DOI: 10.1007/s10535-016-0681-8

De novo transcriptome analysis of *Fraxinus velutina* using *Illumina* platform and development of EST-SSR markers

L.-P. YAN[1,2,3], C.-L. LIU[1,2], D.-J. WU[1,2]*, L. LI[1,2], J. SHU[4], C. SUN[1,2], Y. XIA[1,2]*, and L.-J. ZHAO[3]

Shandong Provincial Academy of Forestry, Jinan 250014, P.R. China[1]
Shandong Provincial Key Laboratory of Forest Tree Genetic Improvement, Jinan 250014, P.R. China[2]
Department of Ornamental Horticulture and Landscape Architecture, China Agricultural University, Beijing 10019, P.R. China[3]
Shandong Agricultural Administrators College, Jinan 250014, P.R. China[4]

Abstract

To facilitate functional genomic analysis and molecular breeding of velvet ash (*Fraxinus velutina* Torr), the *de novo* sequencing was carried out by *Illumina* sequencing technology. The cDNA samples were prepared from eleven different tissues of velvet ash and sequenced by using the *Illumina* genome analyzer. Subsequently, *de novo* assemebly, functional annotation databases, and the screening of expressed sequence tag-simple sequence repeats (EST-SSRs) were performed by comparing with corresponding databases using *BLASTx* and software tools. We obtained 51 698 unigenes with an average length of 661 bp and an N50 length of 980 bp. Among all these unigenes, 41 267 (79.8 %) were annotated in the *NCBI* non-redundant protein database and 25 236 (48.8 %) were annotated in the *Swiss-Prot* database. A total of 31 546 (61.0 %) and 13 281 (25.7 %) unigenes were successfully categorized to 59 and 25 functional groups, respectively, by gene ontology categories and clusters of orthologous group categories. A total of 22 323 (43.2 %) unigenes were assigned to 128 pathways using the Kyoto encyclopedia of genes and genomes pathway database. Additionally, 3 249 EST-SSRs markers were detected in 51 698 unigenes from velvet ash. Based on 3 249 EST-SSRs markers, 1 800 primer pairs were successfully designed using *Primer 3*. In the 50 randomly selected primers, 48 successfully amplified fragments, and 42 showed polymorphisms. We completed a successful application of the *Illumina* platform to *de novo* transcriptome assembly of velvet ash, which has the potential to be used for discovering novel genes and further characterization of gene expression profiles.

Additional key words: RNA-Seq; simple sequence repeats, velvet ash.

Introduction

Salinity is one of the most severe abiotic stresses affecting plant growth, which can reduce or even damage nearly all the functions of plants (Gu *et al.* 2012). One of the approaches to solve the problem is exploiting the existing salt-tolerant trees. Velvet ash (*Fraxinus velutina* Torr.) is considered to be one of the saline-alkali tolerant tree and it is a promising tree for reforestation in saline soils (Yu 1996). Meanwhile, it is also one of the most widely cultivated trees in urban landscape and is an important source of hardwood lumber (Bricker and Stutz 2004). As a member of the family *Oleaceae*, velvet ash is a diploid (2n = 22) dicotyledon (Zhang *et al.* 2007). The long lifecycle is a major limiting factor for improving the yield and quality through conventional breeding approaches. The discovery of novel genes and the development of molecular markers linked to the quality and other traits of plants can speed up the breeding pace (Xu and Crouch 2008, Randhawa *et al.* 2013). However, it is difficult to isolate genes which govern important quality and traits of velvet ash due to the scarcity of available genetic sequences. Therefore, more genomic and transcriptomic sequence data of velvet ash are necessary in order to discover new genes related to the quality and traits of velvet ash.

Submitted 13 May 2015, *last revision* 31 March 2016, *accepted* 10 May 2016.

Abbreviations: EST - expressed sequence tag, COG - clusters of orhtologous group; GO - gene ontology; KEGG - Kyoto encyclopedia for genes and genomes; NGS - next-generation sequencing; PIC - polymorphism information content; SSR - simple sequence repeat.

Acknowledgement: This study was supported by the Study on germplasm resources exploration, innovation and utilization of forest tree species (No.2013BA001B06) and the Good Breeds Extension Program of Shandong Province [No.2010(117)].

* Corresponding author; fax: (+86) 0531 88932824, e-mail: zhaolljun@163.com

Next-generation sequencing (NGS) technology is particularly suitable for non-model organisms without prior genome annotation when analyzing their transcriptomes qualitatively and quantitatively (Collins *et al.* 2008), which can provide high-throughput expression data at a single-base resolution (Morozova *et al.* 2009). In consideration of the high throughput, accuracy and low cost, NGS has been successfully used for *de novo* transcriptome sequencing and assembly in many organisms (Berger *et al.* 2010, Jacob *et al.* 2010, Li *et al.* 2010, Wang *et al.* 2010a). Nevertheless, there is nearly no transcriptomic information for velvet ash.

In this study, we preformed *de novo* assembly and gene annotation of transcriptome datasets derived from cDNA samples of several tissues of velvet ash at various growth stages, and made computational identification of expressed sequence tag-simple sequence repeats (EST-SSRs) loci in unigene sets by using clustering and annotation approaches. In addition, whole set of EST-SSR markers directed primer pairs were designed and evaluated to provide a platform of sequence information for discovery of novel genes in velvet ash.

Materials and methods

Plant material and RNA extraction: Velvet ash (*Fraxinus velutina* Torr) was obtained from Shouguang *Fraxinus* Garden, Shandong Province Academy of Experimental Base, Shadong, China. Eleven tissue samples from different developmental stages and culture conditions were collected, including young leaves, mature leaves, tender shoots, stems, young roots, flower buds, and immature seeds under field condition, as well as tissues of leaves, roots, stems, and shoots under tissue culture conditions, which were snap-frozen and stored in liquid nitrogen until analyses.

Total RNA was extracted using *PureLink*^{TM} plant RNA reagent (*Invitrogen*, Carlsbad, CA, USA) from each sample according to the manufacturer's instructions. The quality and quantity of total RNA were analyzed using *Agilent 2100* (Santa Clara, CA, USA) bioanalyzer with a minimum RNA integrity number of 8. Equal quantities of high-quality RNA from each sample were pooled together for cDNA preparation.

The cDNA library construction for *Illumina* sequencing: The cDNA library was prepared according to *Illumina*'s protocols. Briefly, mRNA was purified from 20 µg total RNA using oligo (dT) magnetic beads. Following the purification, mRNA was digested to produce short fragments using fragmentation buffer (*Ambion*, Austin, TX, USA). The first strand of cDNA was synthesized using random hexamer primers followed by synthesis of the second strand. These cDNA fragments were further processed: the end repaired using T4 DNA polymerase, Klenow DNA polymerase, and T4 polynucleotide kinase, and the ligation of adaptors was conducted using *Illumina*'s adaptor oligo mix and T4 DNA ligase (*Invitrogen*). These products were purified and enriched with PCR, and *Solexa HiSeq2000* was employed to sequence the libraries using PCR amplification (*BGI*, Shenzhen, China). In total, we constructed one individual single-end cDNA library.

***De novo* assembly and sequence data analysis:** Before the transcriptome assembly, we carried out a stringent filtering process of raw sequencing reads. The raw reads were cleaned by removing adaptor sequences, duplication sequences, low-quality sequences (reads with ambiguous bases 'N'), non-coding RNA (such as rRNA, tRNA and miRNA), and reads with more than 10 % Q < 20 bases.

De novo transcriptome assembly of the clean reads was performed using *Trinity* (Grabherr *et al.* 2011). Firstly, the reads were extended into contigs according to ovelaps between sequences. Next, the reads were mapped back to contigs with paired-end reads to detect the contigs from same transcripts and distances between them. Susequently, all the contigs were connected into unigenes, which could not be extended on either end. Finally, after splicing and removing redundancy, the unigenes were divided into singletons and clusters (similarity > 80 %) using *BLASTx*. Sequence orientations were determined in accordance to the best hit in the database (at an e-value threshold 1.0 e-5).

Sequence annotation: Unigene annotations were performed by comparing sequences with those unigenes in public databases. The similarities of protein sequences were obtained *via* comparing against the *NCBI* (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/) non-redundant (*Nr*) protein database, and *SwissProt* database using *BLASTx* under e-value < 1.0 e-5. Protein function information was predicted from annotation of the most similar protein in those databases. To investigate the enriched gene ontology (*GO*) terms from three aspects: biological processes, molecular functions, and cellular components, the obtained results were imported into *Blast2 GO* (Conesa *et al.* 2005, Conesa and Götz 2008), which is a widely used software package for annotations of gene functions. These *GO* terms were futher classified *via WEGO* software (Ye *et al.* 2006) for a broad overview of functional distributions of the unigenes. Pathway assignments were all conducted based on *KEGG* database (http://www.genome.jp/kegg/). Moreover, unigene sequences were aligned to clusters of orhtologous group (COG, http://www.ncbi.nlm.nih.gov/COG/) to predict and classify functions.

EST-SSR identification and polymorphism validation: Mining for EST-SSRs from the contig datasets were performed using the Microsatellite identification tool (*MISA*, http://pgrc.ipk-gatersleben.de/misa/), considering di-nucleotides with minimum 6, and tri-, tetra-, penta-, or hexa-nucleotides with minimum 5 contiguous repeat units (Lulin *et al.* 2012). EST-SSRs with both up- and down-stream sequences longer than 150 bp were selected to design 1 800 primer pairs using *Pimer 5* software (Torre *et al.* 2014), with the following criteria: *1*) no SSR existed in the primers; *2*) no more than 3 mismatched bases in the 5'end of the primers; *3*) all primers could be matched to unigenes; and *4*) the same SSR could be identified using these primers *via* ssr_finder as those *via* MISA. Subsequently, 50 EST-SSR primer pairs were randomly selected to assess the assembly quality by reverse transcriptase (RT)-PCR. Briefly, total RNA was extracted from young leaves of 12 different species or cultivars: *Fraxinus sogdiana, F. chinensis*, *F. velutina* cv. Hongyebaila, *F. hupehensis*, *F. velutina* cv. Lula No.2, *F. velutina* cv. Lula No.3, *F. pennsylvanica*, *F. mandshurica*, *F. rhynchophylla*, *F. americana*, *F. excelsior*, and *F. ornus*. PCR amplification was processed in a 0.1 cm^3 of solution containing 25 mM Mg^{2+}, 10 µM primers, 10 mM dNTP, 5 U mm^{-3} Taq DNA polymerase, 5 - 10 ng mm^{-3} template, 10 mm^3 10× PCR buffer and distilled water up to the final volume. Thermal cycling conditions were: initial denaturation at 94 °C for 3 min; 35 cycles at 94 °C for 1 min, 56 °C for 1 min, and 72 °C for 1 min, followed by a final extension at 72 °C for 5 min. Polyacrylamide gel electrophoresis was performed to analyze the amplification products using the *NTSYS-pc 2.10e* software. UPGMA clustering analysis was conducted to identify the genetic relationships. In addition, the gene polymorphism was also investigated by calculating the percentage of polymorphic site (P) and the polymorphism information content (PIC). P was calculated according to formula: P [%] = (k/n) × 100, where k is the amount of polymorphic sites, and n is the total number of the sites.

Results

Each sequenced sample yielded 2 × 90 bp independent reads from either end of a cDNA fragment. In this research, a total of 30 531 178 raw sequencing reads were generated from the cDNA library. After stringent quality assessment and data filtering, approximately 27 175 750 high quality reads were obtained with 97.08 % Q20 bases (those with a base quality greater than 20) and the CG content of 45.52 %.

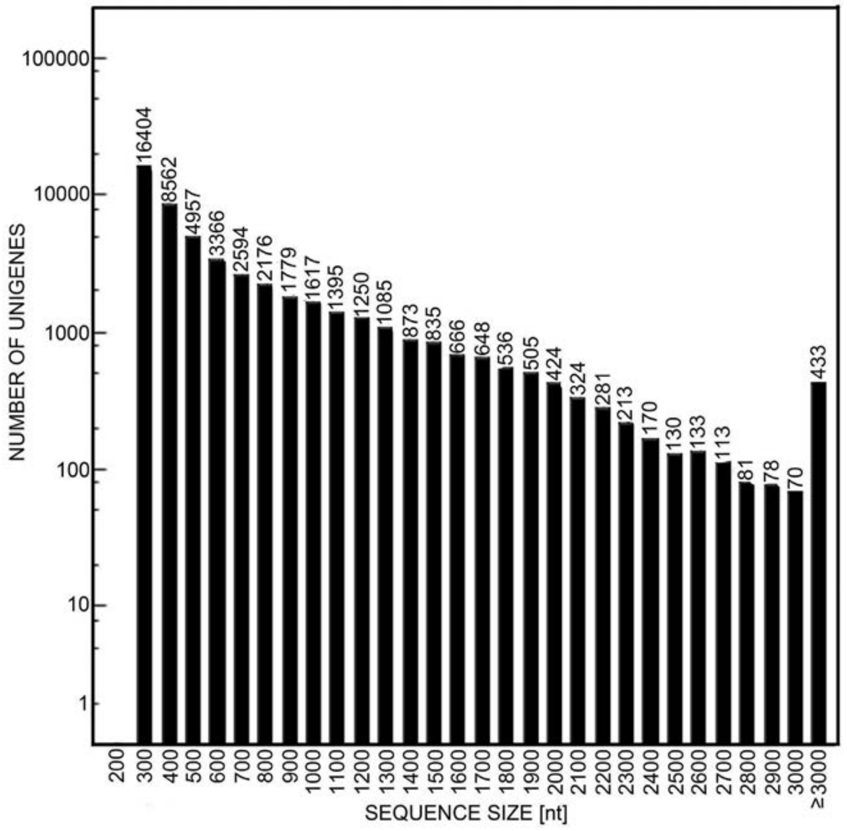

Fig. 1. The size distribution of unigenes in *Fraxinus velutina*. A total of 51 698 unigenes were obtained with 16 489 unigenes between 500 to 1 000 bp, 9 374 unigenes ≥ 1 000 bp, and 2 450 unigenes ≥ 2 000 bp.

We gathered 127 254 contigs, which were longer than 200 bp, with an average length of 301 bp and the N50 length of 452 bp. In total, there were 6 789 contigs longer than 1 kb and 1 171 contigs longer than 2 kb. The assembly of these high quality reads produced 51 698 unigenes with an average length of 661 bp and N50 length of 980 bp. In these unigenes, 16 489 unigenes were between 500 to 1 000 bp, 9 374 unigenes were ≥ 1 000 bp and 2 450 unigenes were ≥ 2 000 bp (Fig. 1).

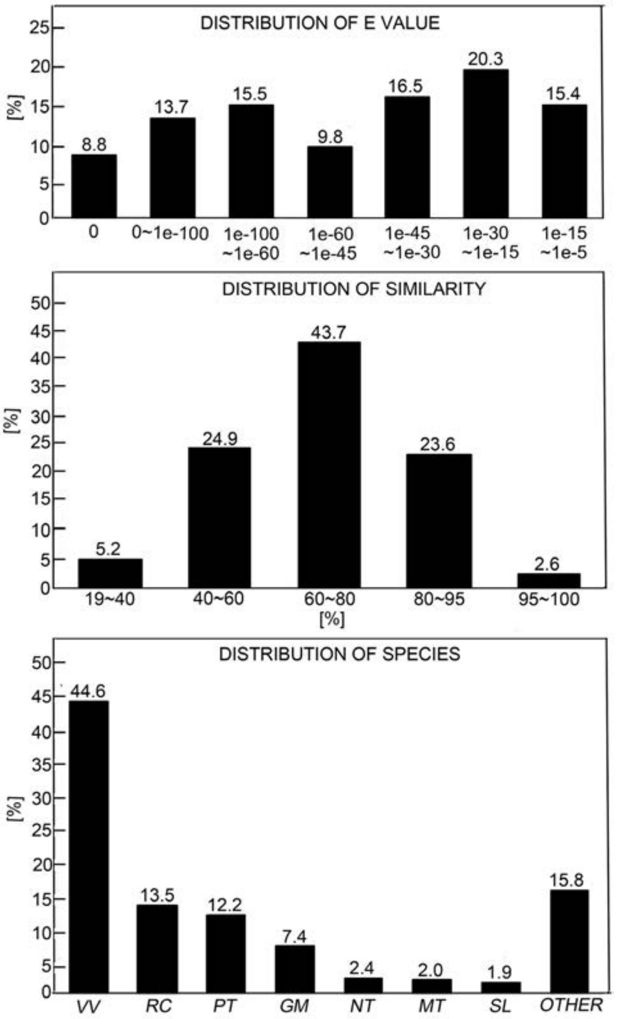

Fig. 2. Distribution of the similarity search of unigenes against the *Nr* database. The distribution of e-value showed 47.8 % of the mapped sequences with strong similarity (smaller than 1.0 e-45), and 52.2 % of the homologous sequences with similarity between 1.0 e-5 and 1.0 e-45. The distribution of similarity showed 26.2 % of the query sequences with a similarity higher than 80 and 68.6 % of the hits with a similarity ranging from 40 to 80 %. The distribution of species showed 44.6 % of the unigenes with the highest similarity to genes from *Vitis vinifera*, followed by *Ricinus communis* (13.5 %), *Populus trichocarpa* (12.2 %), and *Glycine max* (7.4 %). *VV - Vitis vinifera*; *RC - Ricinus communis*; *PT - Populus trichocarpa*; *GM - Glycine max*; *NT - Nicotiana tabacum*; *MT - Medicago truncatula*; *SL - Solanum lycopersicum*.

By comparing with sequences recorded in databases, a total of 41 267 (79.8 %) unigenes were annotated in *NCBI Nr* database, 25 236 (48.8 %) unigenes were annotated in *Swiss-Prot* database, 22 323 (43.2 %) unigenes were annotated in *KEGG* database, 13 281 (25.7 %) unigenes were annotated in the *COG* database, and 31 546 (61.0 %) unigenes were annotated in the *GO* database.

The distributions of the similarity search of unigenes are shown in Fig. 2. The distribution of e-value of the top hits in *Nr* database shows that 47.8 % of the mapped sequences had strong similarity (smaller than 1.0 e-45), while the other 52.2 % of the homologous sequences ranged between 1.0 e-5 and 1.0 e-45 (Fig. 2*A*). The distribution of similarity shows that 26.2 % of the query sequences had a similarity higher than 80 %, and 68.6 % of the hits had a similarity ranging from 40 to 80 % (Fig. 2*B*). The distribution of species of the top *BLASTx* hits against the *Nr* database for the velvet ash transcriptome shows that 44.6 % of the unigenes had the highest similarity to genes from *Vitis vinifera*, followed by other species *Ricinus communis* (13.5 %), *Populus trichocarpa* (12.2 %), and *Glycine max* (7.4 %) (Fig. 2*C*).

Based on *Nr* annotations, we used the *GO* system to classify the possible functions of the unigenes. A total of 31 546 (61.0 %) unigenes were successfully categorized to 59 functional groups (Fig. 3). Of all the three categories of functions, biological process made up the majority of 41.63 %, including “cellular process” and “metabolic process”. There were 40.24 % of unigenes related to cellular component function, of which “cell”, “cell part”, and “organelle” were the top three categories. Within the 18.13 % molecular function categories, genes encoding “binding” proteins and proteins related to “catalytic activity” were the most significant ones. Moreover, we also noticed that a few genes were enriched in the terms of “sulfur utilization”, “cell killing”, “channel regulator activity”, “metallochaperone activity” and “protein tag”.

For the further evaluation of the completeness of transcriptome library and the effectiveness of the annotation process, all unigenes were aligned to the *COG* database to predict and classify possible functions. A total of 13 281 (25.7 %) sequences were assigned to COG classifications (Fig. 4). Among the 25 COG categories, the clusters “general function prediction only” represented the largest group (4 068, 16.7 %), followed by “transcription” (2 111, 8.6 %), “posttranslational modification, protein turnover and chaperones” (1 987, 8.1 %), “replication, recombination, and repair” (1 789, 7.3 %), “translation, ribosomal structure and biogenesis” (1 744, 7.1 %), “signal transduction mechanisms” (1 683, 6.9 %) and “carbohydrate transport and metabolism” (1 551, 6.3 %), whereas 1 080 (4.4 %) unigenes were functionally unknown. Additionally, only a few unigenes were assigned to “nuclear structure” and “extracellular structure”.

To further identify the active biochemical pathways in velvet ash, we mapped the unigenes to the referencing canonical pathways in *KEGG*. In all, the 22 323 unigenes

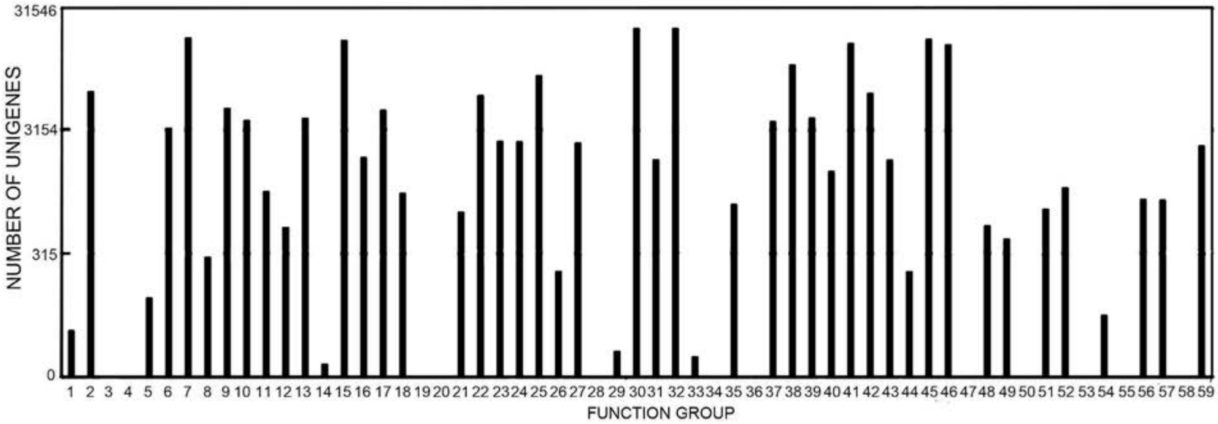

Fig. 3. Gene ontology (*GO*) classification of assembled unigenes. The results are summarized in three main categories: biological process, cellular component, and molecular function. In total, 31 546 unigenes with *BLAST* matches to known proteins were assigned to 59 functional groups.1 - biogical adhesion; 2 - biological regulation; 3 - carbon utilization; 4 - cell killing; 5 - cell proliferation; 6 - cellular component organization or biogenesis; 7 - cellular process; 8 - death; 9 - developmental process; 10 - establishment of localization; 11 - growth; 12 - immune system process; 13 - localization; 14 - locomotion; 15 - metabolic process; 16 - multi-organism process; 17 - multicellular organism process; 18 - negative regulation of biological process; 19 - nitrogen utilization; 20 - pigmentation; 21 - positive regulation of biological process; 22 - regulation of biological process; 23 - reproduction; 24 - reproductive process; 25 - response to stimulus; 26 - rhythmic process; 27 - signaling; 28 - sulfur utilization; 29 - viral reproduction; 30 - cell; 31 - cell junction; 32 - cell part; 33 - extracellular matrix; 34 - extracellular matrix part; 35 - extracellular region; 36 - extracellular region part; 37 - macromolecular complex; 38 - membrane; 39 - membrane part; 40 - membrane-enclosed lumen; 41 - organelle; 42 - organelle part; 43 - symplast; 44 - antioxidant activity; 45 - binding; 46 - catalytic activity; 47 - channel regulator activity; 48 - eletctron carrier activity; 49 - enzyme regulator activity; 50 - metallochaperone activity; 51 - molecular transducer activity; 52 - nucleic acid binding transcription factor activity; 53 - nutrient reservoir activity; 54 - protein binding transcription factor activity; 55 - protein tag; 56 - receptor activity; 57 - structural molecular activity; 58 - translation regulator activity; 59 - transporter activity.

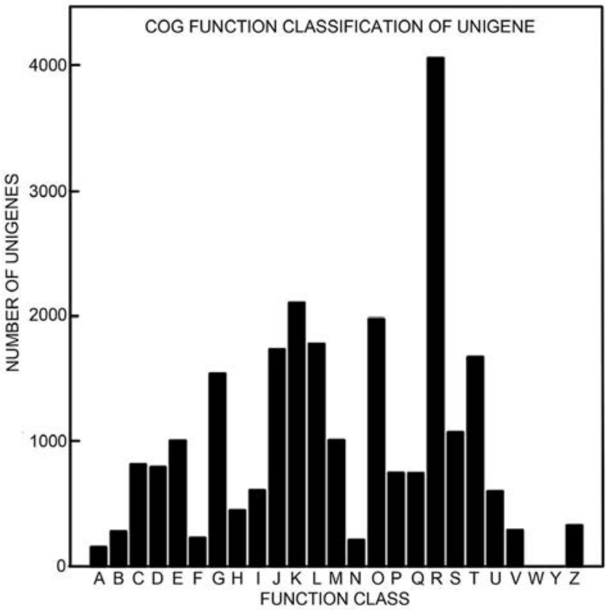

Fig. 4. Clusters of orthologous groups (*COG*) classification. In total, 13 281 (25.7 %) sequences with *Nr* hits were grouped into 25 COG classifications, including "general function prediction only" with the largest group (4 068, 16.7 %), whereas 1 080 (4.4 %) unigenes were functionally unknown. A - RNA processing and modification; B - chromatin structure and dynamics; C - energy production and conversion; D - cell cycle control, cell division, chromosome partitioning; E - amino acid transport and metabolism; F - nucleotide transport and metabolism; G - carbohydrate transport and metabolism; H - coenzyme transport and metabolism; I - lipid transport and metabolism; J - translation, ribosomal structure and biogenesis; K - transcription; L - replication, recombination and repair; M - cell wall/membrane/envelope biogenesis; N - cell motility; O - posttranslational modification, protein turnover, chaperones; P - inorganic ion transport and metabolism; Q - secondary metabolites biosynthesis, transport and catabolism; R - general function prediction only; S - function unknown; T - signal transduction mechanisms; U - intracellular trafficking, secretion, and vesicular transport; V - defense mechanisms; W - extracellular structures; Y - nuclear structure; Z - cytoskeleton.

with significant matches in the *KEGG* database were assigned to 128 *KEGG* pathways. Of those unigenes, 4 978 (21.49 %) were related to metabolic pathways, 2 314 (10.37 %) were related to "biosynthesis of secondary metabolites", 1 393 (6.24 %) to "plant hormone signal transduction", 1 274 (5.71 %) to "plant-pathogen interaction", 862 (3.86 %) to "spliceosome", and 792 (3.55 %) to "RNA transport".

In this study, a total of 3 249 EST-SSR markers were detected in 51 698 unigenes from velvet ash. The number of mono-, di-, tri-, tetra-, penta- and hexa-nucleotide repeats were 702 (21.6 %), 1 259 (38.8 %), 990 (30.5 %), 26 (0.8 %), 71 (2.2 %) and 201 (6.2 %), respectively (Table 1). The EST-SSRs with six tandem repeats (20.6 %) were the most common ones, followed by five tandem repeats (20.0 %) (Table 1). The AG/CT dinucleotide repeat was the most abundant motif (971, 29.9 %), whereas AAG/CTT (319, 9.8 %) and AC/GT (168, 5.2 %)

ranked as the second and third abundant ones, respectively (Fig. 5). Furthermore, two CG/CG (0.06 %) repeats were identified in the databases.

Based on the 3 249 EST-SSR markers, 1 800 primer pairs were successfully designed using *Primer 3*. A total of 50 primer pairs were randomly selected to validate polymorphisms of 12 species/cultivars. Among 50 primer pairs, 42 (P = 84 %) successfully amplified fragments (Table 1 Suppl.), and the PIC ranged from 0.3632 to 0.6725 with 0.543 as the average. Amplification results using primers 3 and 5 well distinguished 12 species/ cultivars (Fig. 6).

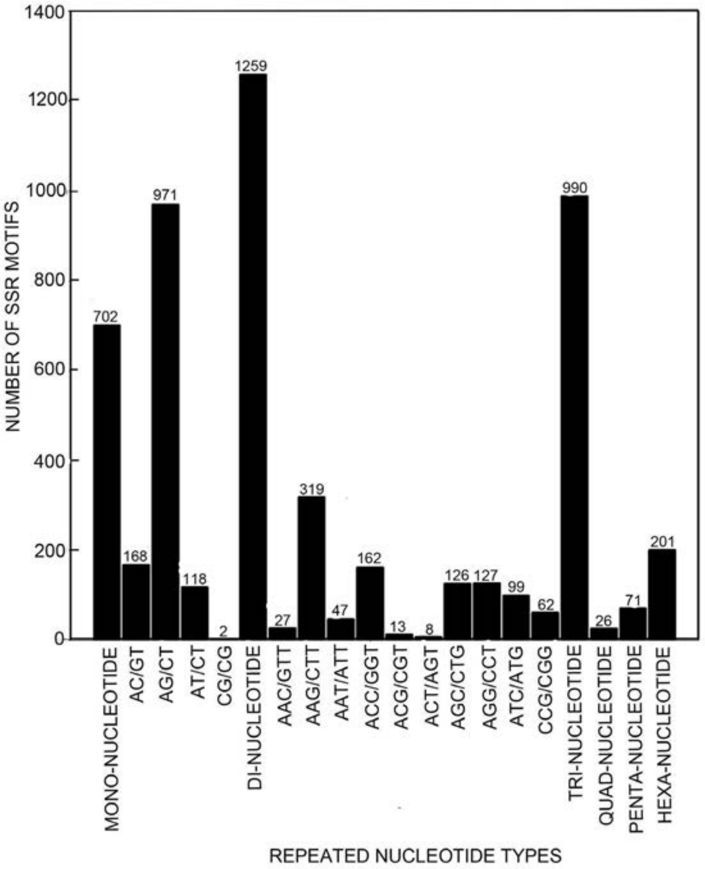

Fig. 5. Characterization of expressed sequence tag-simple sequence repeats (EST-SSR) mining. Among 3 249 EST-SSRs markers, the number of AG/CT dinucleotide repeats was the most abundant motif (971, 29.9 %), followed by AAG/CTT (319, 9.8 %) and AC/GT (168, 5.2 %).

Table 1. Frequency of nucleotide SSR repeat motifs in velvet ash.

Motif types	Repeat numbers										
	4	5	6	7	8	9	10	11	12	≥13	total	[%]
Mono-	-	-	-	-	-	-	-	-	218	484	702	21.6
Di-	-		413	256	210	121	154	95	8	2	1259	38.8
Tri-	-	604	244	113	22	1	0	0	1	5	990	30.5
Tetra-	-	17	9	0	0	0	0	0	0	0	26	0.8
Penta-	56	13	0	1	0	0	1	0	0	0	71	2.2
Hexa-	175	15	3	3	1	3	1	0	0	0	201	6.2
Total	231	649	669	373	233	125	156	95	227	491		
[%]	7.1	20.0	20.6	11.5	7.2	3.8	4.8	2.9	7.0	15.1		

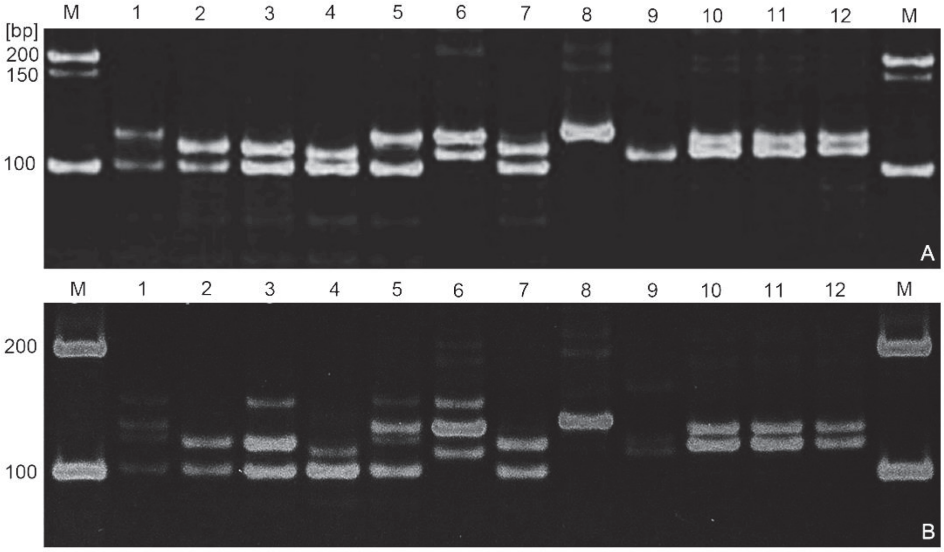

Fig. 6. PCR amplification of genes of 12 *Fraxinus* species/cultivars in electrophoresis gel. Primers 3 (*A*) and 5 (*B*) shows well polymorphisms among them.

Discussion

Transcriptomic information has been applied to a wide range of biological studies including biological processes, the development of molecular markers, such as SSRs and single nucleotide polymorphisms (Cheung *et al.* 2006, Bouck and Vision 2007, Trick *et al.* 2009), as well as the levels of gene expression (Sun *et al.* 2010, Wu *et al.* 2010). In the present study, we adopted NGS technology, which is based on the *Illumina* platform, to obtain the unigenes of velvet ash, as well as predicted and classified possible functions of these unigenes.

We obtained 2.2 Gbp of coverage with 27 175 750 clean sequencing reads. When these sequences were assembled, unigenes with a mean length of 661 bp were obtained, which was longer than those reported in the previous studies using the same technology (Wang *et al.* 2010b, Shi *et al.* 2011, Yang *et al.* 2011). All the velvet ash unigenes were subjected to *BLASTx* analysis against four public databases. In this study, 79.82. % (41 267) unigenes had homologs in *Nr* protein database, whereas it is reported that sesame (Wei *et al.* 2011), Chinese fir (Huang *et al.* 2012), sweet potato (Wang *et al.* 2010c), and whitefly (Wang and Luan 2010) have 54.03, 57.83, 46.21, and 16.20 % unigene homology in *Nr* database, respectively. The higher unigene homology in our study was partly due to the longer unigenes. Specifically, 39.96 % (16 489) of unigenes between 500 and 1 000 bp, 22.65 % (9 374) longer than 1 000 bp, and 5.93% (2 450) longer than 2 000 bp expressed *BLAST* matches, which suggests that longer unigenes were more likely to show *BLAST* hits in protein databases.

SSRs are highly informative and widely used for genetics, evolution and breeding studies. It has been reported that approximately 3 - 7 % of expressed genes contain putative SSR motifs, mainly within the untranslated regions of mRNA (Thiel *et al.* 2003). EST-SSRs are developed as powerful molecular markers for comparative genetic mapping and genotyping since they are ubiquitous in transcriptomes with characteristics of typically locus-specific and codominant, multi-allelic, highly polymorphic, and translatable among species within genera (Yu *et al.* 2004, Varshney *et al.* 2005a,b). In this study, the 3 249 EST-SSRs were identified, which will provide a wealth of markers for further genetic study. Among 50 pairs of high quality PCR primers, 42 of which were successfully yield amplicons at the expected sizes. This result was similar to the successful amplification rates of 60 - 90 % reported previously (Cordeiro *et al.* 2001, Thiel and Michalek 2003, Saha *et al.* 2004), and also provided evidence for the quality validation of our assembled unigenes. Based on these identified EST-SSR-containing sequences, the polymorphism was detected, which might be a valuable resource of genetic markers for future research in *Fraxinus* species.

We successfully constructed the first whole plant unigenes database of velvet ash. The transcriptomic sequences of velvet ash will be used for expanded transcriptome sequencing project of the remaining *Fraxinus* species and applied for more accurate cross-species comparisons with other publicly available genome databases. It turned out that high-throughput RNA-seq is an efficient, inexpensive, and reliable platform for transcriptomic analysis in non-model organisms.

A total of 51 698 unigenes with an average of 661 bp were obtained in velvet ash, which were annotated into 59 *GO* functional categories and 128 pathways. Additionally, 1 800 EST-SSRs were identified as potential molecular markers, and 42 primer pairs were successfully amplified with significant number of polymorphism among 12 *Fraxinus* species/cultivars. Our results provide a new valuable resource for genomic studies on velvet ash.

References

Berger, M.F., Levin, J.Z., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., Johnson, L.A., Robinson, J., Verhaak, R.G., Sougnez, C., Onofrio, R.C., Ziaugra, L., Cibulskis, K., Laine, E., Barretina, J., Winckler, W., Fisher, D.E., Getz, G., Meyerson, M., Jaffe, D.B., Gabriel, S.B., Lander, E.S., Dummer, R., Gnirke, A., Nusbaum, C., Garraway, L.A.: Integrative analysis of the melanoma transcriptome. - Genome Res. **20**: 413-427, 2010.

Bouck, A., Vision, T.: The molecular ecologist's guide to expressed sequence tags. - Mol. Ecol. **16**: 907-924, 2007.

Bricker, J.S., Stutz, J.C.: Phytoplasmas associated with ash decline. - J. Arboricult. **30**: 193-199, 2004.

Cheung, F., Haas, B.J., Goldberg, S.M., May, G.D., Xiao, Y., Town, C.D.: Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. - BMC Genomics. **7**: 272, 2006.

Collins, L.J., Biggs, P.J., Voelckel, C., Joly, S.: An approach to transcriptome analysis of non-model organisms using short-read sequences. - Genome Inform. **21**: 3-14, 2008.

Conesa, A., Götz, S.: Blast2GO: a comprehensive suite for functional analysis in plant genomics. - Int. J. Plant Genomic **2008**: 619832-619844, 2008.

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

Cordeiro, G.M., Casu, R., McIntyre, C.L., Manners, J.M., Henry, R.J.: Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to *Erianthus* and *Sorghum*. - Plant Sci. **160**: 1115-1123, 2001.

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

Gu, J., Weina, L., Akinnagbe, A., Wang, J., Jia, L., Yang, M.: Effect of salt stress on genetic diversity of *Robinia pseudoacacia* seedlings. - Afr. J. Biotechnol. **11**: 1838-1847, 2012.

Huang, H.H., Xu, L.L., Tong, Z.K., Lin, E.P., Liu, Q.P., Cheng, L.J., Zhu, M.Y.: *De novo* characterization of the Chinese fir (*Cunninghamia lanceolata*) transcriptome and analysis of candidate genes involved in cellulose and lignin biosynthesis. - BMC Genomics **13**: 648, 2012.

Jacob, N.M., Kantardjieff, A., Yusufi, F.N., Retzel, E.F., Mulukutla, B.C., Chuah, S.H., Yap, M., Hu, W.S.: Reaching the depth of the Chinese hamster ovary cell transcriptome. - Biotechnol. Bioeng. **105**: 1002-1009, 2010.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O.A., Leung, F.C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Li, Y., Steiner, C.C., Lam, T.T., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M.W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.W., Yiu, S.M., Liu, S., Huang, Y., Yang, G., Jiang, Z., Qin, N., Li, L., Bolund, L., Kristiansen, K., Wong, G.K., Olson, M., Zhang, X., Li, S., Yang, H.: The sequence and *de novo* assembly of the giant panda genome. -Nature **463**: 311-317, 2010.

Lulin, H., Xiao, Y., Pei, S., Wen, T., Shangqin, H.: The first Illumina-based de novo transcriptome sequencing and analysis of safflower flowers. - PloS ONE **7**: e38653, 2012.

Morozova, O., Hirst, M., Marra, M.A.: Applications of new sequencing technologies for transcriptome analysis. - Annu. Rev. Genomics human Genet. **10**: 135-151, 2009.

Randhawa, H.S., Asif, M., Pozniak, C., Clarke, J.M., Graf, R.J., Fox, S.L., Humphreys, D.G., Knox, R.E., Depauw, R.M., Singh, A.K.: Application of molecular markers to wheat breeding in Canada. - Plant Breed. **132**: 458-471, 2013.

Saha, M.C., Mian, M.A., Eujayl, I., Zwonitzer, J.C., Wang, L., May, G.D.: Tall fescue EST-SSR markers with transferability across several grass species. - Theor. appl. Genet. **109**: 783-791, 2004.

Shi, C.Y., Yang, H., Wei, C.L., Yu, O., Zhang, Z.Z., Jiang, C.J., Sun, J., Li, Y.Y., Chen, Q., Xia, T., Wan, X.C.: Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. - BMC Genomics **12**: 131, 2011.

Sun, C., Li, Y., Wu, Q., Luo, H., Sun, Y., Song, J., Lui, E.M., Chen, S.: *De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. - BMC Genomics **11**: 262, 2010.

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

Torre, S., Tattini, M., Brunetti, C., Fineschi, S., Fini, A., Ferrini, F., Sebastiani, F.: RNA-Seq Analysis of *Quercus pubescens* leaves: *de novo* transcriptome assembly, annotation and functional markers development. - PloS ONE **9**: e112487, 2014.

Trick, M., Long, Y., Meng, J., Bancroft, I.: Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. - Plant Biotechnol. J. **7**: 334-346, 2009.

Varshney, R.K., Graner, A., Sorrells, M.E.: Genic microsatellite markers in plants: features and applications. - Trends Biotechnol. **23**: 48-55, 2005**a**.

Varshney, R.K., Sigmund, R., Börner, 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. **168**: 195-202, 2005**b**.

Wang, B., Guo, G., Wang, C., Lin, Y., Wang, X., Zhao, M., Guo, Y., He, M., Zhang, Y., Pan, L.: Survey of the transcriptome of *Aspergillus oryzae via* massively parallel mRNA sequencing. - Nucl. Acids Res. **38**: 5075-5087, 2010a.

Wang, X.W., Luan, J.B., Li, J.M., Bao, Y.Y., Zhang, C.X., Liu, S.S.: *De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development. - BMC Genomics **11**: 400, 2010b.

Wang, Z., Fang, B., Chen, J., Zhang, X., Luo, Z., Huang, L., Chen, X., Li, Y.: *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). - BMC Genomics **11**: 726, 2010c.

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 **12**: 451, 2011.

Wu, T., Qin, Z., Zhou, X., Feng, Z., Du, Y.: Transcriptome profile analysis of floral sex determination in cucumber. - J. Plant Physiol. **167**: 905-913, 2010.

Xu, Y., Crouch, J.H.: Marker-assisted selection in plant breeding: from publications to practice. - Crop Sci. **48**: 391-407, 2008.

Yang, H., Mao, Y., Kong, F., Yang, G., Ma, F., Wang, L.: Profiling of the transcriptome of *Porphyra yezoensis* with Solexa sequencing technology. - Chin. Sci. Bull. **56**: 2119-2130, 2011.

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., Wang, J., Li, S., Li, R., Bolund, L.: WEGO: a web tool for plotting GO annotations. - Nucl. Acids Res. **34**: W293-W297, 2006.

Yu, J.-K., La Rota, M., Kantety, R., Sorrells, M.: EST derived SSR markers for comparative mapping in wheat and rice. - Mol. Genet. Genomics **271**: 742-751, 2004.

Yu, L.: Studies on the cultivation of *Fraxinus velutina* in saline soils of coastal region - J. Liaoning Forest Sci. Technol. **3**: 13-16, 1996.

Zhang, L., Zhang, B.-B., Wang, H.-G.: The karyotype analysis of *Fraxinus velutina*. - J. Wuhan bot. Res. **25**: 513-514, 2007.