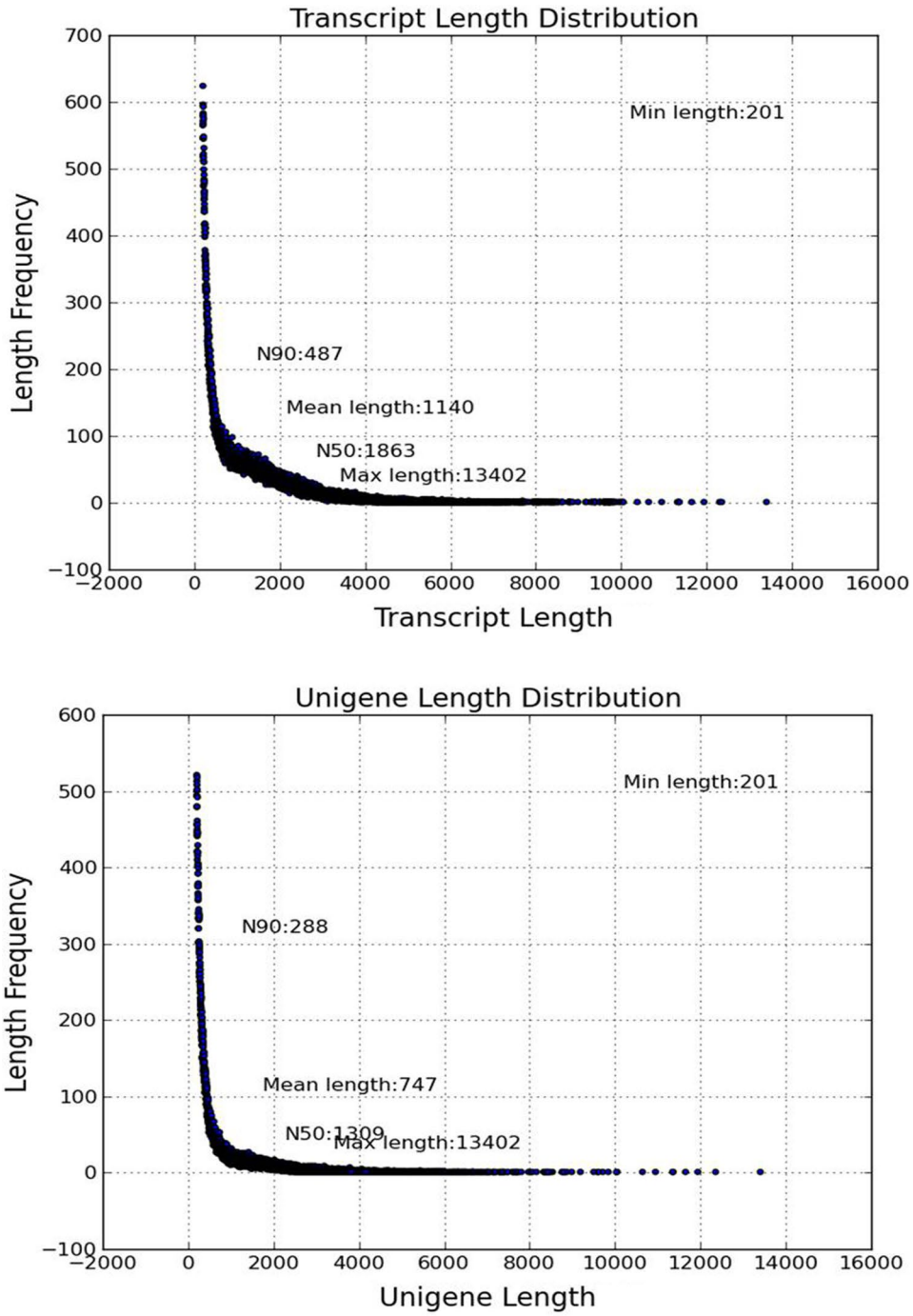

Fig. 1 The length distributions of the transcripts and unigenes of *Juglans mandshurica*

repeats (7.44 %), and >11 tandem repeats (22.78 %). The dominant repeat motif in EST-SSRs was A/T (46.64 %), followed by AG/CT (30.10 %), AT/AT (4.51 %), AAG/CTT (4.26 %), AC/GT (2.43 %), and AGG/CTT (1.83 %) (Table S1).

Gene annotation of *J. mandshurica* transcriptomes

In this study, 39,868 of 99,869 unigene sequences (39. 92 %) were annotated to GO classes (Fig. 4). All unigenes were annotated to three major GO classes: the largest class was biological processes (105,684; 46.13 %), followed by a cellular component class (73,129; 31.92 %), and molecular function class (50,273; 21.95 %). Cellular processes (24,200; 22.90 %) and metabolic processes (22,699; 21.48 %) were the largest and second largest of 22 categories of biological processes, while the least abundant were rhythmic processes (45; 0.04 %) and cell death (34; 0.03 %), respectively. Within the cellular component class, cell (14,911; 20.39 %), cell part (14,822; 20. 35 %), and

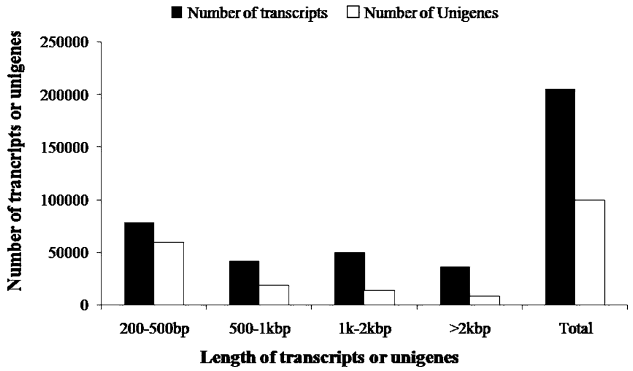

Fig. 2 The assembly length distribution of *Juglans mandshurica*

organelle (10,322; 14.40 %) were the most abundant among the 19 categories (five categories not shown in Fig. 4), while the least abundant were synapse part (8; 0.01 %) and synapse (8; 0.01 %). Within the molecular function class, binding (23,264; 46. 28 %) and catalytic activity (19,752; 39.29 %) were the most abundant among the 13 categories (two categories not shown in Fig. 4), while the least abundant were receptor regulator activity (13; 0.03 %) and metallochaperone activity (9; 0.02 %) (Fig. 4).

Functional classification by orthologous groups (COG)

All unigenes were aligned to the COG database to predict and classify their possible functions. Out of 20,087 Nr hits, 17,806 sequences were assigned to COG classifications (Fig. 5). Among the 25 COG categories, the cluster for general function prediction only (3135; 17.61 %) represented the largest group, followed by post-translational modification, protein turnover, chaperones (2390; 13.42 %), signal transduction mechanisms (1676; 9.41 %), translation, ribosomal structure and biogenesis (1128; 6.33 %), intracellular trafficking, secretion, and vesicular transport (1200; 6.29 %), transcription(1137; 6.31 %), translation (1119; 6.28 %), and RNA processing and modification (1060; 5.95 %), whereas only a few unigenes were assigned to cell motility and unnamed protein (Fig. 5).

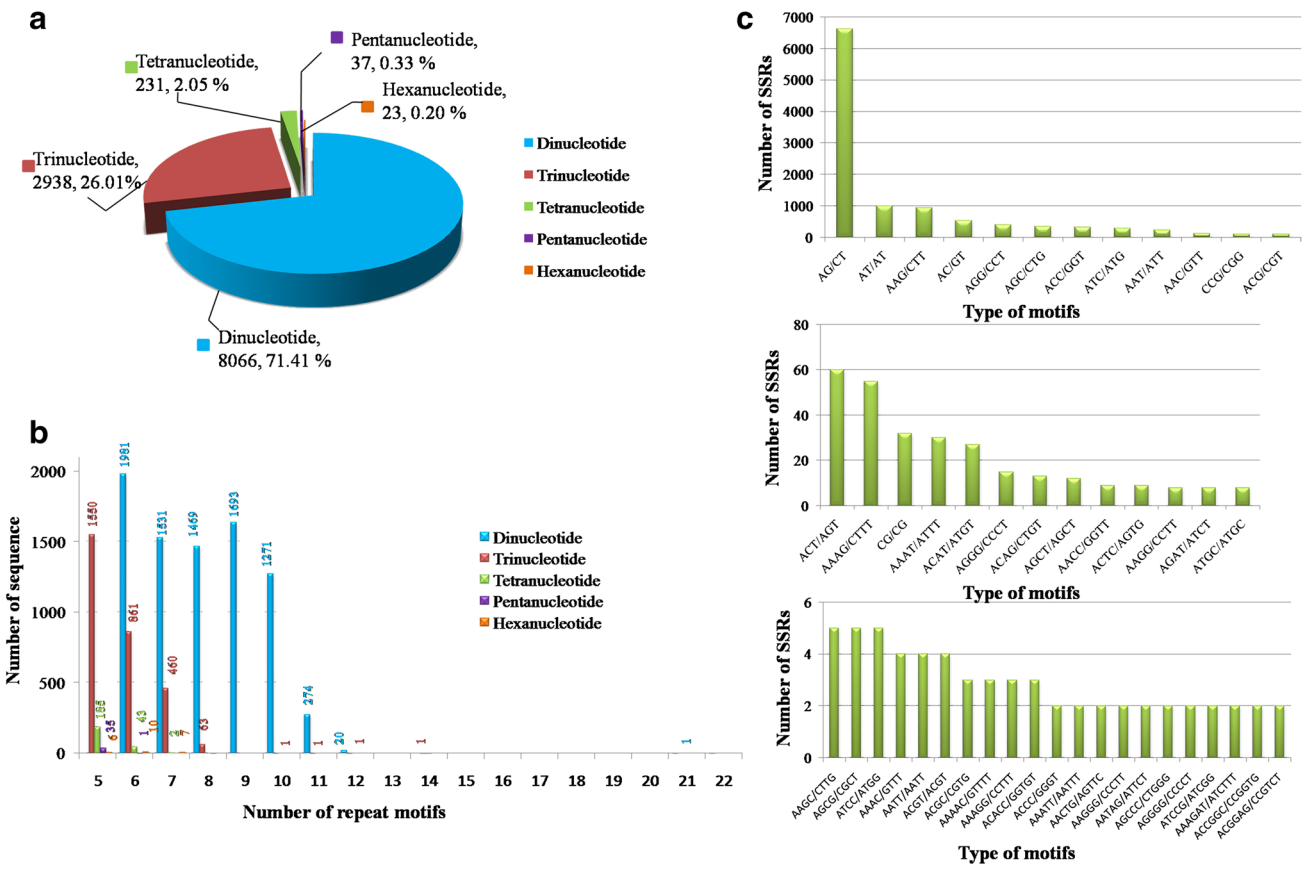

Fig. 3 Characterization of simple sequence repeats (SSRs) in the Manchurian walnut (*Juglans mandshurica*) transcriptome. **a** Distribution of different SSR repeat motif types, **b** number of different repeat motif, and **c** frequency distribution of major SSRs based on main motif type

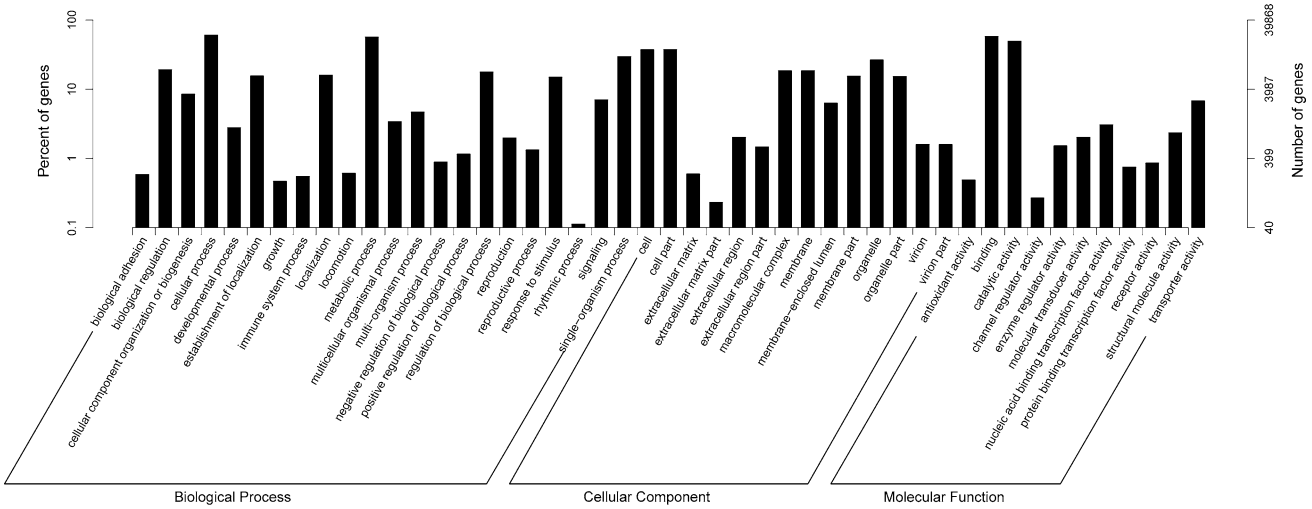

Fig. 4 Gene ontology classification of assembled unigenes. The results are summarized in three main categories: biological process, cellular component, and molecular function. A total of 99,869 unigenes with BLAST matches to known proteins were assigned to gene ontology

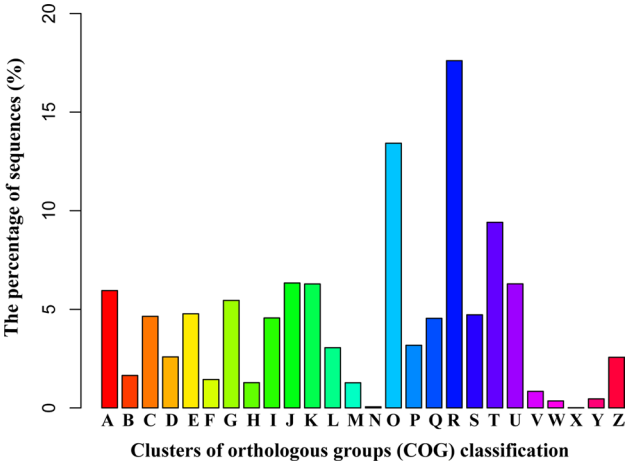

Fig. 5 Histogram presentation of clusters of orthologous groups (COG) classification. All unigenes were aligned to COG database to predict and classify possible functions. Out of 27,435 Nr hits, 18,001 sequences were assigned to 26 COG classifications. RNA processing and modification (*A*), chromatin structure and dynamics (*B*),energy production and conversion (*C*), cell cycle control, cell division, chromosome partitioning (*D*), amino acid transport and metabolism (*E*), nucleotide transport and metabolism (*F*), carbohydrate transport and metabolism (*G*), coenzyme transport and metabolism (*H*), lipid transport and metabolism (*I*), transition, ribosomal structure and biogenesis (*J*), transcription (*K*), replication, recombination and repair (*L*), cell wall/membrane/envelope biogenesis (*M*), cell motility (*N*), posttranslational modification, protein turnover, chaperones (*O*), inorganic ion transport and metabolism (*P*), secondary metabolites biosynthesis, transport and catabolism (*Q*), general function prediction only (*R*), function unknown (*S*), signal transduction mechanisms (*T*), intracellular trafficking, secretion, and vesicular transport (*U*), defense mechanisms (*V*), extracellular structures (*W*), unnamed protein (*X*), nuclear structure (*Y*), and cytoskeleton (*Z*)

Functional classification by the KEGG pathway

Out of the 15,903 unigenes, 6013 (40.3 %) with significant matches were assigned to five main categories divided into 32 KEGG pathways (Fig. 6). Among the 32 pathways, most represented were carbohydrate metabolism (134,1; 8.99 %), and translation (133,2; 8.93 %), followed by signal transduction (1177; 7.89 %), folding, sorting, and degradation (116,9; 7.84 %), and overview processing (994; 6.77 %).

SSR primer screening and verification

Among the 63 microsatellite loci evaluated, only 20 (30.8 %) successfully amplified and also were polymorphic among the four *J. mandshurica* populations tested (Table 2; Fig. 7; Table S2). The polymorphic EST-SSRs ranged from dinucleotide repeats to hexanucleotide repeats, although trinucleotide repeats were not represented, which was surprising because they were relatively common overall (Fig. 3). The number of alleles (N_A) for each locus ranged from 2 to 8 and the mean number of alleles per locus was 4.95. We observed no obvious association between repeat unit size (di- versus penta-nucleotide repeats, for example) and polymorphism or allele number. The observed heterozygosity (H_O) and expected heterozygosity (H_E) varied from 0.309 to 0.813 and from 0.335 to 0.842, with an average of 0.516 and 0.639, respectively (Table 3). The 20 polymorphic SSR-containing unigenes were assigned putative biological function based on BLAST search in the NCBI

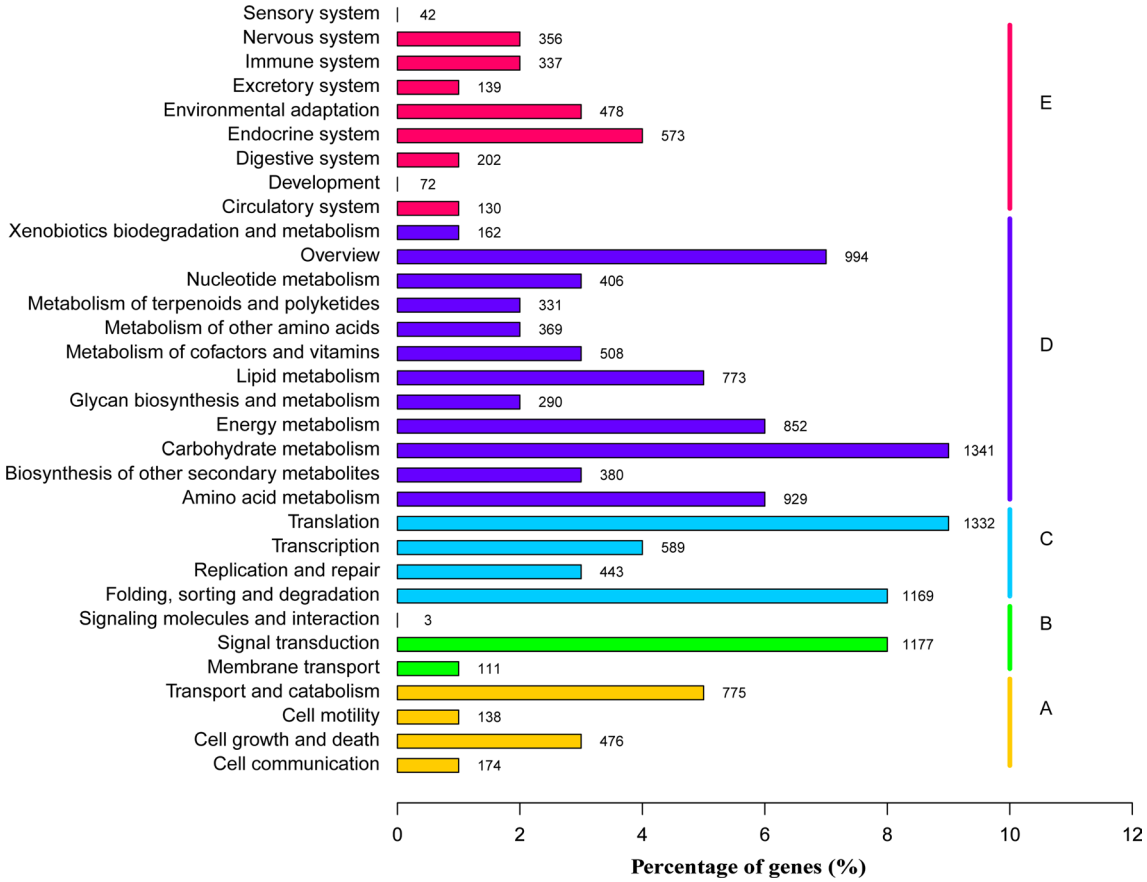

Fig. 6 Pathway assignment based on the Kyoto Encyclopedia of Genes and Genomes (KEGG). *A* Classification based on cellular process categories, *B* classification based on environmental information processing categories, *C* classification based on genetic information processing categories, *D* classification based on metabolism categories, and *E* classification based on organismal systems categories

protein databases (Table 3; Table S3). Eight of the unigenes had no identifiable function, as they matched to "hypothetical protein" or "PREDICTED" results, and five unigenes (JM68820, JM77590, JM9158, JM5446, JM1672) had no significant match to any protein in the database (Table 3).

Four loci out of the 20 polymorphic EST-SSRs, JM59276, JM75325, JM2039, and JM5272, showed significant departures from Hardy–Weinberg equilibrium (HWE) across all samples (Table 3). The Bayesian clustering of 48 individuals was from 4 natural populations where the deltaK spectrum identified $K = 3$ as the most likely number of populations for the sampled genotypes (Fig. 8). The Bayesian clustering approach revealed that samples from XK belonged to a distinct population (red in Fig. 8), as did samples from CB (green in Fig. 8). Samples from sources XL and LM were either from population II (similar to CB samples) or population III (blue in Fig. 8), and samples from LM were admixed (containing elements of population II and III) or assigned to populations II or III exclusively (Fig. 8).

Discussion

The current demographic decline of natural populations of Manchurian walnut is attributed to human activities, including deforestation and global climate change (Ma et al. 2007; Lin et al. 2008; Zhu et al. 2011). As a consequence, there is a need for genetic information to inform management and conservation decisions. Investigation of the genetic diversity of this walnut species has been limited to population genetics and characterization of phenology and gene flow in natural populations (Bai et al. 2007, 2010). Prior to our work, there was little DNA sequence data for *J. mandshurica* in public databases, and our study is the first to produce and characterize transcriptome data of any kind for the species. In the absence of an assembled reference genome, the unigenes described here are the most complete genomic resources for Manchurian walnut. We sequenced expressed sequences (RNA) from three representative organ types (leaves, vegetative buds, and fruit) since the aim of this study was to gain as broad an insight

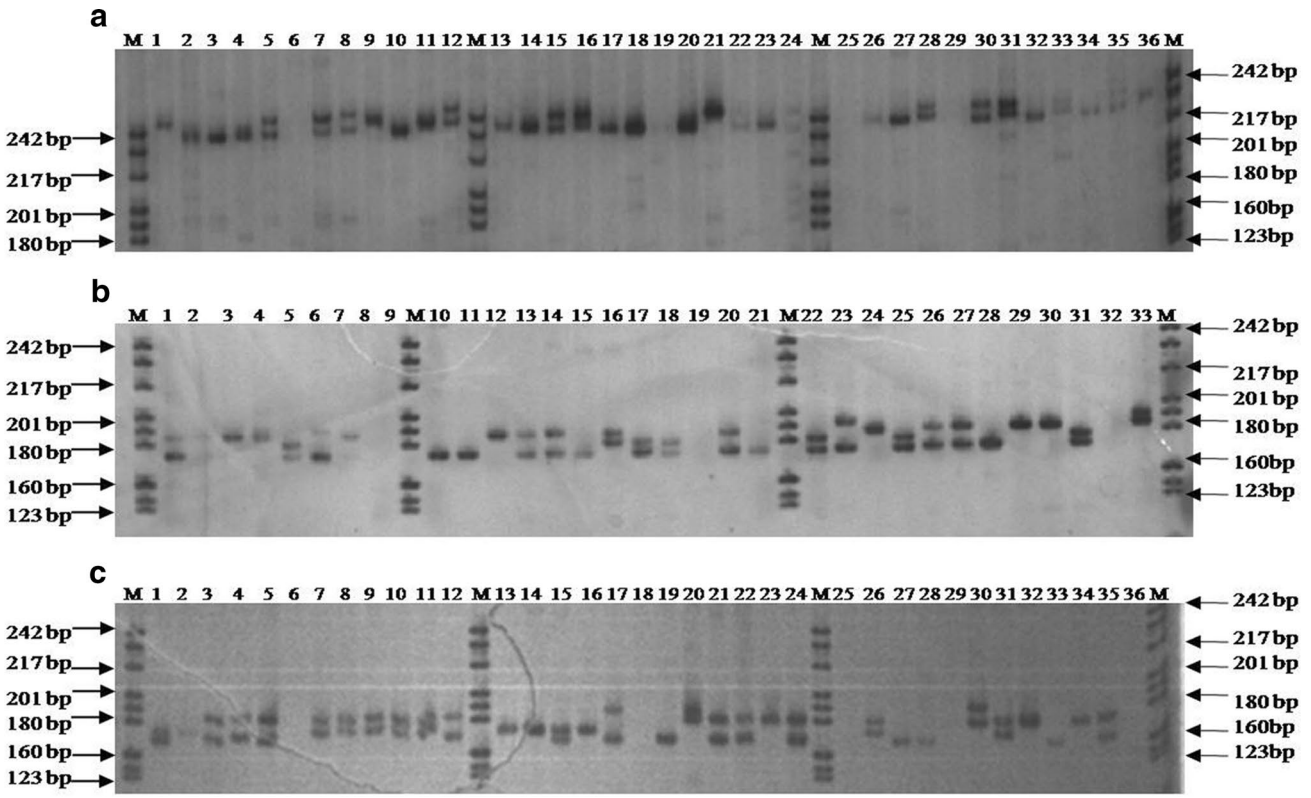

Fig. 7 PCR products and polymorphic patterns of three EST-SSR markers across 48 individuals of *J. mandshurica*. *M* indicates size ladders. "**a**", "**b**", "**c**" represent panels of PCR products for individuals genotyped using primers for the loci JM8820, JM8331, and JM3882, respectively. Lanes 1–36, refer to the different individuals of *J. mandshurica* from four populations (see Table 1)

into the transcriptome of Manchurian walnut as possible. In our case, pooling RNA samples was a cost-effective strategy based on statistical and practical considerations (Sloan et al. 2012; Xu et al. 2012). We examined a total of 5.04 GB of raw reads that corresponded to 4.77 GB of nucleotide sequence. The quality of the resulting transcriptome was high (Table 2). Q20 percentage for our transcriptome data was 97.22, which compares favorably to other recent studies (Wei et al. 2015; Yan et al.2015; Jiang et al. 2015), and the average transcript length of 747 bp was comparable to those observed in other species (Zhang et al. 2015; Jiang et al. 2015; Adal et al. 2015).

Microsatellites are widely distributed in eukaryotic genomes (Li et al. 2002, 2004), although the functions of SSRs in transcribed regions in plants are not well known (He et al. 2015). In this study, we found that Illumina paired-end transcriptome sequencing was an efficient, low-cost, reliable, and rapid tool for transcriptome analysis and EST-SSR marker development in a non-model organism. Dinucleotide repeats were the most frequent SSR motif type (71.41 %) aside from single nucleotide repeats. This finding was consistent with results reported for common walnut (*Juglans regia*) (Zhang et al. 2010), and numerous other species (Kumpatla and Mukhopadhyay 2005; Du et al. 2012), but it is not observed in all species (Adal et al. 2015; Zhang et al. 2015; Wei et al. 2015). Trinucleotide repeats were the most abundant EST-SSR marker for cabbage, red clover, and other plants (Izzah et al. 2014; Yate et al. 2014). Among the dinucleotide repeats (DNRs), AG/CT repeats were the most common in Manchurian walnut, accounting for 58.1 % of the total DNRs. The results were similar to findings reported for *J. regia* (Zhang et al. 2010) and *Ipomoea nil* = 50.32 % (Wei et al. 2015) but not *Lilium* 'Sorbonne' = 23.86 %, (Zhang et al. 2015) and *Vaccinium macrocarpon* = 25.8 % (Schlautman et al. 2015).

From a random sample of 63 unigenes containing microsatellites, we found 20 (30.8 %) that could be easily and consistently amplified and were polymorphic (Table 3). There were at least three reasons the other 43 SSRs did not amplify or appeared monomorphic. First, some of the designed primers may have failed to amplify PCR products because of the presence of introns in the unigenes. Second, SSRs that had no detectable polymorphism may be monomorphic within Manchurian walnut or they may be polymorphic among populations we did not sample. Third, there may have been errors in base calls or assembly that resulted in incorrect flanking sequences from which primers were designed. The

Table 3 Characterization of 20 transcriptome-derived microsatellites of *Juglans mandshurica* Maxim

Locus name	Putative function of Unigene	Repeat motif(s)	Primer sequence (5′–3′)	N_A	Size range (bp)	Tm (°C)	PIC	H_O	H_E	HW
JM68820	None (query length, 1700 bp)	(ACAT)14	F: TCCTTCTGTGTGAGTGCGTG R: GGTCAGGTGAGTGGAGCAAA	3	240–248	55	0.504	0.386	0.577	NS
JM61666	COL domain class transcription factor isoform 1 [*Theobroma cacao*] (query length, 1700 bp; query cover, 50 %; *E* value, 3e-135; Ident, 83 %)	(GA)11	F: AACTGTTGCCGGAGCTTTCT R: TGGGATAACACCACATGCAGT	8	266–280	55	0.813	0.813	0.842	NS
JM78331	Thymidylate synthase 1 isoform 6 [*Theobroma cacao*] (query length, 2567 bp; query cover, 54 %; *E* value, 1e-160; ident, 82 %)	(AACGGC)5	F: GCAGTGCGCTCCTTTTTCAA R: TTCTCGGGTTGAAGCCACAA	4	173–191	55	0.612	0.545	0.681	NS
JM77590	None (Query length, 1679 bp)	(ATGTT)5	F:TGCATGTAAGGGAGCATATGT R: TGCATGAGAAGGTTGTTGGACT	5	278–298	58	0.424	0.646	0.491	NS
JM59276	Hypothetical protein PRUPE_ppa010184 mg [*Prunus-persica*] (Query length, 1317 bp; Query cover, 58 %; E value, 1e-107; ident, 69 %)	(AT)11	F: GAAGCATGCCAACCAAAGCA R: CAGAGGCATTACAGGCAGCT	4	102–108	55	0.681	0.333	0.739	***
JM76238	Glycosyl transferase family 8 family protein [*Populus-trichocarpa]* (query length, 1121 bp; query cover, 73 %; E value, 1e-172; ident, 91 %)	(AATAG)5	F:TCTTTCTTATTTCCAGAGGAAGGA R:TGCTCGCCGGTAAATTCACT	3	242–252	55	0.349	0.391	0.416	NS
JM75325	PREDICTED: F-box protein SKIP14-like isoform X1 [*Fragariavesca* subsp. vesca] (query length, 2403 bp; query cover, 53 %; *E* value, 7e-144; Ident, 58 %)	(GCGA)5	F: CTCAGGAACCCTACGCCAAG R: TTTCTGATTCTCCACCCGCC	4	26–238	55	0.57	0.341	0.643	***
JM7515	Hypothetical protein JCGZ_07992 [*Jatropha curcas*] (query length, 3686 bp; query cover, 27 %; E value, 5e-135; Ident, 82 %)	(TA)9(CA)11	F: ACTGCTTGCGAGATTGCTTG R: CAAATAGGGCAGCCTCGTCT	4	182–188	55	0.544	0.558	0.627	NS
JM7090	F-box protein PP2-A15 [*Glycinesoja*] (Query length, 3220 bp; Query cover, 13 %; *E* value, 2e-79; Ident, 82 %)	(ATGTT)5	F: TGCATGTAAGGGAGCATATGT R: TGCATGAGAAGGTTGTTGGACT	4	205–217	55	0.298	0.401	0.335	NS
JM8463	Haloacid dehalogenase-like hydrolase domain-containing protein 3 [*Morus notabilis*] (query length, 1317 bp; query cover, 43 %; *E* value, 8e-77; ident, 87 %)	(AG)7aac(A)11	F: GTACCTCCCGCATCCAACAA R: CATTCACGATGCAGACCCTT	6	186–196	55	0.556	0.429	0.606	NS
JM9158	None (query length, 257 bp)	(GT)10	F: TCAGACAAAGTTGAGGGCTGA R:TGTGAAAAGTCAGGCAGCCA	8	213–227	53	0.801	0.696	0.832	NS
JM5446	None (query length, 385 bp)	(CTAG)5	F: ATGCATGCAGCTCCTACCTC R: GGACGTGTCCTGGGTTTTCA	5	218–234	53	0.713	0.622	0.76	NS
JM2039	PREDICTED: vacuolar protein sorting-associated protein 35B [*Prunus mume*] (query length, 2809 bp; query cover, 77 %; *E* value, 0.0; ident, 88 %)	(CA)11	F: GAAGGACCTGGATGGAACCG R: GACACCTCCCCATCAAGTGT	4	234–248	55	0.444	0.217	0.492	***
JM5969	Earlier flowering 1 [*Dimocarpus longan*] (query length, 2713 bp; query cover, 46 %; E value, 0.0; Ident, 83 %)	(AG)10	F: ACAATAGTCTCTGCACCGCC R: AGCTTGTACTTACCGCCGAC	6	218–230	55	0.709	0.452	0.758	NS
JM1672	None (query length, 319 bp)	(TC)7cttat(A)11	F: TGCCAGGGGAGCAAGAAAAA R: CTCCCATTGCGAGTCTCCAT	6	205–215	55	0.719	0.409	0.758	NS

Table 3 continued

Locus name	Putative function of Unigene	Repeat motif(s)	Primer sequence (5′–3′)	N_A	Size range (bp)	Tm (°C)	PIC	H_O	H_E	HW
JM3882	Hypothetical protein PRUPE_ppa023884 mg [*Prunus persica*] (Query length, 562 bp; query cover, 50 %; *E* value, 1e-32; ident, 79 %)	(TTTC)5	F: CCTGTGGGCGTAACTTTCCA R: AATCCCCCAATCCCCCATTTT	6	169–185	50	0.684	0.595	0.739	NS
JM9040	Ubiquitin carboxyl-terminal hydrolase isoform 1 [*Theobroma cacao*] (query length, 4910 bp; query cover, 66 %; *E* value, 0.0; ident, 74 %)	(CTCC)5	F: CGGATGGCTTATGCGGGTAT R: TCCTGGCTGAGAGAGGAGAC	5	314–330	50	0.585	0.389	0.659	NS
JM3976	PREDICTED: uncharacterized protein LOC103341189 [*Prunusmume*] (query length, 1103 bp; query cover, 32 %; *E* value, 2e-31; Ident, 63 %)	(TGTC)5	F: GATCAATCGCTCCTACCCCG R: TCAATGAAACCCCAACACCCA	4	215–227	48	0.585	0.439	0.648	NS
JM7109	PREDICTED: uncharacterized protein LOC103341189 [*Prunus mume*] (query length, 1107 bp; query cover, 34 %; *E* value, 8e-32; ident, 63 %)	(ACAG)5	F: TCAATGAAACCCCAACACCC R: GATCAATCGCTCCTACCCCG	4	217–229	50	0.396	0.22	0.437	NS
JM5272	PREDICTED: vacuolar cation/proton exchanger 5 [*Prunusmume*] (query length, 2957 bp; query cover, 45 %; *E* value, 0.0; ident, 82 %)	(CT)10	F: GAGCGAGGGGAGTTTGGAACT R: TGACGGGTTTTTGCGAGTTG	6	168–186	50	0.691	0.209	0.743	***

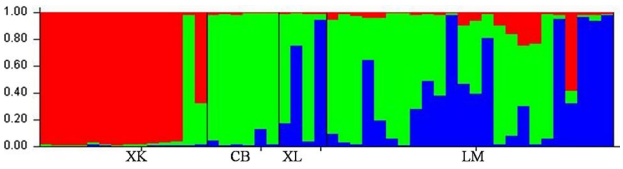

Fig. 8 Results of STRUCTURE analysis (Pritchard et al. 2000) and color-coded grouping at the most likely $K = 3$ as determined using the deltaK method of Evanno et al. (2005). The proportions of cluster memberships for each individual are represented by a *vertical bar*, samples were grouped according to their population of origin as described in Table 1

number of alleles ranged from 2 to 8 (mean = 5), the expected and observed of heterozygosity ranged from 0.209 to 0.813 and 0.335 to 0.842 (mean = 0.455 and 0.639), respectively. The number of alleles ranged from 1 to 17 (mean = 6), whereas the expected heterozygosity and observed heterozygosity were from 0 to 0.925 and 0 to 1.000 (mean = 0.358 and 0.627), respectively (nr-SSRs) (Chen et al. 2013) or the number of alleles ranged from 14 to 49 (mean = 27.5), whereas the observed heterozygosity were 0.578 to 0.916 (mean = 0.734, nr-SSRs) (Bai et al. 2010). The polymorphic information content (PIC) values of these EST-SSR markers ranged from 0.298 to 0.813 (mean = 0.584 ± 0.12). Departures from HWE that we identified for four loci may have been caused by several factors, the most likely include the presence of null alleles, selection, immigration of alleles, and sampling from small populations.

Not all predicted unigenes could be matched to known proteins using BLASTX. Some of the shorter unigenes may have been truncated, lacked a characterized protein domain, or contained a known protein domain that was highly diverged from those in the database, leading to a false-negative result. Some sequence may have been the result of contaminating nuclear DNA. In other cases, non-matching unigenes may have been chimeric reads. In a small number of cases, unigenes from Manchurian walnut were matched to animal proteins, either because of the types of errors described above, or because certain plant proteins are highly similar to proteins associated with the nervous system, proteins used for cell-to-cell communication, or immune system proteins (Baluška et al. 2005; Kwon et al. 2008a, b). If a high-quality reference sequence for Manchurian walnut becomes available it should improve our ability to identify and use EST-SSRs from this species.

There are several advantages to using EST-derived SSRs for genetic studies. They are generally more transferrable to other related species than are genomic-SSRs. Unigenes containing validated SSRs can often be shown to have a function related to development or metabolism, making polymorphisms in these genes useful for identifying

genetic variation in adaptation or physiology, and making them important tools for studying speciation and population genetic differentiation of Manchurian walnut. Identification of SSRs based on expressed sequences from the NCBI databases has been demonstrated (Zhang et al. 2010), but the putative gene function of polymorphism in these SSRs was not described (Chen et al. 2013).

Our analysis of the genetic structure of *J. mandshurica* showed that most of the genetic variation was within populations (87.5 %); the different results reported in this species were based on 22 populations (Bai et al. 2010). The microsatellites we developed were polymorphic enough to distinguish among the 48 individuals and to reveal affinities and differences among the four sampled populations ($F_{ST} = 0.091$, $P < 0.001$). Surprisingly, populations XK and LM, which were sampled from locations relatively close to one another, and showed considerable divergence. The results are not similar with population genetic study using genomic-SSRs (Bai et al. 2010). It is possible that the EST-SSRs we used to analyze the population genetics of *J. mandshurica* produced results that reflect selection pressure on the genetic regions in which the SSRs reside. Despite the differentiation among populations that were observed, there was clear evidence for admixture, probably due to pollen flow, especially in population LM. The basis for genetic divergence among *J. mandshurica* populations is most likely isolation by distance combined with regional adaptation. The heavy seeds of *Juglans* species are typically not dispersed far, but geographic isolation in *Juglans* can be overcome by long-distance pollen flow (Victory et al. 2006), which effectively minimize genetic divergence among wild populations in the Juglandaceae (Bai et al. 2010; Hoban et al. 2010; Pollegioni et al. 2014). A full understanding of the genetic structure and diversity of Manchurian walnut will require additional sampling and the establishment of common gardens. The EST-SSRs described here and other EST-SSRs found within the transcriptome data we generated should prove useful for understanding local adaption, gene flow, and genetic structure in Manchurian walnut and other members of the Juglandaceae.

This study presents the first transcriptome for *J. mandshurica.* It should serve as a foundation and a resource for understanding walnut development and for future genetic and genomic studies of walnut species. Most of the novel set of polymorphic microsatellite loci developed here were within transcribed genes and, as such, may be particularly useful for determining if population differentiation is related to local adaptation. In addition to population genetic studies, the loci may be useful for identifying population units for walnut conservation and management.

Acknowledgments This study was funded by the National Natural Science Foundation of China (Grant No. 41471038, Grant No. 31200500, and Grant No. J1210063), the Program for Excellent Young Academic Backbones funding by Northwest University (Grant No. 338050070), and the Northwest University Training Programs of Innovation and Entrepreneurship for Undergraduates (Grant No. 2015159 and Grand No. 2016171). Mention of a trademark, proprietary product, or vendor does not constitute a guarantee or warranty of the product by the U.S. Department of Agriculture and does not imply its approval to the exclusion of other products or vendors that also may be suitable.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest concerning this article.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

Adal AM, Demissie ZA, Mahmoud SS (2015) Identification, validation and cross-species transferability of novel *Lavandula* EST-SSRs. Planta 241(4):987–1004

Bai WN, Zeng YF, Zhang DY (2007) Mating patterns and pollen dispersalin a heterodichogamous tree, *Juglans mandshurica* (Juglandaceae). New Phytol 176:699–707

Bai WN, Liao WJ, Zhang DY (2010) Nuclear and chloroplast DNA phylogeography reveal two refuge areas with asymmetrical gene flow in a temperate walnut tree from East Asia. New Phytol 188(3):892–901

Baluška F, Volkmann D, Menzel D (2005) Plant synapses: actin-based domains for cell-to-cell communication. Trends Plant Sci 10:106–111

Bodénès C, Chancerel E, Gailing O, Vendramin GG, Bagnoli F, Durand J, Goicoechea PG, Villani F, Mattioni C, Koelewijn HP, Murat F, Salse J, Roussel G, Boury C, Alberto F, Kremer A, Plomion C (2012) Comparative mapping in the Fagaceae and beyond with EST-SSRs. BMC Plant Biol 12:153

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) The SWISSPROT protein knowledgebase and its supplement TrEMBL in 2003. Nucl Acids Res. 31:365–370

Chen CM, Han SJ, Yuan SS, Wang CJ, Yu J (2013) Isolation and characterization of 20 polymorphic microsatellite markers for *Juglans mandshurica* (Juglandaceae). Appl Plant Sci 1:1200009

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

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

Du Q, Gong C, Pan W, Zhang D (2012) Development and application of microsatellites in candidate genes related to wood properties in the Chinese white poplar (*Populus tomentosa* Carr.). DNA Res 20:31–44

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359–361

Ellis JR, Burke JM (2007) EST-SSRs as a resource for population genetic analyses. Heredity 99:125–132

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

Farrar M (2007) Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics 23:156–161

Fu N, Wang Q, Shen HL (2013) De novo assembly, gene annotation and marker development using illumina paired-end transcriptome sequences in celery (*Apium graveolens* L.). PLoS One 8:e57686

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36:3420–3435

Grabherr M, Haas B, Yassour M, Levin J, Thompson D, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Palma FD, Birren BW, Nusbaum C, Lindblad-Ton K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652

He X, Zheng J, Zhou J, He K, Shi S, Wang B (2015) Characterization and comparison of EST-SSRs in Salix, Populus, and Eucalyptus. Tree Genet Genomes 11:820

Hoban SM, Borkowski DS, Brosi SL, McCleary T, Thompson LM, McLachlan JS, Pereira M, Schlarbaum SE, Romero-Severson J (2010) Range-wide distribution of genetic diversity in the North American tree *Juglans cinerea*: a product of range shifts, not ecological marginality or recent population decline. Mol Ecol 19:4876–4891

Izzah NK, Lee J, Jayakodi M, Perumal S, Jin M, Park BS, Ahn K, Yang TJ (2014) Transcriptome sequencing of two parental lines of cabbage (*Brassica oleracea* L. var. *capitata* L.) and construction of an EST-based genetic map. BMC Genom 15:149

Jiang B, Xie D, Liu W, Peng Q, He X (2013) De novo assembly and characterization of the transcriptome, and development of SSR markers in wax gourd (*Benicasa hispida*). PLoS One 8:e71054

Jiang Q, Wang F, Tan HW, Li MY, Xu ZS, Tan GF, Xiong AS (2015) De novo transcriptome assembly, gene annotation, marker development, and miRNA potential target genes validation under abiotic stresses in *Oenanthe javanica*. Mol Genet Genomics 290:671–683

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol Ecol 16:1099–1106

Kaur S, Cogan NO, Pembleton LW, Shinozuka M, Savin KW, Materne M, Forster JW (2011) Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. BMC Genom 12:265

Kim SH, Lee KS, Son JK, Je GH, Le JS, Lee CH, Cheong CJ (1998) Cytotoxic compounds from the roots of *Juglans mandshurica*. J Nat Prod 61:643–645

Kumpatla SP, Mukhopadhyay S (2005) Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome 48:985–998

Kwon C, Neu C, Pajonk S, Yun HS, Humphry M, Bau S, Straus M, Kwaaitaal M, Rampelt H, El Kasmi F, Parker J, Panstruga R, Lipka V, Schulze-Lefert P (2008a) Co-option of a default secretory pathway for plant immune responses. Nature 451:835–840

Kwon C, Panstruga R, Schulze-Lefert P (2008b) Les liaisons dangereuses: immunological synapse formation in animals and plants. Trends Immunol 29:159–166

Lesur I, Le Provost G, Bento P, Da Silva C, Leplé JC, Murat F, Ueno S, Bartholomé J, Lalanne C, Ehrenmann F, Noirot C, Burban C, LégerV Amselem J, Belser C, Quesneville H, Stierschneider M, Fluch S, Feldhahn L, Tarkka M, Herrmann S, Buscot F, Klopp C, Kremer A, Salse J, Aury JM, Plomion C (2015) The oak gene expression atlas: insights into Fagaceae genome evolution and the discovery of genes regulated during bud dormancy release. BMC Genom 16(1):112

Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol 11:2453–2465

Li YC, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21:991–1007

Lin D, Xian D, Xing S, Gui G, Zhang W (2008) Interspecific association analysis in dominant species of *Juglans mandshurica* community in Wuling Mountain nature reserve of Beijing. Journal of Beijing Forestry University 30:154–158 **(in Chinese)**

Long Y, Zhang J, Tian X, Wu S, Zhang Q, Zhang J, Dang Z, Pei XW (2014) 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 Genom 15:1111

Lu AM (1982) The geographical dispersal of Juglandaceae. Acta Phytotaxon Sin 20:257–274

Ma WL, Luo JC, Jing T, Joni K (2007) Study on dynamics of *Juglans mandshurica* population from Changbai mountain. Bulletin of Botany Research 27:249–253 **(in Chinese)**

Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28:2537–2539

Poczai P, Varga I, Laos M, Cseh A, Bell N, Valkonen JP, Hyvönen J (2013) Advances in plant gene-targeted and functional markers: a review. Plant Methods 9:6

Pollegioni P, Woeste KE, Chiocchini F, Olimpieri I, Tortolano V, Clark J, Hemery GE, Mapelli S, Malvolti ME (2014) Landscape genetics of Persian walnut (*Juglans regia* L.) across its Asian range. Tree Genet Genomes 10:1027–1043

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. Mol Ecol Notes 4:137–138

Rousset F (2008) Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Resour 8:103–106

Schlautman B, Fajardo D, Bougie T, Wiesman E, Polashock J, Vorsa N, Steffan S, Zalapa J (2015) Development and validation of 697 novel polymorphic genomic and EST-SSR markers in the American Cranberry (*Vaccinium macrocarpon* Ait.). Molecules 20:2001–2013

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145

Sloan DB, Keller SR, Berardi AE, Sanderson BJ, Karpovich JF, Taylor DR (2012) De novo transcriptome assembly and polymorphism detection in the flowering plant *Silene vulgaris* (*Caryophyllaceae*). Mol Ecol Resour 12:333–343

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting and genotyping errors in microsatellite data. Mol Ecol Notes 4:535–538

Victory ER, Glaubitz JC, Rhodes OE Jr, Woeste K (2006) Genetic homogeneity in *Juglans nigra* (Juglandaceae) at nuclear microsatellites. Am J Bot 93:118–126

Wang R, Xu L, Zhu X, Zhai L, Wang Y, Yu R, Gong Y, Limera C, Liu L (2015) Transcriptome-wide characterization of novel and heat-stress-responsive microRNAs in radish (*Raphanus sativus* L.) using next-generation sequencing. Plant Mol Biol Rep 33:867–880

Wei C, Tao X, Li M, He B, Yan L, Tan X, Zhang Y (2015) De novo transcriptome assembly of *Ipomoea nil* using Illumina sequencing for gene discovery and SSR marker identification. Mol Genet Genomics 290:1873–1884

Woeste K, Burns R, Rhodes O, Michler C (2002) Thirty polymorphic nuclear microsatellite loci from black walnut. J Hered 93:58–60

Xu HL, Yu XF, Qu SC, Zhang R, Qu XR, Chen YP, Ma XY, Sui DY (2010) Anti-proliferative effect of Juglone from *Juglans mandshurica* Maxim. on human leukemia cell HL-60 by inducing apoptosis through the mitochondria-dependent pathway. Eur J Pharmacol 645:14–22

Xu DL, Long H, Liang JJ, Zhang J, Chen X, Li JL, Pan ZF, Deng GB, Yu MQ (2012) De novo assembly and characterization of the root transcriptome of *Aegilops variabilis* during an interaction with the cereal cyst nematode. BMC Genom 13:133

Yan X, Zhang X, Lu M, He Y, An H (2015) De novo sequencing analysis of the *Rosa roxburghii* fruit transcriptome reveals putative ascorbate biosynthetic genes and EST-SSR markers. Gene 561:54–62

Zhang R, Zhu AD, Wang XF, Yu J, Zhang HR, Gao JS, Cheng YF, Deng XX (2010) Development of *Juglans regia* SSR markers by data mining of the EST database. Plant Mol Biol Rep 28:646–653

Zhang J, Liang S, Duan J, Wang J, Chen S, Cheng Z, Zhang Q, Liang X, Li Y (2012) De novo assembly and characterisation of the transcriptome during seed development, and generation of genic-SSR markers in peanut (*Arachis hypogaea* L.). BMC Genom 13:90

Zhang MF, Jiang LM, Zhang DM, Jia GX (2015) De novo transcriptome characterization of Lilium 'Sorbonne' and key enzymes related to the flavonoid biosynthesis. Mol Genet Genomics 290:399–412

Zhao P, Woeste KE (2011) DNA markers identify hybrids between butternut (*Juglans cinerea* L.) and Japanese walnut (*Juglans ailantifolia* Carr.). Tree Genet Genomes 7:511–533

Zhou ZC, Dong Y, Sun HJ, Yang AF, Chen Z, Gao S, Jiang JW, Guan XY, Jiang B, Wang B (2014) Transcriptome sequencing of sea cucumber (*Apostichopus japonicus*) and the identification of gene-associated markers. Mol Ecol Resour 14:127–138

Zhu H, Zhao Y, Lin S, Yang Y, Zhao S, Song L, Jiang Y, Xu B, Liang H, Geng L (2011) The progress of *Juglans mandshurica* Maxim. Chinese Agricultural Science Bulletin 27:1–4 **(in Chinese)**