Plant Biochemistry & Physiology

Kong et al., J Plant Biochem Physiol 2013, 1:2
http://dx.doi.org/10.4172/jpbp.1000109

Short Communication Open Access

Analyzing the Nucleotide Variations within the Expressed Sequence Tags of Loblolly Pine (*Pinus taeda*)

Fanming Kong, Xiaolong Wang, Yingnan Chen, Aihong Bian, Jin Xu and Tongming Yin*

Faculty of Forest Resources and Environmental Sciences, Nanjing Forest University, Nanjing 210037, China

Keywords: Insertions; Deletions; Nucleotide variation; Transition; Substitution; Pine

Introduction

Single nucleotide polymorphisms (SNPs) represent the most frequent variations in eukaryotic genomes. For example, the frequency of SNPs is one per kilobase in human [1], one every 78 base pair (bp) in grapevine [2], and one every 43 bp in maize [3]. SNPs have proven to be useful genetic tools in many genetic studies, including molecular breeding, population genomics, genetic diversity, and cultivar identification [4].

Pine forests are the dominant ecosystem in many regions of the world and supply important industrial materials. Pines have enormous genomes that are about 10 and 40 times larger than human and poplar genomes, respectively [5]. Given current sequencing technologies, sequencing the whole genome of pine is infeasible in the near future. Expressed sequence tags (ESTs) provide an alternative approach for functional studies of the expressed genes in pine. In recent years, large numbers of pine EST sequences, especially of loblolly pine, have been deposited in public databases (www.ncbi.nlm.nih.gov/search/EST/), providing valuable resources for analyzing the nucleotide variations within expressed genes in the pine genome.

In this paper, we detected and analyzed a large number of nucleotide variations in the expressed genes of loblolly pine using existing EST resources. Our objectives were to (1) detect nucleotide variations within expressed sequences of loblolly pine, and (2) characterize these variations.

Materials and Methods

Acquisition and assembly of EST sequences

EST sequences of loblolly pine were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov/search/EST/) on May 10, 2012. Prior to assembly, we used the program SEQMAN NGEN (v. 1.2) [6] to trip adaptors, primers, and poly-A tails and to filter by sequence quality (threshold quality score = 20). The EST sequences were then assembled using GS De novo Assembler program (v. 2.7) [7] with default parameters, except for the minimum overlap length (50).

Discovery and analysis of nucleotide variations

To detect nucleotide variation in ESTs, we used the consensus EST sequences of loblolly pine as the reference sequences. Roche gsMapper (v. 2.7) [8] was used to align individual ESTs to the reference sequences with default settings, except for minimum overlap length (50) and minimum overlap identity (96%). As described in the Single Nucleotide Polymorphism Database (dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/), we included SNPs, short multi-base polymorphisms (MNPs), single- and short multi-base indels, and tandem repeat variation in our analysis. Candidate SNPs detected with gsMapper were filtered according to the following criteria: (1) at least four non-duplicated ESTs shared the polymorphism, and both forward and reverse ESTs had the change if the total depth was lower than 7, and (2) no other SNPs were detected within five bp on either side of the candidate SNP.

Results and Discussion

EST sequence assembly and detection of nucleotide variations

Altogether, 328,662 loblolly pine ESTs were downloaded from NCBI and assembled into 12,515 contigs, with 26 ESTs per contig on average. A total of 23,881 nucleotide variations, including 22,682 (94.98%) SNPs and 1,199 MNPs were detected in 3,697 of the assembled contigs, while 8,818 contigs contained no base changes. The ratio of contigs containing base variations was 29.54% (Table 1). The detected nucleotide variations and their characteristics were listed in Supplementary Table 1. Compared with MNPs, SNPs are the most common nucleotide change in many plant species, including crops like maize [9] and Brassica [10] and forest trees, like *Pinus* [9]. MNPs result in greater changes than SNPs in translated proteins, affecting function. Therefore, the predominance of SNPs may be a general phenomenon in coding genes.

As expected, the majority of single-base variations were SNPs (22,529, 99.33% of single-base changes), while indels were relatively infrequent (153, 0.67%). In contrast, the ratios of substitutions (626, 52.21%) and indels (573, 47.79%) were similar among the 1,199 multiple base variations (Table 1).

Analysis of SNPs

Of the 22,682 SNPs, 22,529 (99.33%) were single-base substitutions

Variation type	Number	Percentages	
single-nucleotide changes	22,682	94.98%	
SNPs	22,529		99.33%
indels	153		0.67%
multiple-nucleotide changes	1,199	5.02%	
MNPs	626		52.21%
indels	573		47.79%
Totals	23,881	100%	

Table 1: Nucleotide variations detected in expressed sequence tag assemblies of loblolly pine.

***Corresponding author:** Tongming Yin, Faculty of Forest Resources and Environmental Sciences, Nanjing Forest University, Nanjing 210037, China, Tel: 15261427928; E-mail: tmyin@njfu.edu.cn

Received April 17, 2013; **Accepted** June 10, 2013; **Published** June 19, 2013

Citation: Kong F, Wang X, Chen Y, Bian A, Xu J, et al. (2013) Analyzing the Nucleotide Variations within the Expressed Sequence Tags of Loblolly Pine (*Pinus taeda*). J Plant Biochem Physiol 1: 109. doi:10.4172/jpbp.1000109

Substitution Type	Number	Percentages	
Transitions	13,507	59.95%	
A/G	6,482		47.99%
C/T	7,025		52.01%
Transversions	9,022	40.5%	
A/C	2,207		24.46%
A/T	1,885		20.89%
C/G	2,727		30.23%
G/T	2,203		24.42%
Total	22,529	100%	

Table 2: Single-nucleotide substitutions detected in expressed sequence tag assemblies of loblolly pine.

involving 13,507 (59.95%) transitions and 9,022 (40.05%) transversions, while only 153 single nucleotide changes (0.67%) were indels (Table 2). Among 22,529 substitutions, about 6,364 were non-synonymous, accounting for 28.25% of the total; about 16,165 were synonymous, accounting for 71.75% of the total. There were more transitions than transversions, because base changes between two pyrimidines or two purines are biochemically easier than changes between a pyrimidine and a purine. A high frequency of transitions was also observed in other SNP discovery studies [11]. Among the SNP substitutions, C/T transitions occur most frequently. A higher rate of C/T transitions occurs in other organism [11], consistent with the experimental observation that cytosine demethylation was the most common mutational event. In the coding genes of pine, G/A and G/C substitutions were more common than G/T, A/C, and A/T transversions, although the underlying mechanism could not be explained.

We detected a total of 153 single-base indels, including 40, 43, 37, and 33 for A, C, G, and T, respectively. There was no significant indel bias for the different nucleotides.

Analysis of MNPs

Of the 1199 multiple-base changes, 626 (52.21%) were MNPs, all of which caused amino acid changes. There were also 573 (47.79%) multi-base indels, of which the frequency generally decreased with indel lengths (Figure 1), excepted for indels of 6, 9, 12, 15, and 18 bases, which were in length of integrated codons. Most multi-base indels were 2–4 bases long, accounting for 67.19% of the total. This result might imply that short indels occurred at a higher frequency than long indels, or short indels are more likely to be selected against over evolutionary time. This trend was also observed in a previous study on maize [12].

As previously reported, the general trend that longer indels are less frequent is affected by indels of codon length [11,12]. Codon-length indels of 6-18 bases occurred at higher frequencies than expected, because such indels would only cause slight changes in the open reading frames of the corresponding genes. Indels may result from errors in DNA synthesis, repair, or recombination or may be caused by insertion and excision of transposable elements, which often leave behind a characteristic DNA remnant of several bases. Notably, the frequency of eight-base indels was also higher than expected. A similar scenario was observed in maize by Bhattramakki et al. [12] and was thought to be related to sequence duplication during insertion and excision of Ac/Ds transposable elements [13].

Although the frequency of multi-base variations was very low, they could be used as molecular markers whose polymorphisms are much easier to detect than those of single-base variations. The multi-base

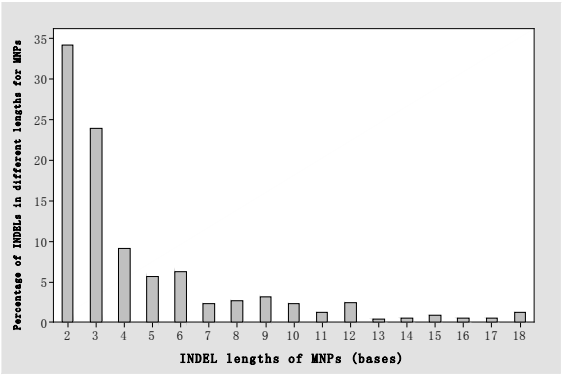

Figure 1: Frequencies of multiple base indels detected in EST assemblies of loblolly pine

changes detected in this study offer a valuable resource for developing easily-detectable markers in coding genes.

Acknowledgments

Funding for this work was provided by the Key Forestry Public Welfare Project of China (201304102), the Natural Science Foundation of China (31270661), and the Program for Innovative Research Team in University (PCSIRT) of the Educational Department and Jiangsu Province of China.

References

1. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.
2. Salmaso M, Faes G, Segala C, Stefanini M, Salakhutdinov I, et al. (2004) Genome diversity and gene haplotypes in the grapevine (Vitis vinifera L.), as revealed by single nucleotide polymorphisms. Mol Breed 14: 385-395.
3. Jones E, Chu WC, Ayele M, Ho J, Bruggeman E, et al. (2009) Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (Zea mays L.) germplasm. Mol Breed 24: 165-176.
4. Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5: 94-100.
5. Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, et al. (2009) Evolution of genome size and complexity in Pinus. PLoS One 4: e4332.
6. Galindo J, Grahame JW, Butlin RK (2010) An EST-based genome scan using 454 sequencing in the marine snail Littorina saxatilis. J Evol Biol 23: 2004-2016.
7. Zhang W, Chen J, Yang Y, Tang Y, Shang J, et al. (2011) A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PLoS One 6: e17915.
8. Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, et al. (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics 9: 312.
9. Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, et al. (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3: 19.
10. Li XJ, Zhang G, Gu AX, Xuan SX, Wang YH, et al. (2010) Detection and analysis EST-SNPs in Brassica. J Plant Genet Res 11: 772-776.
11. Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol 132: 84-91.
12. Bhattramakki D, Dolan M, Hanafey M, Wineland R, Vaske D, et al. (2002) Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Mol Biol 48: 539-547.
13. Sutton WD, Gerlach WL, Peacock WJ, Schwartz D (1984) Molecular analysis of ds controlling element mutations at the adh1 locus of maize. Science 223: 1265-1268.