J. Microbiol. Biotechnol. (2010), **20**(3), 474–479
doi: 10.4014/jmb.0909.09014
First published online 19 December 2009

JOURNAL OF MICROBIOLOGY AND BIOTECHNOLOGY

Characteristics of Microsatellites in the Transcript Sequences of the *Laccaria bicolor* Genome

Li, Shuxian[1†], Xinye Zhang[2†], and Tongming Yin[1*]

[1]*The Key Laboratory of Forest Genetics and Biotechnology and Jiangsu Key Laboratory for Poplar Germplasm Enhancement and Variety Improvement, Nanjing Forestry University, Nanjing 210037, China*
[2]*Hubei Forestry Academy, Wuhan 430079, China*

Received: September 10, 2009 / Revised: October 15, 2009 / Accepted: October 23, 2009

In this paper, we analyzed the microsatellites in the transcript sequences of the whole *Laccaria bicolor* genome. Our results revealed that, apart from the triplet repeats, length diversification and richness of the detected microsatellites positively correlated with their repeat motif lengths, which were distinct from the variation trends observed for the transcriptional microsatellites in the genome of higher plants. We also compared the microsatellites detected in the genic regions and in the nongenic regions of the *L. bicolor* genome. Subsequently, SSR primers were designed for the transcriptional microsatellites in the *L. bicolor* genome. These SSR primers provide desirable genetic resources to the ectomycorrhizae community, and this study provides deep insight into the characteristics of the microsatellite sequences in the *L. bicolor* genome.

Keywords: *L. bicolor*, microsatellites, SSR primer development, transcript sequences

Mycorrhizal symbioses, the union of roots and soil fungi, are universal in terrestrial ecosystems. Ectomycorrhizae has long being found to play a significant role in many aspects of plant growth and physiological procedures, such as nutrition uptake, endogenous hormone content, water utilization, disease and hardiness resistance, biomass production, and so on. It has been widely used in agricultural and forestry plantations to increase plant productivity, disease resistance, and hardiness endurance [8]. However, the genetic mechanisms for ectomycorrhizae interacting with plant roots remain largely unknown. Supported by the Department of Energy (DOE), U.S.A., the genome of the ectomycorrhizal basidiomycete *L. bicolor* has been completely sequenced and publicly released [10]. This accomplishment of the *L. bicolor* genome will facilitate identification of the primary factors that regulate symbiotic development and metabolic activity, and therefore open the door for our better understanding of the role of ectomycorrhizae in plant development and physiology [10]. The availability of the *L. bicolor* genome sequence leads to its emergence as one of the model fungi in the genetics and functional genomics studies in ectomycorrhizae. However, the applicability of the *L. bicolor* genome sequence to studies of alternate *L. bicolor* genotypes and related species remains undetermined. We need to establish a series of platforms and genetic tools to extend the genome information in the varied fields of genetic studies on *L. bicolor*. The genetic map is one of the essential platforms to identify and locate genetic loci underlying important economic and ecological traits [17]. Through the joint effort of INRA–Nancy University and Oak Ridge National Lab. (ORNL), a moderate density genetic map was built for *L. bicolor* [5]. This map was mainly constructed by amplified fragment length polymorphism (AFLP) markers, with integration of a few microsatellites. Although anonymous markers, like RAPDs and AFLPs, are useful for fast construction of a genetic map, they are not as efficient in comparative mapping and in validating the genetic results derived from different mapping studies. By contrast, microsatellites or simple sequence repeats (SSRs) are one of the most efficient tools to unite the genetic studies on related species and communicate them with the genome sequence information. By aligning the priming sequences of SSR primers to the genome sequence, we can infer their physical positions in the genome, and thereafter, to define the genome sequences and the gene contents in a particular genetic interval. Combining QTL analysis and SSR priming sequences alignment, we can generate a list of candidate genes in the QTL interval, and thereafter to facilitate gene identification [17]. Microsatellites have been proven to be the most efficient genetic tools for

*Corresponding author
Phone: +01186-25-85428165; Fax: +01186-25-85428165;
E-mail: tmyin@njfu.com.cn

†Li and Zhang contributed equally to this paper.

communicating genetic information among genomes of related species [6]. Microsatellite loci possess high mutability, which lends high polymorphism to SSR markers. Meanwhile, the priming sequences of microsatellites are normally conserved within species, and many of them even can be transferable among *taxa* of *genus* [17]. Using SSRs, it is feasible to build a platform to study different individuals or related species as a macrogenetic system and to validate genetic findings from different studies. Although microsatellites are desirable molecular markers, conventional procedures for microsatellite primers development are time-consuming and expensive [2]. The availability of the genome sequences of *L. bicolor* provides enormous sequence resources for SSR primers development at low cost. According to the study in other organisms, SSR primers developed from genic sequences possess higher transferability than that designed from the nongenic sequences [17], and these microsatellites are relating to functional genes directly. In this paper, our objectives were to analyze the microsatellite sequences in all of the transcripts of the whole *L. bicolor* genome, to learn the characteristic and constituents of the microsatellite repeats, to compare the microsatellite signatures in transcript sequences with that in random genome sequences of *L. bicolor* and that in transcripts of the poplar genome, and subsequently, to develop SSR primers for all the detected microsatellites in the transcripts of the *L. bicolor* genome, thus providing a genetic resource of great potential value to the ectomycorrhizae community.

MATERIALS AND METHODS

Sputnik program coding with C language was used to search DNA sequence files in Fasta format for microsatellites (C. Abajian, University of Washington). The program uses a recursive algorithm to search for repeated units of nucleotides of length between 2 and 5. The minimum score for repeats searching is set at 9. The transcript and genome sequences were obtained from the *L. bicolor* genome browser (http://genomeportal.jgi-psf.org/Lacbi1/Lacbi1.home.html). Since gaps and the ambiguous sequences (*i.e.*, N readings) contain no information in the analysis, they were excluded in the related statistics in this study. Information of polar genic SSRs was collected from the SSR primer development paper by Yin *et al.* [18]. All the above sequences were analyzed under the same criteria and by the same program. Thus, comparison of the microsatellites in the three sequence resources was only determined by the sequences themselves. SSR primers were subsequently designed by Primer 3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). For each SSR primer pair, both the left and right primers were supposed to be in sites priming with transcript sequences.

RESULTS

Microsatellites in the Transcript Sequences of the *L. bicolor* Genome

The sequencing project of the *L. bicolor* genome revealed that its genome contained approximately 65-megabase oligonucleotides and 20,000 predicted protein-encoding genes [10]. Totally, we analyzed the sequences of 20,614 transcripts. These transcripts cover a physical length of 23,406,136 bp. Among them, 6,024 bp are ambiguous readings (N). Thus, unambiguous A,T,G,C readings in transcript sequences of the *L. bicolor* genome are 23,400,112 bp, whereas the total length of unambiguous readings of the whole *L. bicolor* genome are 58,683,470 bp. Thus, unambiguous transcript sequences account for about 39.9% of the total unambiguous genome sequences. With the Sputnik program, we detected 3,233 microsatellites on 2,362 genes. It estimated that about 11.5% of genes in the *L. bicolor* genome contain one or more microsatellites, based on our search criteria. On average, microsatellites occur at every 7,242 bp in the transcript sequences of the *L. bicolor* genome. The repeat motif of these microsatellites includes di-, tri-, tetra-, and pentanucleotides. Microsatellites are ordered as trinucleotide repeat microsatellites (75.1%)> pentanucleotide repeat microsatellites (12.2%)>tetranucleotide repeat microsatellites (8.2%)>dinucleotide repeat microsatellites (4.5%), according to their abundance (Fig. 1A). Trinucleotide repeat microsatellites accounted for the majority of the detected microsatellites in the transcripts. The richness of the other three classes of microsatellites was found to positively correlate with their repeat motif lengths.

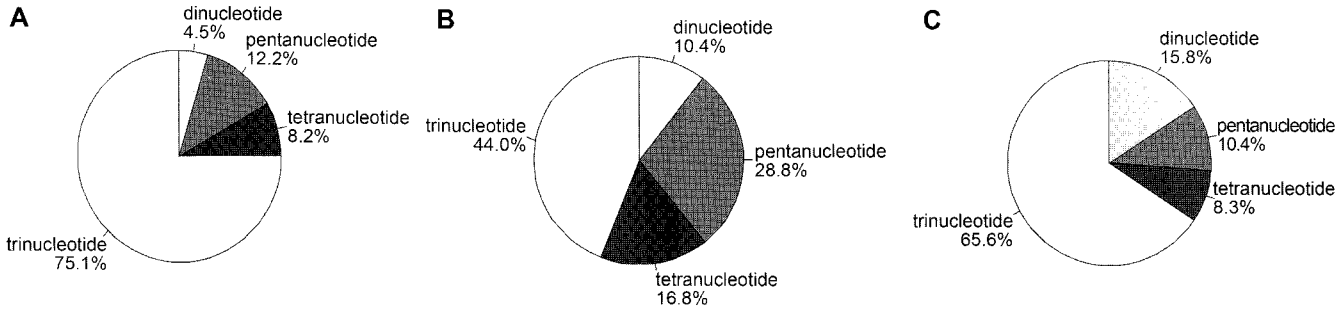

Fig. 1. The proportions of microsatellites with different repeat motif lengths in the transcript sequences of the *Laccaria bicolor* genome (**A**) in the random genome sequences of *Laccaria bicolor* (**B**), and in the transcript sequences of poplar genome.

In the *L. bicolor* genome, the length of the microsatellite sequences diverges greatly, ranging from 12 to 118 bp. The average length is 15 bp. Microsatellites ranging from 12 to 15 bp account for 73.9% of the total detected microsatellites, with each of them accounting for more than 10% in richness, whereas in microsatellites in other lengths, each of them accounts for less than 5%. The microsatellite lengths distribution demonstrates an overabundance of short-length microsatellites. By contrast, long microsatellites are scarce, and those greater than 20 bp only account for 9.9% of the total microsatellites. The length diversification of microsatellites reflects the activity of microsatellites gain/loss repeat motifs. It is a feature related to the polymorphisms of the microsatellites loci. We subsequently explored the lengths diversification of the microsatellites with different repeat motif lengths (Fig. 2), and found that pentanucleotide repeat microsatellites vary the most in their lengths. On the contrary, the lengths of dinucleotide repeat microsatellites vary the least. The length diversification of trinucleotide repeat microsatellites is similar to that of the tetranucleotide repeat microsatellites, with the former slightly higher than the later. In general, microsatellites with odd repeat motif lengths (penta- and tri-) vary more in their lengths than those with even repeat motif lengths (di- and tri-). Apart from the trinucleotide repeat microsatellites, a positive correlation was observed between microsatellite lengths diversification and their repeat motif lengths.

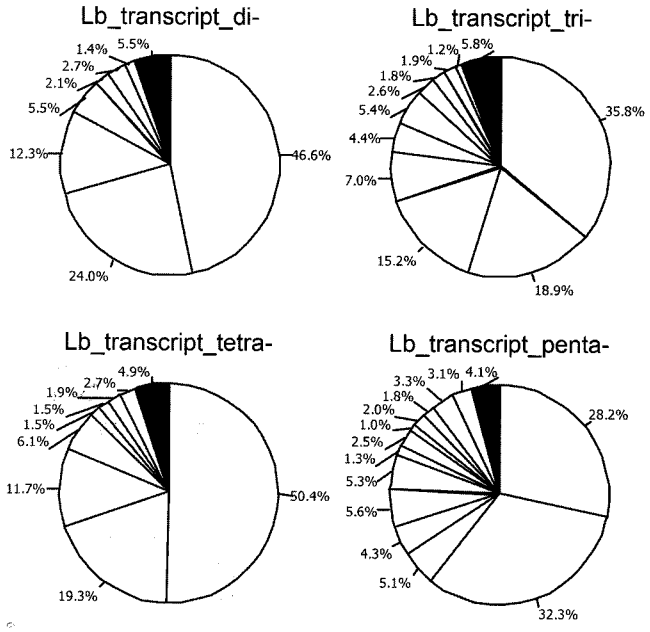

Fig. 2. Length diversification of the di-, tri-, tetra-, and penta-microsatellites in the transcript sequences of the *L. bicolor* genome.

In each pie chart, each white slice corresponds to microsatellites with the same length, and long microsatellites with percentage less than <1% are combined and shown in the black slice. The slice sizes in each pie chart are scaled according to the percentage of microsatellites in different lengths. In these pie charts, the more slices there are, the more divergence there is in the lengths of the corresponding type of microsatellites.

Comparison of Microsatellites in the Random Genome Sequences vs. in Transcript Sequences of *L. bicolor*

To compare microsatellites in the transcript sequences and in other regions of the *L. bicolor* genome, we subsequently analyzed 645 assembled sequence scaffolds that cover 32,547,773 bp in length. Among these sequences, 4,710,461 bp are captured gaps or ambiguous readings (N), and the unambiguous readings are 27,837,312 bp, which account for 47.4% of the unambiguous readings of the total genome. Under the same criteria, we detected 6,088 microsatellites within these sequences. On average, microsatellites occur at every 4,572 bp in the random genome sequences, which are about 1.58-folds higher than the microsatellite frequency observed in transcript sequences. Apart from trinucleotide repeat microsatellites, the richness of microsatellites with different motif lengths is in the same order as revealed for that in transcript sequences (Fig. 1B). Trinucleotide repeat microsatellites are significantly lower compared with their percentage in transcript sequences (44.0% vs. 75.1%). Since transcript sequences account for about 39.9% of the genome sequences, and we know the percentage of each class of microsatellites in transcript and random genome sequences, we can deduct their percentages in the nongenic sequences of the *L. bicolor* genome. The calculation demonstrated that the richness of microsatellites in the nongenic regions is ordered as pentanucleotide repeat microsatellites (39.8%)>trinucleotide repeat microsatellites (23.4%)>tetranucleotide repeat microsatellites (22.5%)> dinucleotide repeat microsatellites (14.3%). The percentages of trinucleotide repeat microsatellites and tetranucleotide repeat microsatellites are close to each other in the nongenic regions. Apart from trinucleotide repeat microsatellites, the richness of the other three classes of microsatellites positively correlates with their repeat motif lengths. This trend is the same in the nongenic regions as in the genic regions of the *L. bicolor* genome.

The length diversification of microsatellites was relatively faster in random genome sequences than in transcripts sequences. Within the investigated genome sequences, the lengths of microsatellite sequences range from 12 to 162 bp (vs.11–118 bp in transcript sequences). The average length is 16 bp (vs.15 bp in transcript sequences). Microsatellites ranging from 12 to 15 bp account for 54.3% of the total detected microsatellites (vs.73.9% in transcript sequences). The lengths diversification rate of microsatellites with different repeat motif lengths (Fig. 3) demonstrates that dinucleotide repeat microsatellites are similar to the pentanucleotide microsatellites, and trinucleotide repeat microsatellites are similar to the tetranucleotide microsatellites, with the former two classes being slightly higher than the latter two classes. The microsatellite lengths diversification

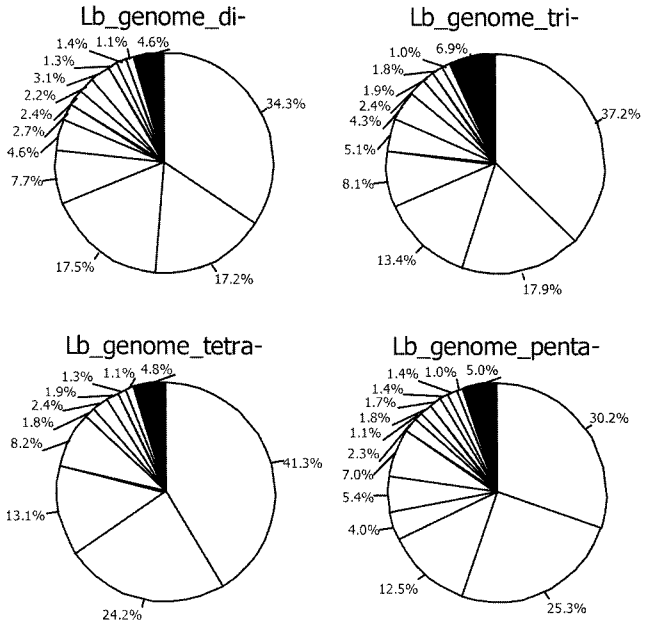

Fig. 3. Length diversification of the di-, tri-, tetra-, and penta-microsatellites in the random genome sequences of *Laccaria bicolor*.
Information about the slices in each pie chart is the same as that of Fig. 2.

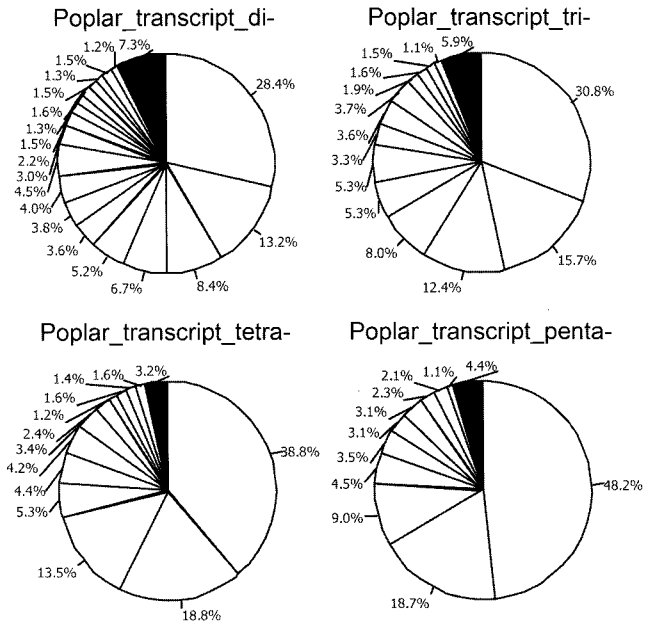

Fig. 4. Length diversification of the di-, tri-, tetra-, and penta-microsatellites in the transcript sequences of the poplar genome.
Information about the slices in each pie chart is the same as that of Fig. 2.

pattern was different from that revealed in the transcript sequences, and no obvious trend was observed between length diversification rate and the repeat motif length.

Comparison of Microsatellites in Transcript Sequences in *L. bicolor* vs. in the Poplar Genome

We used poplar as the representative of higher plants. The criteria and programs for searching and analyzing microsatellites in this study were exactly the same as that used in analyzing the poplar genome by Yin *et al.* [18]. Hereby, we particularly collected information of microsatellites in transcript sequences of the poplar genome and used them in comparison with that of the *L. bicolor* genome. In poplar, 5,989 microsatellites were found in transcript sequences (Vista models). On average, microsatellites occur at every 5,878 bp in the transcript sequences of the poplar genome (vs. 7,242 bp in transcripts of *L. bicolor* genome), which is 19% less frequent than in transcript sequences of the *L. bicolor* genome. Richness analysis on microsatellites with different motif lengths revealed that trinucleotide repeat microsatellites were also the major type of microsatellites in the transcript sequences of the poplar genome (65.6%). However, no obvious trend was observed between the frequencies of the other three classes of microsatellites and their repeat motif lengths (Fig. 1C).

The length diversification of microsatellites in the transcripts sequences was similar in poplars as in *L. bicolor*. Microsatellites range from 12 to 102 bp (vs. 12–118 bp in transcript sequences in *L. bicolor*) and the average length is 14.8 bp (vs. 15 bp in transcript sequences in *L. bicolor*). The length diversification of microsatellites with different repeat motif lengths demonstrated that dinucleotide repeat microsatellites vary the most in their length (Fig. 4). In general, the length diversification of microsatellite sequences negatively correlates with their repeat motif lengths in the transcript sequences of the poplar genome, which is opposite to the trends observed in transcripts of the *L. bicolor* genome.

SSR Primers Design for Microsatellites in Transcript Sequences of the *L. bicolor* Genome

A total of 2,985 SSR primer pairs were designed from the transcript sequences of the *L. bicolor* genome. The complete list of the designed SSR primers is available at ftp://202.119.210.12 (user: toyi;pin: 666666). The principal metrics in this table include a description of the gene name, microsatellite repeat motif length, repeat sequences, sequences and melting temperature (Tm) of each primer, GC content of the primer sequences, and expected PCR product sizes in the sequenced monokaryotic genome of *L. bicolor* "S238N-H82".

DISCUSSION

Microsatellites are simple sequence repeats, ubiquitously interspersed in eukaryotic genomes [13]. Because of their high mutability, microsatellites are thought to play a significant role in genome evolution by creating and

maintaining quantitative genetic variation [4, 14]. Although microsatellites are most commonly found in positions relaxing from genetic selection, recent studies have attributed various functional roles to microsatellites. Microsatellites are believed to be involved in gene expression, regulation, and function [3]. There is direct evidence that microsatellites can function as transcriptional activating elements [7] and there is evidence suggesting that even microsatellites in noncoding regions may also be of functional significance [4]. Thus, the microsatellite content is an important signature of the genome of a focal organism. However, the output of microsatellite sequences dramatically depends on the study criteria and the searching engines employed. To make the results comparable, sequences need to be analyzed under the same situation. In this study, we analyzed three sequence resources under the same criteria and by the same program to make sure our comparisons would not be biased. Investigation of the microsatellites content the within *L. bicolor* genome revealed that, apart from trinucleotide repeat microsatellites, the richness of the other three classes of microsatellites is positively correlated with their repeat motif lengths both in the genic and nongenic regions. The contents of trinucleotide repeat microsatellites were found to vary significantly in the genic and nongenic regions, and they account for the majority of microsatellites in transcript sequences (75.1%). By contrast, they only account for 23.4% of microsatellites in the nongenic regions of the *L. bicolor* genome. When compared with microsatellites in transcript sequences in the poplar genome, we also observed that trinucleotide repeat microsatellites accounted for the majority of the detected microsatellites (65.6%), but no obvious trend was detected for the microsatellites contents and their repeat motif lengths in the poplar genome. Although richness trends were not consistent in the two species, trinucleotide repeat microsatellites were found to account significantly high proportions of the transcriptional microsatellites in both species. The same scenario is also observed in many other species [15] and this subset of microsatellites is of great interest in the human genome because of the role it plays in many human neurodegenerative disorders [11] and in some human cancers [9]. The alteration responsible for these genetic diseases is the expansion of trinucleotide repeats [15]. The overabundance of trinucleotide repeat microsatellites is supposed to be driven by the genetic code selection, since exons seemed to tolerate only trinucleotide repeats among the microsatellites we detected. It is also noteworthy that microsatellites occured at a lower frequency in transcript sequences than in the nongenic regions of the *L. bicolor* genome. It has been shown that microsatellites in genic regions are less abundant than in nongenic regions, and it can be explained on the basis of differential selection [1]. Our study has confirmed this observation. Genes in fungi are supposed to have higher mutation rates than those in higher plants. Under the same searching criteria, the microsatellites frequency in transcript sequences is higher in the *L. bicolor* genome than in the poplar genome, which would cause higher mutability of genes in the *L. bicolor* genome than in the poplar genome.

The high polymorphism of microsatellites is due to their high frequency of variation in the number of repeat motifs in different individuals or species. This kind of polymorphism is easily detectable *via* the polymerase chain reaction (PCR) using specific primers on the flanking regions of the repeated sequence [16]. The high mutation rates of microsatellite sequences can be explained most frequently by slippage during DNA replication on a single DNA strand [12]. Therefore, the length diversification of microsatellites reflects the activity of microsatellites gain/loss repeat motifs. It is a feature related to the polymorphisms of the microsatellites loci. In poplar, the length diversification of microsatellites in transcript sequences was found to negatively correlate with the lengths of the repeat motifs, which suggested that the shorter repeat motifs gained/lost repeats much faster than the longer ones, and the microsatellites with a longer repeat motif would be more stable. By contrast, an opposite trend was observed with te length diversification and repeat motif lengths of microsatellites in transcript sequences of the *L. bicolor* genome (apart from the trinucleotide repeat microsatellites), and length diversification is found to positively correlate with the repeat motif lengths of microsatellites in its transcript sequences in *L. bicolor*. Thus, in transcript sequences of *L. bicolor*, pentanucleotide repeat microsatellites have the highest mutation rates, and the dinucleotide repeat microsatellites are supposed to mutate at the slowest rate. In *L. bicolor*, the lengths of microsatellites in random genome sequences diverse faster than that in transcript sequences, but no obvious trend was observed with the length diversification rate and repeat motif lengths of microsatellites in random genome sequences of *L. bicolor*.

From the above discussion, microsatellites in transcripts sequences are found to be less abundant and less divergent in their lengths than those detected in the random genome sequences in *L. bicolor*, which suggests that the former are under stronger convergent selection and would be less variable than the latter ones. In comparison with higher plant, such as poplar, microsatellites in transcripts sequences diverse at a similar rate in their lengths. However, the correlation trends observed in poplar and in *L. bicolor* were opposite. Thus, apart from the trinucleotide repeat microsatellites, transcriptional microsatellites with longer repeat motifs are supposed to be more polymorphic in *L. bicolor*, and by contrast, they are expected to be less polymorphic in poplar. Although our analysis cannot measure microsatellite polymorphism *per se*, the lengths of microsatellites and their diversification rates provide clues to select high polymorphic microsatellites loci in genetic studies of *L.*

bicolor. A large number of SSR primer pairs were also developed in this study. These primer resources provide desirable genetic tools to extend the genome sequence of *L. bicolor* to the genetic studies in alternative *L. bicolor* genotypes and in the related mycorrhizal fungi, which will facilitate our understanding of the genetic mechanism underlying the role of ectomycorrhizae in plant development and physiology.

Acknowledgments

Funding for this research was provided by the Natural Science Foundation of China (30971609) and National Forestry Nonprofit project of China (200904002).

REFERENCES

1. Hancock, J. 1995. The contribution of slippage-like processes to genome evolution. *J. Mol. Evol.* **41:** 1038–1047.
2. He, P. 1998. Abundance, polymorphism and applications of microsatellite in Eukaryote. *Hered. China* **20:** 42–47.
3. Jewell, E., A. Robinson, D. Savage, T. Erwin, C. G. Love, G. A. C. Lim, *et al.* 2006. SSR primer and SSR taxonomy tree: Biome SSR discovery. *Nucl. Acids Res.* **34:** 656–659.
4. Kashi, Y., D. King, and M. Soller. 1997. Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet.* **13:** 74–78.
5. Labbé, J., X. Zhang, T. Yin, J. Schmutz, J. Grimwood, F. Martin, G. A. Tuskan, and F. Le Tacon. 2008. A genetic linkage map for the ectomycorrhizal fungus *L. bicolor* and its alignment to the whole-genome sequence assemblies. *New Phytol.* **180:** 316–328.
6. Li, S. and T. Yin. 2007. Map and analysis of microsatellites in genome of *Populus*: The first sequenced perennial plant. *Sci. China C Life Sci.* **50:** 690–699.
7. Li, Y. C., A. B Korol, T. Fahima, A. Beiles, and E. Nevo. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: A review. *Mol. Ecol.* **11:** 2453–2465.
8. Liu, H. N. and J. Q. Zhu. 2001. Advances on VA-mycorrhizal research in horticultural plant. *J. Hubei Agric. College* **21:** 274–278.
9. Lothe, R. 1997. Microsatellite instability in human solid tumors. *Mol. Med. Today* **3:** 61–68.
10. Martin, F., A. Aerts, D. Ahrén, A. Brun, E. G. J. Danchin, F. Duchaussoy, *et al.* 2008. The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* **452:** 88–93.
11. Reddy, P. and D. Housman. 1997. The complex pathology of trinucleotide repeats. *Curr. Opin. Cell Biol.* **9:** 364–372.
12. Schlotterer, C. and D. Tautz. 1992. Slippage synthesis of simple sequence DNA. *Nucl. Acids Res.* **20:** 211–215.
13. Tautz, D. and M. Renz. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucl. Acids Res.* **12:** 4127–4137.
14. Tautz, D., M. Trick, and G. A. Dover. 1986. Cryptic simplicity in DNA is a major source of genetic variation. *Nature* **322:** 652–656.
15. Toth, G., Z. Gaspari, and J. Jurka. 2000. Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res.* **10:** 967–981.
16. Weber, J. L. and P. E. May. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44:** 388–396.
17. Yin, T., S. P. Difazio, L. E. Gunter, X. Zhang, M. M. Sewell, S. A. Woolbright, *et al.* 2008. Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. *Genome Res.* **18:** 422–430.
18. Yin, T. M., X. Y. Zhang, L. E. Gunter, S. X. Li, S. D. Wullschleger, M. R. Huang, and G. A. Tuskan. 2009. Microsatellite primers resource developed from the mapped sequence scaffolds of Nisqually-1 genome. *New Phytol.* **181:** 498–503.