A Meta-Analysis of EST-SSR Sequences in the Genomes of Pine, Poplar and Eucalyptus

Maomao Yan, Xiaogang Dai, Shuxian Li, Tongming Yin

Corresponding author email: tmyin@njfu.com.cn; Authors

Abstract Microsatellites are the kinds of sequences in the genome of living organism that have the fastest variation frequency. The variations of numbers of microsatellite repeat units in the structure gene cause frame-shift mutation of the gene, resulting in gene expressing in fully difference or expressing truncated protein. Thus, in the evolutionary process, microsatellite within the gene region would be influenced by a strong selection. In order to study the variations of microsatellite within the of gene in different tree species, we performed a meta-analysis by using SPUTNIK program to analyze the 30 000 express sequence tag (EST) sequences of pine (Pinus spp.), Poplar (Populus spp.) and eucalyptus (Eucalyptus spp.) downloaded from NCBI database.

The results showed that the proportion of EST sequences containing microsatellites was 18.7% in eucalyptus and 15.3% in poplar, whereas only 8.2% occurred in pine tree that exhibited bigger differentiation. The study found that three-base repeat type was is a major repeat type of microsatellite in the coding sequence of these three tree species. In addition to three-base repeat microsatellite, the abundance of other types of microsatellites in EST sequences of eucalyptus and poplar was decreased with increase of the repeat unit length, but in the pine tree occurring in the opposite situation. It was worth noticeable that the amount of microsatellite with fast frequency of variation of EST sequences in pine (>20 bp) was significantly less than that in eucalyptus and poplar. The study also found that the rate of repeating unit losing or gaining decreased with increases of the repeating units in the microsatellite of three tree species. In this study, we reported the comparative studies of microsatellites within the gene region in different tree species, revealing the similarities and differences in abundance and variation frequency of microsatellites in the EST sequences in pine, poplar and eucalyptus. Microsatellite sequence would have an important influence on the function of the gene containing the microsatellite. The results of this study would provide some parameters for understanding the characteristics of microsatellites with the gene region in the different species, as well as useful references for developing microsatellite markers with high polymorphism by using the EST sequence of the studying tree species.

Keywords Microsatellites; Express sequence tag (EST); Repeat unit; Pine (Pinus spp.); Poplar (Populus spp.); Eucalyptus (Eucalyptus spp.)

Background Microsatellite, also known as Simple Sequence Repeats (SSR), refers to short tandem repeating sequences of 2–6 base pairs nucleotide units (He, 1998). It is the most rapid variable DNA sequence in genome, not only with high polymorphism but also with high conservation and universality among in the relatives (Kashi and King, 2006). Therefore, microsatellite markers would be one of the most effective genetic tools for the integration of information among the genomes of closely relative species (Li et al., 2009), that has become the most widely applied molecular markers in genetic research.

Microsatellite markers have been widely used for genetic mapping and germplasm identification etc in many plants. In recent years, microsatellite markers were also rapidly developed in trees, having been widely used in the tree studies of fingerprinting, genetic map building and population genetic analyzing etc. (González-Martínez et al., 2006). In the studies of modern molecular genetics, microsatellite markers
would be an important tool in the different research fields for genotyping, genetic map building, QTL mapping, as well conservation of genetic germplasm, protection of endangered species, gene flow monitoring, genetic drift research, gene mutation and systematic analysis (Huang, 2002). The typical procedures for development of microsatellite markers would be including library construction, enrichment of microsatellite sequences, sequencing, microsatellite searching and primer designing that no doubt might be cumbersome and expensive processes limiting the development and application of microsatellite markers. Although microsatellites can also mined by using the resourced sequences in many different species, however, comparing to the model animals and plants as well as important crops, it is much more limited of genomic information in most of forest tree species.

EST sequencing is an important means of functional studies of genes. A large-scale EST sequencing of trees have been done in the forest representatives of pine, poplar and eucalyptus, there was an abundant EST sequence information of these species deposited in public databases (Sterky et al., 2004; Allona et al., 1998; Keller et al., 2009), which provides important sequence resources for developing the microsatellite marker in these species. In fact, pine, eucalyptus and poplar are known leading industrial timber species in China and even in the world, developing the resources of microsatellite markers would be an important to strengthen the genetic studies of these species.

The characteristics of microsatellite sequences are an important indicator to understand the genomic differences of different species. Microsatellite is widely distributed in the genome of eukaryote, exiting not only in the region of introns and intervals of genes, also presenting in the coding sequence of gene. Because microsatellite sequences are prone to variations, the genes containing the microsatellite are more subjected to being mutations than that of non-microsatellite.

A large number of studies have shown that microsatellite instability is related to human cancer and neurological diseases occurring (Lothe, 1997; Toth et al., 2000). In the evolutionary process, the SSR within the gene region will be influenced by the selection of convergence.

The studies on the microsatellite in the exon sequence of the poplar genome revealed that the microsatellites with three-base repeat units in the exons of poplar genome were far more than other types of microsatellites (Li et al., 2009). Comparing to microsatellites with other repeat types, the microsatellite with three-base repeat unit had minimum impacts on open reading frame of gene, enrichment of three-base repeat microsatellite showed that microsatellite in the exon region were influenced by the selection of genetic code in the evolutionary process.

Currently, the studies of microsatellite in gene region mostly focused on humans and model animals and plants, especially in human cancer research (Lothe, 1997; Brinkmann et al., 1998; Toth et al., 2000). It rarely reported on the research of microsatellite in gene regions in forest tree species. Through comparison of microsatellite in the gene region in different tree species, it not only provides important parameters for understanding the differentiation of the genomes of different tree species, but also provides a reference of bioinformatics for how to make use of these sequence resources to develop microsatellite markers with high polymorphism.

1 Results and analysis
1.1 Analysis of microsatellite abundance among the expressed sequence pine, poplar and eucalyptus

There were 2,465, 4,599 and 5,612 of EST sequences containing microsatellites with repeating units from 2 to 5 bp in length by using Sputnik software to search each 30,000 EST sequences of pine, poplar and eucalyptus, respectively. The frequency of microsatellite occurring in EST sequences was 8.2% in pine, 15.3% in poplar and 18.7% in eucalyptus respectively. The results showed that the frequency of microsatellite occurring in EST sequences was quite close between eucalyptus and poplar, but EST-SSR abundance of the pine was significantly lower than that of the poplar and eucalyptus.

Analysis on repeat unit length of EST-SSR in three
trees found that microsatellite abundances in eucalyptus and poplar were very close except the abundance of 5 base repeat unit microsatellites, but quite different in pine (Table 1). The results also showed that the frequency of microsatellite with three-base repeat unit were more than other types of microsatellites in three tested trees. It is noteworthy that the abundance of other three repeat unit microsatellites except microsatellites with three-base repeat unit in eucalyptus and poplar reduced with the increases of the length of repeat units, whereas increased along with the increase of length of repeating unit in pine. Genetic code consists of three base nucleotides. Generally speaking, gene would have higher fault-tolerant to mutations of three base insertion or deletion, long-term selection might lead to enrich the microsatellites with three-base repeat unit in the gene region, the results of this study might provide new direct evidence for this statement. We also noted that enrichment of three-base repeat unit microsatellites in pine (64.0%) were greater than that in poplar (45.2%) and eucalyptus (48.2%), the reason might be that the time of pine genome formation was far earlier than that of poplar and eucalyptus, this indicated that the pine genome might be subjected to a very strong impact in codon selection during the gene evolution.

Table 1 Comparative analysis of different types of microsatellites in expressed sequences in pine, poplar and eucalyptus

<table>
<thead>
<tr>
<th>Type of microsatellite</th>
<th>Pine</th>
<th>Poplar</th>
<th>Eucalyptus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Di-nucleotide</td>
<td>285 (10.1%)</td>
<td>1 532 (27.0%)</td>
<td>2 208 (30.1%)</td>
</tr>
<tr>
<td>Tri-nucleotide</td>
<td>1 809 (64.0%)</td>
<td>2 559 (45.2%)</td>
<td>3 539 (48.2%)</td>
</tr>
<tr>
<td>Tetra-nucleotide</td>
<td>330 (11.7%)</td>
<td>806 (14.2%)</td>
<td>1 031 (14.0%)</td>
</tr>
<tr>
<td>Penta-nucleotide</td>
<td>400 (14.2%)</td>
<td>772 (13.6%)</td>
<td>567 (7.7%)</td>
</tr>
<tr>
<td>Percentage of ESTs containing microsatellites</td>
<td>2 465 (8.2%)</td>
<td>4 599 (15.3%)</td>
<td>5 612 (18.7%)</td>
</tr>
</tbody>
</table>

Note: The percentage values in this table are the ratio of the microsatellites with different lengths of repeat units accounting for the total numbers of microsatellites detected in the tested tree species

Law of abundance of microsatellite in gene region and in different types varied with the length of repeat unit was significantly differences in pine from eucalyptus and poplar; it is yet to be identified in many tree species whether this is common differences of microsatellite in gene region among needle trees and broad-leaf trees. Although we couldn’t give a definitive conclusion to the puzzle due to EST sequences of other species at present very limited, the results of this study revealed a very interesting phenomenon.

1.2 Nucleotide composition of pine dominant repeat unit of EST-SSR in pine poplar and eucalyptus

Studies have shown that SSR site is a hotspot region of gene recombination occurring (Jeffreys et al., 1998), some repeat sequences of microsatellite such as repeat units of GT, CA, CT, and GA often direct impact on the DNA recombination through affecting the structure of DNA (Biet et al., 1999). And also variations of the numbers of repeat units of microsatellite in expressed sequences facilitate the open reading frame changing, thus affecting the gene encoded product. So, because of the influence of the selection, the differentiation of nucleotide composition in dominant repeat unit in different species will occur in the long-term evolutionary process.

Analysis on nucleotide composition of microsatellite repeat unit in expression sequence (repeat motif) in three species showed that the base composition of microsatellite with highest frequency were two base repeat unit of AG/TC in poplar and eucalyptus, while TA/AT in the pine. Whereas the three-base repeat unit of AAG/TTC occurred in the three species with the highest frequency. The composition of dominant four repeating unit was AAAG/TTTC in pine and poplar, but AGTG/TCAC in the eucalyptus. There were different base compositions of dominant penta-nucleotide repeat unit in three tree, CTGGC in pine, AAAAG/TTTTC in poplar and CAAAG in eucalyptus. So overall, the base compositions of dominant repeat units with different length had species-specific characteristics, there was no any significant regularity of differentiation in pine, poplar and eucalyptus (Table 2).

As the variations of the repeating numbers of repeat
unit of microsatellite in gene region might affect the gene expressed products, mutations in repeat unit of microsatellite might lead to gene silencing or the occurrence of selective removing. Therefore, analysis of base compositions of the dominant repeat unit could help to understand the gene of which the nucleotide composition of the repeat unit had a higher mutation receptivity.

Table 2 The nucleotide compositions and frequency of the dominant repeat units of microsatellites in pine, poplar and eucalyptus

<table>
<thead>
<tr>
<th>The type of repeats</th>
<th>The composition and proportion of nucleotides in dominant repeat units</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pinus</td>
</tr>
<tr>
<td>Di-nucleotide</td>
<td>TA/AT (38.6%)</td>
</tr>
<tr>
<td>Tri-nucleotide</td>
<td>AAG/TTC (9.8%)</td>
</tr>
<tr>
<td>Tetra-nucleotide</td>
<td>AAAG/TTTC (6.7%)</td>
</tr>
<tr>
<td>Penta-nucleotide</td>
<td>CTGCG (6.3%)</td>
</tr>
</tbody>
</table>

Note: The percentage values in this table are the ratio of the listed dominant repeat units in different lengths of repeat units.

1.3 Length variations of microsatellites with different length of repeat units in pine, poplar and eucalyptus

The length variations of microsatellites containing repeat unit of two bases, three bases, four and five bases were shown in figure for pine, poplar and eucalyptus (Figure 1). A different part of each pie chart (Figure 1) represented the repeating numbers of different length of each repeat unit, more repetitions showed greater differentiation. Overall, the longer of the length of repeat unit of microsatellite in the EST sequences was, the less variation of repetitions was in pine, eucalyptus and poplar trees. However, in three species, the rate of repeat unit losing or gaining was not significantly different in five nucleotide repeat unit microsatellites as well as in four repeat unit microsatellites. As the differentiation of the sequence length has been exhibited the rate of repeat unit losing or gaining of microsatellite, this feature was closely associated with the polymorphisms of microsatellite sites. The results of the figure showed that there was higher polymorphism in microsatellites containing short repeat unit than in microsatellites containing long repeat unit in the three species. This would be a kind of reference for developing highly polymorphic microsatellite markers.

We had the searched microsatellites divided into two categories, the first class included the SSRs with 20 bp or more in length, the second class including these SSRs with 12 bp more in and 20 bp less (Temnykh et al., 2001). The first class SSRs had a higher polymorphism than the second class SSRs, of which the law that Weber (1990) firstly discovered in experimental data of microsatellite in human had been confirmed in many organisms. The second class generated less mismatching sites in slipping chain matching because of the short fragment length of SSR, so the polymorphism of the second class SSRs was not as the first class. Mutation rate of SSR sequences with less than 12 bp fragment length was no any difference with other sequences, which would have a random variation in trend. In this study, the 33.3% of EST-SSRs in eucalyptus belonged to in the first category and 18.9% in poplar but only 8.1% in the pine. As mentioned above data, the largest proportion of Eucalyptus EST containing high polymorphic SSRs was eucalyptus, followed by poplar and pine trees. And the number of first class SSR in eucalyptus and pine both were higher than in poplar (Figure 1). Above mentioned results showed that the abundance of pine SSR in gene region was significantly lower than that in poplar and eucalyptus, while enrichment of three-base repeat microsatellites was also significantly higher than that of pine and poplar, these results might be connected to codon selection, which indicated that microsatellite in gene region of pine should be accumulated with very strong selection pressure. Microsatellite length variations also showed that the content of microsatellites with high polymorphism in the pine trees was significantly lower, which might also be due to codon selection. If the microsatellites in gene region had some strong cumulative selection pressure in pine in the evolutionary process, the content of microsatellites in gene region and enrichment of
microsatellites with three-base repeat would be made clear differentiation in pine compared to the other two species, and the proportion of microsatellites with high polymorphism would be diminished accordingly.

Figure 1 The variation of repeat numbers for different types of EST-SSRs in pine, poplar and eucalyptus
Note: Pies indicated with A, B and C corresponding to pine, poplar and eucalyptus respectively. The numbers in brackets are the repeat numbers of corresponding microsatellites. The percentage values stand for the proportion of microsatellites with certain repeat numbers in the corresponding type of SSRs. In these figures, except for the first section with gray shadow, each section corresponding to microsatellites with different number of repeat motif. In different types of microsatellites, the occurrence frequency of very long microsatellites was low, long microsatellites with frequency lower than 1% were pooled in the first section with gray shadow in each pie. The summed sections with gray shadow corresponding to the total amount of microsatellite with 20 bp or more

2 Discussions
Microsatellites widely distributed in eukaryotic genomes (Ding and Tong, 1999). The latest studies have found that microsatellites have many different functions, such as affecting gene expression, regulation and function. Compared to other sequences in the genome, microsatellites have high frequency of variation that is considered one important factor leading to generate and maintain the variation of quantitative trait in the process of genome evolution (Tautz et al., 1986; Kashi et al., 1997). While a large number of studies had found that instability of microsatellite closely related to human cancer and neurological diseases (Lothe, 1997; Toth et al., 2000), such as microsatellite instability leading to the incidence of breast cancer, colorectal cancer, stomach cancer, prostate cancer, esophageal cancer, thyroid cancer etc. (Lin and Sun, 2003, Wujing Yixueyuan Xuebao, 12(3): 231-233). Therefore, the characteristics of microsatellite sequences would be the important parameters to understand of the genome
in different species.

A large number of microsatellite analyses on bioinformatics have been reported. Representatively, Dieringer and Schlotterer (2003) found two significantly different patterns in the course of microsatellite variation through bioinformatics analysis on a large number of microsatellites in the nine species. However, microsatellite analysis of bioinformatics mainly focused on fungi, human beings and mode of plant (Brinkmann et al., 1998; Lothe, 1997; Toth et al., 2000). For forest tree species, most of microsatellite analysis were yet limited to be experimental analysis on a small number of microsatellite loci (Wyman et al., 2003), whereas a large number of microsatellite bioinformatics analysis were also limited to a single species, such as whole genome sequenced poplar (Tuskan et al., 2004; Li et al., 2009). So far, it is yet to report comparative study of the characteristics of microsatellite more than one species, which might be due to lack of the genomic sequence resources. Instead of whole genome sequencing, EST sequencing of forest tree species was much more popular (Sterky et al., 2004; Allona et al., 1998; Keller et al., 2009). A plenty of EST sequences of pine, poplar and eucalyptus in public databases would facilitate this research.

Overall, the dominant microsatellites in gene regions displayed the similar trends in the length of repeat unit and the frequency of repeat unit losing and gaining in this study, but the abundance and the number of microsatellites with high-frequency variation had a significant differentiation occurring between pine and poplar or eucalyptus compared. Pine is a kind of conifer species, while eucalyptus and poplar are kinds of the broadleaf species. Our findings did reveal whether there is the common difference between the genomes of needle and broadleaf species, it does no doubt need to be analyzed in more tree species. However, the released public databases of EST sequences of other needle and broadleaf species are still very limited and insufficient to carry out the relevant analysis of bioinformatics. In recent years, with the rapid development of next generation high-throughput sequencing technology, transcriptome sequencing has been conducted on more and more tree species, of which will be able to draw a clear answer to this question.

3 Materials and methods

3.1 The sequence resources and microsatellite sequence finding

EST sequences of Pine, poplar and eucalyptus were downloaded from NCBI database by the website of http://www.ncbi.nlm.nih.gov/dbEST/index.html. There were differences in the number of EST sequences sequenced in these species, in order to ensure the results with comparability to each species, 30,000 sequences were randomly selected to search microsatellite sequences by using the program of Sputnik developed by C. Abajian of University of Washington. The finding process followed with the default threshold, and the minimum Score was set to nine, the ranges of all microsatellites with repeating units from 2 to 5 bp in length were covered.

3.2 Analysis on microsatellite length variation and nucleotide composition of dominant repeating units

We adopted the EXCEL tool to classify the nucleotide compositions of repeat units of microsatellites, in order to find the highest proportion of the repeat unit with the same base composition in different types of microsatellites and figure out the dominant base composition in different types of microsatellite units through the corresponding repeat unit. Of microsatellite length variation was carried out by the mapping functions of the EXCEL. Pie charts were drawn based on microsatellites different length repeat unit of which each sector corresponding to the different microsatellite in length (the microsatellite with the 1% frequency or less was merged, each sector corresponding to the length of the microsatellite and the proportion of the microsatellite were marked with tie line in the corresponding sector), the number of sectors stood for the variation of microsatellite length, the more the sectors had, the faster the rate of repeat unit losing and gaining in the corresponding type of microsatellites, thus the corresponding type of microsatellites generally has higher polymorphism.

Author’s contributions

MMY, DXG and SXL are the persons who carried out this study. TMY conceived the project and designed the analysis procedures as well as wrote and revised manuscript. All authors had read and agreed the final text.
Acknowledgements

This research was jointly sponsored by National Foundation of Natural Sciences (31070543, 30971609) and Key Project of Jiangsu Foundation of Natural Sciences for High Education (10KJA180018). Authors appreciated two anonymous reviewers with their critical comments. In this paper we mentioned some chemical and reagent suppliers and sequencing service providers, that doesn’t mean we would like to recommend or endorse the production of theirs.

References


Morgante M., Hanafey M., and Powell W., 2002, Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes, Nature Genetics, 30(2): 194-200 http://dx.doi.org/10.1038/ng822 PMid:11799393


Tautz D., Trick M., and Dover G.A., 1986, Cryptic Simplicity in DNA is a major Source of genetic variation, Nature, 322(6080): 652-656 http://dx.doi.org/10.1038/322652a0 PMid:3748144


Toth G., Gaspary Z., and Jurka J., 2000, Microsatellites in different eukaryotic genomes: Survey and analysis, Genome Research, 10: 967-981 http://dx.doi.org/10.1011/gr.184001 PMid:11843586 PMCid:310977


Kashi Y., and King D.G., 2006, Simple sequence repeats as advantageous markers in evolution, Trends in Genetics, 22(5): 253-259http://dx.doi.org/10.1016/j.tig.2006.03.005 PMid:1657018