EST–SSR marker development and transcriptome sequencing analysis of different tissues of Korean pine (Pinus koraiensis Sieb. et Zucc.)

Jia Du, Zhen Zhang, Hanguo Zhang and Tang Junhong

Department of Environmental Engineering and Science, Hangzhou Dianzi University, Hangzhou, P. R. China; Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Fuyang, P. R. China; School of Forestry, Northeast Forestry University, Harbin, P. R. China

ABSTRACT
Korean pine is a gymnosperm, and gymnosperms have relatively large genome sequences and lack a model organism reference genome. Understanding the important gene expression in the tissue growth process of needles (T1), stems (T2), female flowers (T3) and cones (T4) of the Korean pine is necessary to develop and compound the enzyme genes related to secondary metabolite synthesis. Simple sequence repeat (SSR) markers were explored based on the expressed sequence tag (EST). High-throughput sequencing technology was used to compare and analyse transcriptomes of four different tissue parts of the Korean pine, yielding many differentially expressed unigene sequences. Fluorescently labelled SSR primers were designed to analyse the polymorphism level of 60 open-pollinated families from the Heilongjiang province of China. The research showed that (1) 21.3 GB of data was obtained from the transcriptome sequencing, and 41,476 candidate unigenes were identified based on sequence annotation using various databases. Clusters from orthologous groups and gene ontology function classification tools were used to divide the annotated transcript sequences into 56 functional categories. (2) Cones had the highest number of expressed genes during puberty, with rich expression as they were being formed. (3) By pathway enrichment analysis, 16 key enzyme genes related to fatty acid synthesis in other homologous species were obtained. (4) Ten novel polymorphic fluorescence labelling were used to identify 60 open-pollinated families with a medium polymorphism level. The research showed that high-throughput sequencing technology could analyse the transcriptome expression level between different organisms, and SSR markers were successfully developed.

KEYWORDS
Gene annotation; secondary metabolite biosynthesis; genetic diversity

INTRODUCTION
The tree species Korean pine (Pinus koraiensis), commonly called the Korean pine, is in the genus Pinus of the family Pinaceae. The Korean pine is a grade II national key preserved wild plant of China and is a native tree species in the north-east China forest region [1]. Pinus koraiensis Sieb. has gained more and more attention as the most important tree species for timber and forestation in north-east China. In addition, Korean pine nuts are famous for their medical uses and nutritional value. Korean pine seeds are the most popular among current pine-seed products [2–4], and they contain large amounts of crude protein, crude fat, total sugar and crude fibre, as well as vitamins, minerals and trace elements (calcium, phosphorous, manganese, cobalt, copper and zinc) [4,5].

Currently, in the process of genetic breeding of the Korean pine, edible nut grove breeding and strain improvement are given priority. However, there are few studies on the molecular biology, diversity level and genetic markers of the Korean pine. Korean pine is a gymnosperm, and gymnosperms have relatively large genome sequences and lack a model organism reference genome. New generation sequencing has enabled the sequencing of several conifer genomes despite their very large size, estimated at 20–24 Gbp [6,7]. Understanding the genetic characteristics of Korean pine growth and development, the impact of gene expression on the growth of the Korean pine, the development of enzyme genes in the synthesis process of secondary metabolites and the development of EST–SSR markers can prove fundamental for hybridization breeding, germplasm resource preservation and new germplasm excavation [8,9]. Expressed sequence tag (EST) technology is an effective method for developing transcriptome research. It has been widely applied in the analysis of gene expression and gene function prediction in animals, plants and the microbiology spectrum [10–12].
can reduce the cost and time of sequencing, and abundant data can be obtained with the development of sequencing technology and the applications of high-throughput sequencing to the study of transcriptomes. This research studied transcriptome data from Pinus trees obtained by HiSeq2000 sequencing technology developed by the Illumina company. To study gene expression in the process of growth and development of the Korean pine, we performed research such as building a homogenization cDNA library of the Korean pine, sequencing and gene functional annotation and classification, etc. Then, we obtained difference unigenes, which was helpful for selective breeding and genetic engineering research on the Korean pine and its similar species. We used high-throughput approaches to obtain transcriptome data from the Korean pine. We developed a large collection of transcriptome SSR markers by using computer assistance [13–17]. Simultaneously, we analysed its composition, distribution and characteristics, combined with the SSR fluorescence labelling capillary electrophoresis method. Finally, we found suitable primers for SSR markers of the Korean pine and detected the polymorphism level of 60 open-pollinated families.

Materials and methods

Sample source

Fresh samples were collected from superior clones (No.15 and No. 20) in Linkou seed orchard of the Korean pine, located in Heilongjiang province. The seed orchard was grafted and built in 1979. The clones were propagated by grafting, and no relationships existed between their ortets. We collected four tissues, needles (T1), the top of the stem (T2), female flowers (T3) and cones (T4), in the same period in June 2013. These samples were immediately frozen in liquid nitrogen in the field and then stored at −80 °C for RNA extraction. We extracted RNA from each of the four materials. The extracted RNA was used for transcriptome sequencing. Sixty genome samples were used for EST–SSR amplification. The samples were collected in clonal seed orchards from Weihe, Teili, Hegang and Linkou in the Heilongjiang province (Figure 1). From each seed orchard, 15 open-pollinated families were chosen. Stock plants from four seed orchards were grafted and planted in 1979. The plant spacing was 4 m × 6 m, with a dislocated arrangement in the seed orchard. We collected seeds in the autumn of 2013, and seed-accelerated germination was performed in the winter of 2013. Breeding was conducted in the culturing room. Each family was individually collected and then stored at −80 °C.

RNA isolation, cDNA library creation and HiSeq2000 sequencing

The total RNA extraction method for T1–T4 tissues was performed according to the instructions of Trizol Reagent (Invitrogen, USA). Then, the total RNA was
delivered to Biomarker Technology Company (Beijing, China) for cDNA library construction and sequencing. We used the NEB Next Poly (A) mRNA Magnetic Isolation Module (NEB, E7490) kit to collect mRNA after identifying the quality, integrity and purity of the total RNA. Using RNA as a template, we built a computer library using NEB Next mRNA Library (Library effective concentration~2 nmol/L). Clusters were created with a detection-qualified library. Finally, the sequencing was processed by HiSeqTM 2000.

**Assembly and annotation**

We converted image data into corresponding nucleotide sequence data by using the Illumina platform sequencing. Trinity software was used for the mixed assembly of sample data [18]. Contigs were obtained by assembling overlap information between sequences. According to the similarity between the paired-ends of the sequence information and contigs, we could obtain transcripts by assembling. Through the assembly of components from the selection of potential alternatively spliced transcripts, we selected the longest transcript as the unigene sequence of the sample.

Unigene sequences were compared between databases non-redundant protein, Swiss-Prot, gene ontology (GO), COG and Kyoto Encyclopedia of Genes and Genomes (KEGG) by using BLAST software [19–23]. Comment information on the unigenes was obtained (E-value ≤ 1e-05). Then, the sequences were divided into different categories according to different functions. Finally, gene function information notation by KEGG was commented and predicted by biological pathways.

**Analysis of differential unigene expression in different tissues**

A reads database and a unigene database were obtained by sequencing each sample; then, both of these databases were compared using Bowtie comparison software. Expression levels were estimated using RSEM [24]. According to comparison information, we used RPKM to reflect the expressive abundance of unigenes [25]. Differential expression analysis was selected by EBSeq [26]. We used the False Discovery Rate (FDR) method to analyse unigene expression. In the screening of differential genes, we selected an FDR <0.01 and a fold change ≥2 as the standard definition for differential unigenes. Significant enrichment analysis by GO function could perform the main biological function of differentially expressed genes (DEGs). Through pathway, significant enrichment could identify (DEGs) involved in the main biochemical pathways and signal transduction pathways.

**Identification of SSRs and primer design**

MISA (http://pgrc.ipk-gatersleben.de/misa/misa.html) software was used to find SSR loci [27]. Search criteria included minimum repetitions of 6, 5, 5, 5, 5 times for di-, tri-, tetra- and pentanucleotides, respectively. In total, 101 primer pairs (containing 45 di-, 26 tri-, 25 tetra- and 5 pentanucleotides) were designed using Primer3.0 version 4.0.0 with a preferred amplicon size of 100–500 bp, a GC content of 30%–70%, a Ts value of 50–60 °C and primer lengths of 18–25 bp [28].

**Genome extraction, primer screening and sequencing validation**

The cetyl-trimethyl ammonium bromide method with minor modifications was used for total genomic DNA extraction [29]. A total of 101 pairs of primer were designed for amplification and detection. The Polymerase Chain Reaction (PCR) reaction mixture was 25 μL: 2.0 μL of template DNA (25 ng/μL), 2.5 μL of 10 × PCR buffer, 0.5 μL 1 U Taq enzyme, 0.5 μL of primer (10 μmol/L), 0.5 μL of dNTPs and 18.5 μL of ddH2O. The thermal cycle was as follows: initial denaturation for 5 min at 95 °C, followed by 35 cycles of denaturation for 30 s at 95 °C and annealing at the optimal temperature 30 s, and then 72 °C for 10 min. To detect polymorphic loci with SSR primers, PCR amplification was performed; 6% denaturing polyacrylamide gel electrophoresis was separated and tested; amplification products were collected, followed by sequencing validation (Shanghai Sangon Biological Engineering, Shanghai, China).

Four families of DNA templates were selected for the preliminary screening of primers. Ten pairs of polymorphic loci primers, with stable amplification reactions and clear bands, were chosen for the synthesis of fluorescence labelling primers. After PCR amplification, 1 μL of PCR products was mixed with 8.5 μL of ion formamide and 0.5 μL of LROX3730-500 internal standard, and the mixture was denatured for 5 min at 95 °C, stored at 4 °C for 5 min, and centrifuged for 1 min at 3000 r/min. Automatic fluorescence detection of capillary electrophoresis was performed with an ABI3730DNA analyzer, and then, data were collected and analysed by GeneMapper 4.0 software [30,31].

**High-throughput sequencing and functional annotation**

Prior to sequencing, cDNA samples obtained from four tissues and from individuals were normalized to increase the sequencing efficiency of rare transcripts. Subsequently, a total of 21.3 GB of data were generated from a
full Illumina Hiseq 2000. 4901 106 contig were obtained through sequence splicing (removing low quality and uncertain sequence from the sequencing process). Among these segments, there were 35,331 segments bigger than 300 bp and 14,661 segments bigger than 1 kbp. A total of 71,238 transcripts and 41,476 unigenes were obtained through the Trinity assembly. The N50 of transcripts and unigenes was 1855 and 1826 bp, respectively. As shown in Figure 2, after BLAST sequence alignment, 25,824 (62.26% of the total number) unigenes with annotation information were eventually obtained. In total, 41,476 unigenes of the Korean pine were obtained by assembling, and then, they were compared with the Nr database. A total of 11,837 unigenes (45.84%) showed sequence homology with the spruce (*Picea sitchensis*). A total of 2830 unigenes (10.96%) showed sequence homology with the grape (*Vitis vinifera*). In total, 41,476 unigenes were distributed into one or more GO terms, with 23.75% for cellular components, 27.24% for molecular functions and 49.01% for biological processes (Figure 3). In addition, compared with the KEGG database, 5356 unigenes (19.95%) were annotated by a pathway and related to 118 KEGG pathways.

**GO enrichment analysis of differentially expressed genes (DEGs)**

Annotated genes from the different tissues were used to perform a differential comparative analysis. The results showed that the three tissues of needles (T₁), stems (T₂) and flowers (T₃) showed higher differential gene numbers than fruits (T₄) (Figure 4). Counting of variable genes revealed that the number of DEGs with down-regulated expression was larger than those with up-regulated expression for the three tissues and fruits. Needles and fruits had the highest number of variable genes, which suggested that fruits had the highest number of expressed genes during the developmental period, and cone expression was rich in the formation process.

GO enrichment analysis of the DEGs showed that in the different tissues, the unigenes were divided into three groups: components, molecular function and biological process. The proportion of the three sub-classes, cell, organelle and cell parts, was the highest in the cellular components group. The proportion of the two subclasses, catalytic activity and binding, was the highest in the molecular function group. Metabolic process and cellular process were the highest proportion in biological process group (Figure 5). GO enrichment correlation analysis of unigene variation was performed between different tissues based on this grouping. (T₁) vs. (T₄) were used as examples (Figure 6). There were significant differences between leaf tissue and cone unigenes regarding sequence-specific DNA binding transcription factor activity (122), oxidoreductase activity (63), metabolic processes (183), oxidation-reduction processes (298), carbohydrate metabolic processes (83), the extracellular region (129), plant-type cell wall (95), cytoplasmic membrane-bounded vesicles (190) and apoplasts (91) in the formative period (P-value ≤ 0.05).
**Figure 3.** Gene ontology (GO) distributions for the *Pinus koreansis* transcriptome.

**Figure 4.** MA plot of differentially expressed genes in annotations of different tissues.
Pathway annotation analysis of different unigene in different organizations

We used the KEGG database to do the pathway enrichment analysis (Figure 7), in order to study the metabolic pathways of different unigenes involved. For example, the research found out that needles (T1) and fruit (T4) had significant differences in the pathway of phenylpropanoid biosynthesis (61), phenylalanine metabolism (43), photosynthesis (31), circadian rhythm (17), flavonoid biosynthesis (25), diterpenoid biosynthesis (13) and protein processing in endoplasmic reticulum (47). The data of the top three different unigene pathway showed that the pathways were consistent between the T1 and T2, T1 and T3, the shared pathways were phenylpropanoid biosynthesis, phenylalanine metabolism and photosynthesis. In addition, the pathways were consistent between T1, T2, T3 and T4. The shared pathways were phenylpropanoid biosynthesis and protein processing in endoplasmic reticulum. In the end, the shared pathways were phenylpropanoid biosynthesis, phenylalanine metabolism and plant hormone signal transduction between T2 and T3.

Unigenes of pathway in fatty acid synthesis

There were 669 DEGs annotated to metabolic pathways of secondary metabolite unigenes in the transcriptome database, and the metabolic pathways had 25 different
branches. Forty-eight unigenes could be associated with the fatty acid biosynthesis pathway by pathway enrichment analysis from the transcriptome data of the different tissues. Fatty acid synthesis is the foundation of fat synthesis. Study of the genes involved in the fatty acid synthesis pathway would help to further understand the mechanism of fat synthesis. A pairwise comparison of DEGs was performed on each tissue. Some of the DEGs in each group were classified as associated with the fatty acid biosynthetic pathway after pathway enrichment analysis. Sixteen key enzyme genes, homologous with other species, were obtained after a comparison between 48 unigene sequences in the KEGG database. The average fragment length was 1397.65 bp. Among the genes, a ketoacyl-ACP synthase (KASIII) fragment sequence was the longest, with 2542.5 bp. Palmitoyl-ACP thioesterase (FatB) and stearoyl-ACP desaturase (SAD) had the highest expression level. Moreover, these two genes were at the highest level of expression in cone tops (Figure 8).

The number and distribution of SSR loci

A total of 1757 SSRs conforming to the definitions were recovered from 1549 sequences. The distribution frequency (the number of SSRs compared with the total number of sequences) of SSR loci was 4.24%. One SSR locus was found in an average of 17.38 kbp. The SSR loci contained 1–5 repeat types. The types of repeats were mainly concentrated among mononucleotide to trinucleotide repeats, which accounted for 98.68% of the total SSRs. The average unit length for di-, tri-, tetra- and pentanucleotide repeats were 11.81, 14.34, 16.37, 21.11 and 20 bp, respectively. The most abundant repetitive sequence length was 10–14 bp, which reached a 52.82% proportion; the proportion of 15–18 bp was 37.56%; the proportion of 19–22 bp was 7.63%; the proportion of ≥23 bp was 1.94%.

The 1757 SSR loci contained 31 unit types. Among them, mono-, di-, tri-, tetra- and pentanucleotide unit types numbered 4, 4, 10, 9 and 4, respectively. Five unit types of high frequency were A/T, AT/AT, AGC/CTG, AGG/CCT and AAG/CTT, which accounted for 46.1%, 10.76%, 7.63%, 6.66% and 6.55% of the total number of SSRs, respectively.

Screening and verification of EST–SSR primers

In total, 101 pairs of EST–SSR primers were designed and synthesized, which included 2–5 nucleotide repeat units. In the prescreening of all primer pairs with an individual pair randomly chosen from each seed orchard, we successfully amplified 21 SSRs (20.8%). Twenty-one pairs of primers were tested by detecting polymorphic loci, and then, PCR amplification was performed; 6% denaturing polyacrylamide gel electrophoresis was used to separate
and test the PCR products, and amplification products were collected. Sequencing validation was then conducted (Shanghai Sangon Biological Engineering, Shanghai, China). The results showed 16 SSRs (76.19%, containing 2 di-, 12 tri-, 2 tetra-) among the 21 pairs of primers with the expected sizes (Table 1).

**Polymorphism analysis of SSR loci**

Ten SSRs were selected to synthesize fluorescence labeling and detect polymorphisms in 60 open-pollinated families [32]. Four capillary electrophoresis templates were amplified by the P92 primers (Figure 9). Data analysis by GeneMapper4.0 software showed that the fragment size of the amplification products was between 110 and 432 bp. Detection of 10 SSR loci of genetic variants in 60 samples is shown in Table 2. The polymorphism information content (PIC) detected from these 10 SSR loci was 0.1168–0.6913, with an average of 0.4500. According to the criteria of Botstein et al., half of the 10 novel polymorphic SSRs demonstrated highly informative scores (PIC > 0.50), and the other polymorphic SSRs had reasonably informative scores (0.50 > PIC > 0.25), except for the SSRs P.74 and P.70 (PIC < 0.25) [33]. The results suggested that these markers could be considered an important marker resource in Korean pine population genetic studies.

**Results and discussion**

A large number of ESTs for pines have been sequenced to date. Large numbers of ESTs have been sequenced

---

**Table 1. SSR primer sequences and repeat types.**

| Primer No. | Primer sequence (5′—3′) | SSR type | Tm/°C |
|------------|-------------------------|----------|-------|
| P11        | F: TGAGAATAGGGCGAATCG T: GAGAAGAAGTTAAGGGGA | (CT)_6 | 53    |
| P25        | F: AAAGTACAGATTGCACTC T: TCAGTCAAGCAGCAGAAG | (CT)_6 | 55    |
| P44        | F: TTTGCTTCTCTGCTGCT T: CCGTGTGGTACATGAC | (ATG)_1, (AGC)_1 | 53    |
| P47        | F: GTTTGGAGATCCGGACAC T: CAGTATATATCCGGTTT | (AT)_5 | 51    |
| P49        | F: GAGATGGACAGTCAGGGG T: TACAGTGTCACCTACGG | (AAAC)_2 | 52    |
| P60        | F: AAAGCCAGCTGGAGGAAG T: AACTCAGAGACTTGGTG | (CTG)_6 | 52    |
| P62        | F: AGTTGCTACGTGCTGAG T: AACATTGAAGCTGGAGG | (GCA)_6 | 51    |
| P63        | F: GCACGAGATCGAGGGAGG T: CAGCCACAAACTTGGTACAC | (CAAC)_5 | 56    |
| P67        | F: TGACGCAAGGCAGAAGTT T: GCGAGCGAAGTGGAAA | (ACAA)_5 | 52    |
| P70        | F: CAACATGCAAGAHTCCTC T: CCTACCTCGCTCGCT | (CTCA)_6 | 54    |
| P72        | F: TGCTGTACCTCCTTCAC T: CAATGAGCTTGGAGCA | (CTCA)_6 | 52    |
| P74        | F: AGCTTACCGATCTCCTACC T: GTTCCGGCTACTACAT | (GAA)_6 | 52    |
| P79        | F: CCACCGCAGATGCCTCATTA T: GCCTTTGTTAGCGTCGCG | (CAA)_6 | 55    |
| P82        | F: GCAAGATGACAGCTGGAAACC T: ACACCCGGCCAAGAGCAGCA | (CCGG)_6 | 54    |
| P90        | F: CGCCAAATCTCGAGACTGAT T: GCAGCAGCAAGATAGAAACC | (CCA)_6 | 56    |
| P92        | F: ACTTTGGCTGTAATCGAGACC T: AAAGTTAGGGGCTGGTACAG | (CAAC)_4 | 53    |

---

**Figure 9. No. P92 SSR locus genotype of HG2, LK19, TL1048 and WH056.**
and analysed in pine to discover growth and wood quality trait-related genes [34–38]. But researches are few in the different organizational development system of Korean pine. Especially, Korean pine nuts are the product of an edible woody oil tree species with oil content as high as 56.76%. The nutritional components of the seeds of the *P. koraiensis* trees exhibited rich variation; the variation coefficient of the fatty acids was 2.24%–66.83% [39]. Our research obtained 41,476 unigenes and 25,824 annotation information unigenes by assembly. The DEGs of different Korean pine tissues were subjected to GO function classification and enrichment analysis of unigene differences. The results showed that the most unigene differences were observed between cones and needles, tender stems and flowers. The results also suggested that cones had abundant biochemical processes and a high level of gene expression during the developmental period. Li et al. [40] used high-throughput sequencing technology to perform a transcriptome analysis between different tissues and to find saponin biosynthesis genes. Some researchers have used high-throughput sequencing for transcriptome sequencing and to excavate gene resources in Norway spruce, lopede pine and *Pinus tabulaeformis* [34–36]. This study confirmed that high-throughput sequencing of the Korean pine is feasible for studying the transcription levels between different tissues.

Genetic improvement of the Korean pine had, so far, mainly involved traditional selection and crossing based on phenotypes. Preliminary showed that phenotypic traits have abundant genetic variation [40]. The developed SSR markers could be used for fingerprinting of superior clones, paternity analysis and pedigree reconstruction for breeding populations [8,13,14,41,42]. Thus, we expect these markers to significantly enhance the accuracy and progress of the Korean pine breeding program. Furthermore, our subsequent development of large-scale EST-based SSR markers should greatly facilitate Quantitative Trait Loci (QTL) mapping, Molecular Assisted Selection (MAS) and comparative genomics studies. SSRs have emerged as an important marker system in many areas of molecular genetics because they exhibit a high degree of allelic diversity. The identification and development of SSR markers represents significant challenges for non-model organisms and organisms with extremely large genomes [8].

Although traditional sequencing yields longer EST sequences, it has little advantage compared to new assembly technology based on the large-scale ESTs. Additionally, most previous studies used a cDNA library of one tissue, whereas we used a normalized cDNA library comprising multiple tissues and individuals. These large-scale ESTs will provide more comprehensive pine transcriptome information and facilitate the assembly of *Pinus koraiensis* ESTs in the future. The Korean pine is a gymnosperm with relatively large genome sequences. It is a model organism without a suitable reference genome that must be assembled *de novo* from transcriptome data. We examined the characteristics of 2–5 bp unit-type microsatellites in the Korean pine based on next-generation genome sequencing technology. The distribution frequency of SSRs was 4.24%, which is less than the American black poplar (*Populus deltoides*) (14.83%), and Oil-tea (*Camellia oleifera*) (6.7%) but greater than the Loblolly pine (*Pinus taeda*) (4.32%) and the Masson pine (*Pinus massoniana*) (1.4%) [15,43–45].

In the study of different tree species, the proportion and type of SSR repetitive sequences in ESTs vary. This study could amplify polymorphic loci with primers, and priority was given to dinucleotide and trinucleotide repeat units. Codons are a functional unit with three nucleotides; therefore, a three-nucleotide displacement would not have a large effect on the reading frame of expressed genes. Therefore, the trinucleotide repeat SSR types are the most frequent in EST sequences, while dinucleotide and trinucleotide types are the common type of transcriptome SSR repeat unit in most species [46]. Our study is consistent with the above research. Trinucleotide repeats predominated, accounting for 34.66% of the total number of SSRs; dinucleotide repeats accounted for 17.12% of the total number of SSRs. At the same time, the results verified that the polymorphism of low-level SSR units was higher than that of senior units, so designed SSR primers should focus on containing the sequence of the lower unit [47,48].

EST–SSRs have a great advantage in characterizing and protecting the genetic diversity of germplasm resources, but they also have some shortcomings. Because EST–SSRs come from functional gene sequences of relative conservation; therefore, species-specific genomic SSR markers generally have higher polymorphism

**Table 2.** Products amplified by the SSR primers and their polymorphisms.

| SSR primer | Overall allele frequency | Shannon’s Information index (I) | Polymorphism information content (PIC) |
|------------|--------------------------|--------------------------------|--------------------------------------|
| P92        | 0.1083–0.625             | 0.8770                         | 0.4601                               |
| P60        | 0.075–0.7333             | 0.8666                         | 0.4103                               |
| P63        | 0.05–0.325               | 1.4291                         | 0.6913                               |
| P74        | 0.0667–0.9333            | 0.2449                         | 0.1168                               |
| P79        | 0.2083–0.7917            | 0.5117                         | 0.2754                               |
| P90        | 0.1333–0.8667            | 0.3927                         | 0.2043                               |
| P11        | 0.2417–0.425             | 1.0731                         | 0.5757                               |
| P25        | 0.3083–0.375             | 1.0947                         | 0.5900                               |
| P44        | 0.30–0.35                | 1.0961                         | 0.5909                               |
| P47        | 0.2583–0.375             | 1.0853                         | 0.5837                               |
| Mean       |                          | 0.8681                         | 0.4500                               |
levels than EST-based ones, which is useful for fingerprinting and the pedigree reconstruction of closely related individuals [49–51]. SSR polymorphism molecular markers are based on the different branches of fragments amplified in micro-satellite repeat number, but they cannot effectively distinguish between SSRs of the same length and different repeat sequence and SSRs with a similar length. With increasing EST sequencing and the development of new EST–SSR markers in public databases, ESTs are obtained based on more effective high-throughput sequencing technologies, helping to increase the number of polymorphic SSR primers and to make up for the low polymorphism of EST–SSR markers. At the same time, the reliability and accuracy of the test can be improved by combining with other comparative studies on molecular markers, such as markers, Single Nucleotide Polymorphisms (NP), etc. In this research we adopted the ABI3730DNA analyzer instrument, using capillary electrophoresis, PCR product segments amplified by 10 pairs of fluorescence labelling SSR primers; this approach overcomes the artificial estimation error caused by fragment size inherent to polyacrylamide gel electrophoresis methods.

Conclusions

In this research, high-throughput sequencing technology was used to analyse the transcriptome expression level of different organizations. The DEGs of different Korean pine tissues were subjected to GO function classification and enrichment analysis of unigene differences, and SSR markers were successfully developed. EST was obtained based on effective high-throughput sequencing technologies. High-throughput sequencing technologies were contributed to increase the numbers of polymorphic SSR primers, and made up polymorphism of EST–SSR marker. EST–SSR markers are useful for the studies of population genetics in Korean pine.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The authors would like to acknowledge the financial support from the special research of Public Welfare Forestry Industry ‘Korean pine improved type selection and directive breeding technology research [grant number 201204320]’, and Project supported by National Natural Science Foundation of China [grant number 31600533].

References

[1] Lim TK. Edible medicinal and non-medicinal plants. Netherlands: Springer. 2013, p. 297–303.
[2] Wolff RL, Pedrono F, Pasquier E, et al. General characteristics of Pinus spp. seed fatty acid compositions, and importance of Delta 5-olefinic acids in the taxonomy and phylogeny of the genus. Lipids. 2000;35:1–22.
[3] Yoon TH, Im KJ, Koh ET, et al. Fatty-acid compositions of Pinus Koraiensis seed. Nutr Res. 1989;9:357–361.
[4] Nergiz C, Donmez I. Chemical composition and nutritive value of Pinus pinea L. seeds. Food Chem. 2004;86:365–368.
[5] Zadernowski R, Naczk M, Czaplicki S. Chemical composition of Pinus sibirica nut oils. Eur J Lipid Sci Technol. 2009;111:698–704.
[6] Julien P, Fukuta-Pekka V, John JM. Conifer genomics and adaptation: at the crossroads of genetic diversity and genome function. New Phytol. 2016;209:44–62.
[7] Neale DB, Wegrzyn JL, Stevens KA, et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 2014;15:RS59.
[8] Wei YC, Liu QH, Dong HY, et al. Selection of reference genes for real-time quantitative PCR in Pinus massoniana post nematode inoculation. PLoS One. 2016;11:e0147224.
[9] Yang HB, Zhang R, Jin GQ, et al. Assessing the genetic diversity and genealogyal reconstruction of cypress (Cupressus funebris Endl.) breeding parents using SSR markers. Forests. 2016;7:2–14. DOI:10.3390/f7080160
[10] Adams MD, Kelley JM, Gocayne JD, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 1991;252:1651–1656.
[11] Boguski MS, Tolstoshev CM, Bassett DE. Gene discovery in dbEST. Science. 1994;265:1993–1994.
[12] Bouchez D, Herman H. Functional genomics in plants. Plant Physiol. 1998;118:725–732.
[13] Yi G, Lee JM, Lee S, et al. Exploitation of pepper EST-SSRs and an SSR-based linkage map. Theor Appl Genet. 2006;114:113–130.
[14] Portis E, Nagy I, Sasvari Z, et al. The design of Capsicum spp. SSR assays via analysis of in silico DNA sequence, and their potential utility for genetic mapping. Plant Sci. 2007;172:640–648.
[15] Bai TD, Xu LA, Xu M, et al. Characterization of Masson pine (Pinus massoniana Lamb.) microsatellite DNA by 454 genome shotgun sequencing. Tree Genet Genomes. 2014;10:429–437.
[16] Parchman TL, Geist KS, Grahnen JA, et al. Transcriptome sequencing in an ecologic ally important tree species: assembly, annotation, and marker discovery. BMC Genomics. 2010;11:180–195.
[17] Pootakham W, Chanprasert J, Jomchai N, et al. Development of genomic-derived simple sequence repeat markers in Hevea brasiliensis from 454 genome shotgun sequences. Plant Breed. 2012;131:555–562.
[18] Manfred GG, Brian JH, Moran Y, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–652.
[19] Deng YY, Li JQ, Wu SF. Integrated nr database in protein annotation system and its localization. Comput Eng. 2006;32:71–74.
[20] Apweiler R, Bairoch A, Wu CH, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2004;32:115–119.

[21] Michael A, Catherine AB, Judith AB, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–29.

[22] Roman LT, Michael YG, Darren AN, et al. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2004;32:277–280.

[23] Minoru K, Susumu G, Shuichi K, et al. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32:115–119.

[24] Glaubitz JC. CONVERT: a user-friendly program to reformat PCR fragments. Nat Biotechnol. 2000;18:233–236.

[25] Ning L, John AD, James AT, et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013;29:1035–1043.

[26] Guo SG, Zheng Y, Joung JG, et al. Transcriptome sequencing and comparative analysis of Monocotyledonous plants. Plant Mol Biol. 2005;57:203–224.

[27] Zhang Z, Zhang HG, Yang CP, et al. Clonal variation in nutritional components of Korean pine (Pinus Koraiensis) seed collected from seed orchards in northeastern China. J Forest Res. 2016;27(2):295–311.

[28] Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. 2009;25:41–44.

[29] Porebski S, Bailey LG, Baum BR. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol Biol Rep. 1997;15:8–15.

[30] Gloubitz JC. CONVERT: a user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. Mol Ecology Notes. 2004;4:309–310.

[31] Liu KJ, Muse SV. PowerMarker: integrated analysis environment for genetic marker data. Bioinformatics. 2005;21:2121–2129.

[32] Markus S. An economic method for the fluorescence labeling of PCR fragments. Nat Biotechnol. 2000;18:233–236.

[33] Botstein D, White RL, Skolnick M. Construction of a genetic linkage maps in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980;32:314–331.

[34] Chen J, Severin U, Nicolas G, et al. Sequencing of the needle transcriptome from Norway spruce (Picea abiesKart L.) reveals lower substitution rates, but similar selective constraints in gymnosperms and angiosperms. BMC Genomics. 2012;13:859–604.

[35] Dawn EH, Macaire MSY, Sharon J, et al. Transcriptome resources and functional characterization of monoterpene synthases for two host species of the mountain pine beetle, lodgepole pine (Pinus contorta) and jack pine (Pinus banksiana). BMC Plant Biol. 2013;13:80–94.

[36] Niu SH, Li ZX, Yuan HW, et al. Transcriptome characterisation of Pinus tabuliformis and evolution of genes in the Pinus phylogeny. BMC Genomics. 2013;14:263–275.

[37] Liu QH, Zhou ZC, Wei YC, et al. Genome-wide identification of differentially expressed genes associated with the high yielding of oleoresin in secondary xylem of Masson pine (Pinus massoniana Lamb) by transcriptomic analysis. PLoS One. 2015;10:e0132624. DOI:10.1371/journal.pone.0132624.

[38] Pavy N, Laroche J, Bousque J, et al. Large-scale statistical analysis of secondary xylem ESTs in pine. Plant Mol Biol. 2005;57:203–224.

[39] Zhang Z, Zhang HG, Yang CP, et al. Clonal variation in nutritional components of Korean pine (Pinus Koraiensis) seed collected from seed orchards in northeastern China. J Forest Res. 2016;27(2):295–311.

[40] Li CF, Zhu YJ, Guo X, et al. Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in Panax ginseng C. A. Meyer. BMC Genomics. 2013;14:245–256.

[41] Guan L, Suharyanto SS. Isolation and characterization of tetranucleotide microsatellite loci in Pinus massoniana (Pinaceae). Am J Bot. 2011;98:216–217.

[42] Hung KH, Lin CY, Huang CC, et al. Isolation and characterization of microsatellite loci from Pinus massoniana (Pinaceae). Bot Stud. 2012;53:191–196.

[43] Zhang XY, Song CW, Zhang YD, et al. Development of EST-SSR in Populus deltoides and P. euramerica. Scientia Silvae Sinicae. 2009;45:53–59.

[44] We Q, Xu LC, Jiang XM. Survey and analysis of microsatellites from RNA sequences in Camellia species using 454 pyrosequencing. Scientia Silvae Sinicae. 2013;49:43–50.

[45] Lin YZ, Guo H, Liu CX. EST-SSR prediction of heat induced cDNA library of Pinus teada. J South Chin Agric Univ. 2009;30:41–44.

[46] Varshoey RK, Thiel T, Stein N. In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett. 2002;7:537–546.

[47] Dreisigacker S, Zhang P, Warburton ML, et al. SSR and pedigree analyses of genetic diversity among CIMMYT wheat lines targeted to different mega-environments. Crop Sci. 2004;44:381–388.

[48] Sonali S, Lin YC, Zuccolo A, et al. Comparative in silico analysis of EST-SSRs in angiosperm and gymnosperm tree genera. BMC Plant Biol. 2014;14:220–230.

[49] Eujayl IA, Sorrells ME, Baum M, et al. Assessment of genotypic variation among cultivated durum wheat based on EST-SSRs and genomic SSRs. Euphytica. 2001;119:39–43.

[50] Chagné D, Chaumeil P, Ramboer A, et al. Cross-species transferability and mapping of genomic and cDNA SSRs in pines. Theor Appl Genet. 2004;109:1204–1214.

[51] Liewalaksaneeyawan C, Ritland CE, El-Kassaby YA, et al. Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. Theor Appl Genet. 2004;109:361–369.