Front. For. China 2008, 3(3): 352–356
DOI 10.1007/s11461-008-0055-2

RESEARCH ARTICLE

Wei LI, Hui LI, Xiaoyang CHEN, Harry WU

Single nucleotide polymorphism discovery of *Pinus radiata* with chromosome walking PCR method

Abstract In this paper, the basic principle of chromosome walking is presented and we used an *actin* gene of radiata pine (*Pinus radiata*) as an example to conduct upstream and downstream chromosome walking for EST sequences. The full genomic sequence (2154 bp) of the *actin* gene, including promoters 5′ UTR, CDS and 3′ UTR, was identified by chromosome walking. PCR amplification and DNA band sequencing from 200 unrelated radiata pine trees revealed a total of 21 SNPs for the *actin* gene, three in the promoter region, 15 in CDS and 4 in 3′ UTR. The results of this experiment provide a technical framework for SNPs discovery in none coding regions of candidate genes.

Keywords radiata pine, single nucleotide polymorphism (SNP), chromosome walking, none coding region

1 Introduction

Single nucleotide polymorphism (SNP) is a polymorphism at the nucleotide level, caused by genomic DNA mutations, including transition, transversion and nucleotide insertion and deletion. Generally, locus is viewed as SNP when it exists in at least two variants and the allele frequency of the most common variant is $<99\%$ (Landegren et al., 1998). SNPs have characteristics of occurring in large amounts, extensive distribution, high heritability and easy classification of its genotype. This has made SNPs the preferred third generation of molecular markers, after restriction fragment length polymorphism (RFLP) and simple sequence repeat (SSR) for the genomic study in humans, plants and animals (Gupta et al., 2001). With the completion of genomic sequencing of higher plants such as *Arabidopsis thaliana*, *Oryza sativa* and *Populus tremula*, it was observed that variations among plants in performance, susceptibility to diseases and resistance to environmental stresses were mostly correlated with SNP (David et al., 1998; Goff et al., 2002; Tuskan et al., 2006). Therefore, SNP discovery among the candidate genes and expressed sequence tags (ESTs) and their association with variation of biological traits have become a major research area in recent years.

The common procedure for SNP detection is to select candidate genes or ESTs first. For novel genes, EST is usually sequenced from screening of microarray, differential display PCR technology (DDRT) experiments and primers are designed accordingly. The targeted DNA fragments are cloned and sequenced and then applied to discover an SNP using a small sample of a population and detection software (Feltus et al., 2006). However, for the non-coding regions of the candidate genes such as the promoter regions, 5′ UTR and 3′ UTR, their cDNA sequences are not available from microarray or DDRT experiments. This makes SNP detection in these regions difficult.

Chromosome walking can overcome this difficulty by cloning the upstream or downstream DNA from the known sequence (such as EST) of novel genes, through the construction of gene specific primers and a chromosome walking library. Chromosome walking provides a new technology for SNP discovery in non-coding regions of the candidate genes. We have used the EST sequence of *actin* gene from radiata pine as an example to obtain a full genomic sequence, including promoter, 5′ UTR, intron, exon and 3′ UTR based on chromosome walking technology. After this, we designed SNP primers and conducted a genomic PCR for SNP discovery. A total of 21 SNPs were

Translated from *Acta Botanica Boreali-Occidentalia Sinica*, 2007, 27 (8): 1571–1576 [译自: 西北植物学报]

Wei LI
College of Biological Sciences and Biotechnology, Beijing Forestry University, Beijing 100083, China;
CSIRO Forest Bioscience, Canberra 2600, Australia

Hui LI
College of Biological Sciences and Biotechnology, Beijing Forestry University, Beijing 100083, China

Xiaoyang CHEN (✉)
College of Forestry, South China Agricultural University, Guangzhou 200064, China
E-mail: xychen@scau.edu.cn

Harry WU
CSIRO Forest Bioscience, Canberra 2600, Australia

detected in the full sequence of *actin* gene from radiata pine. Our objective was to demonstrate the application of chromosome walking on SNP discovery in non-coding regions.

2 Materials and methods

2.1 Materials

The EST sequence of the radiata pine *actin* gene was derived from a microarray experiment in CSIRO, Australia. A collection of 200 unrelated, 30-year-old radiata pine trees (*Pinus radiata*) was used for SNP discovery.

2.2 Methods

DNA extraction: plant DNA used for the construction of the chromosome walking library was isolated from needles of radiata pine, using a modification of the CTAB method described by Doyle et al. (1990). Template DNA for SNP detection was isolated from an amount of evenly mixed needles of 200 unrelated radiata pine trees with the same method described earlier (Sanger et al., 1997).

Chromosome walking library construction: four kinds of restriction endonucleases of *Dra*I, *Pvu*II, *Ssp*I and *Eco*RV were selected for radiata pine genomic DNA digestion. The volume for DNA digestion was 100 μL with DNA quality of 10 μg. The digested DNA was purified by a mixture of phenol, chloroform and isoamyl alcohol in a ratio of 25:24:1. An adaptor was then added and ligated to the digested DNA according to the protocol by Doyle et al. (1990).

Primer design: primary and secondary primers were designed based on the *actin* gene EST sequence of radiata pine with primer3 software (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). The length of the primer was 25–28 n, the GC content 40%–60% and the Tm value 67–72°C. The adaptor primer 1 (AP1) is: 5′-GTAATACGACTCACTATAGGGC-3′; adaptor primer 2 (AP2) is: 5′-ACTATAGGGCACGCGTGGT-3′; gene-specific primer for upstream GSP1 is: 5′-GTCACAAGGGGTCATAACAACCAGCAG-3′ and GSP2 is: 5′-CCTTACCAATTCCGACCTAAACGTCCA-3′; gene specific primer for downstream GSP1 is: 5′-TGAATCCCAAGGCAAACAGAGAGAAGA-3′; and GSP2 is: 5′-CAATGTGCCTGCTATGTATGTTGCCATT-3′.

Polymerase chain reaction (PCR): AP1 and GSP1 were used as primers for primary PCR on four constructed chromosome walking libraries. The PCR reaction volume was diluted 50 times after the PCR was finished. The primary PCR products were used as the template for the secondary PCR with the primer AP2 and GSP2. For genomic PCR, the DNAs, isolated from the evenly mixed needles, were used as a template. PCR reactions were carried out in a 25-μL volume containing 200 μmol/L of each dNTP, 1 μmol/L of each oligonucleotide primer, 2.5 U Taq polymerase per 100 μL and 1 μL template DNA. The following schedule was used for chromosome walking PCR: 1 cycle at 94°C for 1 min, 39 cycles for primary PCR and 25 cycles for secondary PCR at 94°C for 30 s, 68°C for 30 s, 72°C for 2.5 min, 1 cycle at 72°C for 10 min and finally at 4°C. Genomic PCR adopted the following schedule: 1 cycle at 94°C for 1 min, 25 cycles of 94°C for 30 s, 60°C for 30 s, 72°C for 2.5 min and 1 cycle at 72°C for 10 min and finally at 4°C.

DNA sequencing: DNA sequencing was carried out according to the Chain Termination Method (Kumar et al., 2004).

SNPs detection: Mega3.1 in Cluster and DnaSP software was used for SNP detection (Chasman et al., 2001; Rozas et al., 2003).

3 Results and analysis

3.1 Chromosome walking

From the microarray experiment, the size of the selected EST sequence of the radiata pine *actin* gene was 391 bp. After a nucleotide blast in the GenBank, the most similar hit was the *actin* gene cDNA from *Picea rubens* with a similarity of 97%. However, the full cDNA sequence of the *Picea rubens actin* gene was 1694 bp. By comparison, the EST sequence of the *actin* gene from radiata pine was presumed to be only a small part relative to the full sequence. A full sequence is required before detecting SNPs in the promoter or other unknown regions of the radiata pine *actin* gene. To obtain the full genomic sequence, chromosome walking primers were designed for upstream and downstream walking based on the known EST sequence, as shown in Fig. 1.

Primary and secondary PCR results are shown in Figs. 2 and 3 for the four restriction endonucleases used for library construction. Figure 2 shows the primary PCR results. Lanes 1 to 4 were for upstream of the four libraries constructed by restriction endonucleases of *Dra*I, *Pvu*II, *Ssp*I and *Eco*RV respectively and lanes 5 to 8 were for downstream of the same four libraries. Figure 3 shows the secondary PCR results in which lanes 1 to 4 were for upstream and lanes 5 to 8 for downstream. The secondary PCR produced clear bands in all the constructed libraries, but the size of the PCR products was quite different between the libraries. We selected the largest bands (lane 3) for upstream and lane 7 for downstream for sequencing. The promoter region and start code were predicted by an NCBI nucleotide blast.

The full length of genomic sequence from the radiata pine *actin* gene was found to be 2154 bp, by connecting the sequences from upstream and downstream genomic

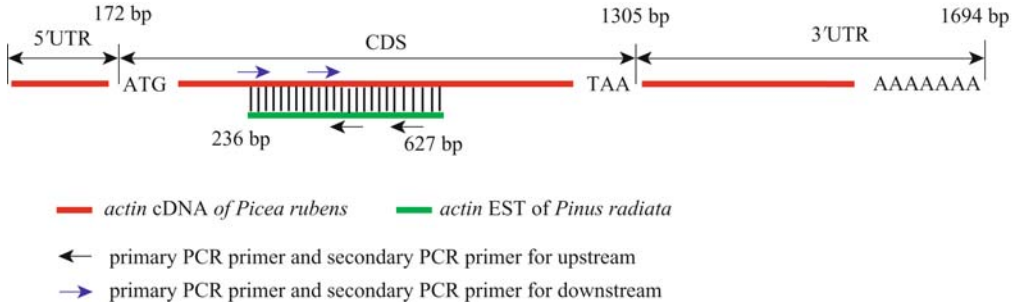

Fig. 1 Blast result of *actin* EST of *Pinus radiata*

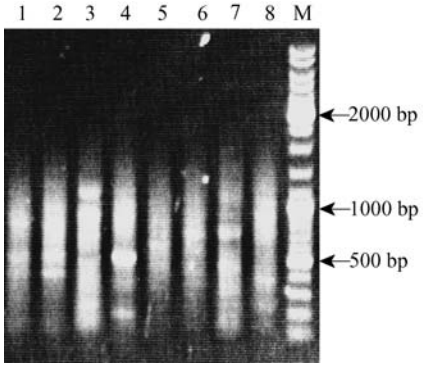

Fig. 2 Primary PCR results

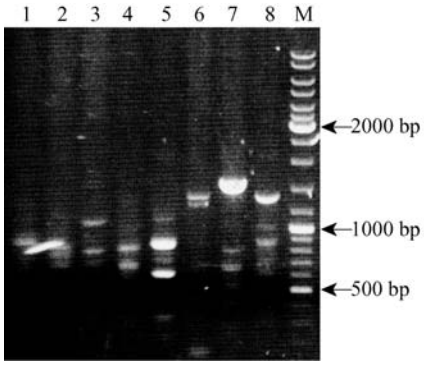

Fig. 3 Secondary PCR Results

walking. Given the NCBI nucleotide blast, the most similar cDNA sequence was the *actin* gene from *Brassica rapa* with a similarity of 72%. Alignment with *Brassica rapa actin* gene, the start and stop codes of the radiata pine *actin* gene were confirmed in the 231st and 1653rd bases. The genomic sequence contained 231 bp of 5′ UTR in the promoter region and 501 bp of 3′ UTR. Based on the genomic sequence, two pairs of primers were designed for SNP detection. Primer1 was: left 5′-CCTGCTATGTATGTTGCCATTC-3′ and right 5′-AGCCTCAAGTTTCAAGATGC-3′. Primer 2 was: left 5′-TTCTAATAGTTGGCCCTGTGGT-3′ and right: 5′-AACAAACAAAGCAATCACATGC-3′. Figure 4 shows the position of the primer of the *actin* gene.

3.2 Genomic PCR and SNP marker discovery

The DNA templates for genomic PCR were isolated from the evenly mixed needles of 200 unrelated radiata pine trees. The primers for SNP detection were designed from the known genomic DNA. For the radiata pine *actin* gene, two pairs of primers were designed for genomic DNA. The size of the PCR products was 600 bp (Fig. 5, lanes 1–7) and 1500 bp (Fig. 5, lanes 8–15). To reduce nucleotide mutation caused by PCR, the reaction cycle number was set at 25.

Genomic PCR results were cleaned up, ligated to T-vector and transferred to *E. coli*. Forty white blots were selected for PCR identification and sequencing. Figure 6 shows the genomic PCR results where lanes 1 to 27 contained 1500 bp bands and lanes 28 to 54 contained 600 bp bands. The sequencing results were blasted to the GenBank and the fake clones were deleted. True SNPs were detected by using the Mega 3.1 software.

A total of 40 *actin* gene sequences were run on DnaSP software with a minimum frequency of 5%. A total of 21 SNPs were detected in the genomic sequences of 2154 bp after excluding insert and missing nucleotide and the

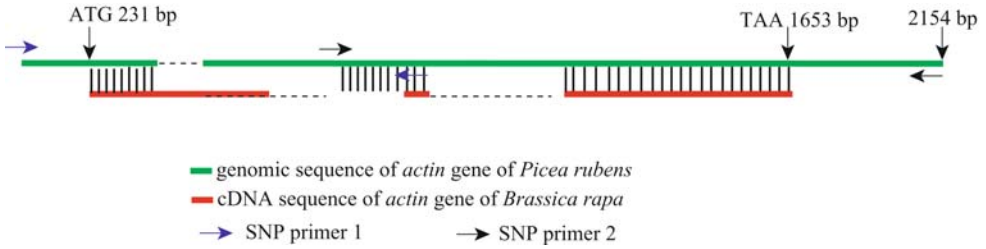

Fig. 4 Blast result of actin genomic DNA of *Pinus radiata*

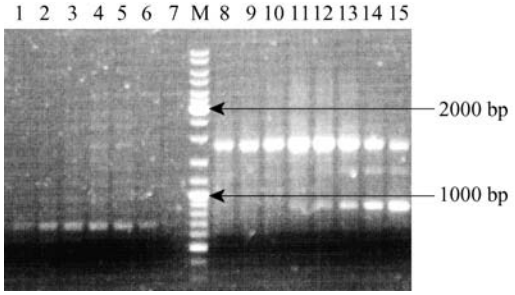

Fig. 5 Genomic PCR results

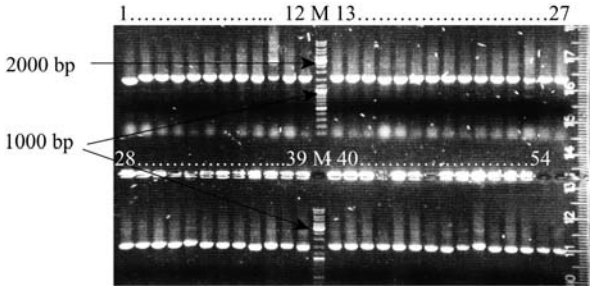

Fig. 6 Plasmid PCR results

SNPs with distance less than 50–60 bp from another SNPs. Among the 21 SNPs, three were in the promoter region, five in 3′ UTR, 13 in the coding region, with frequencies between 7.5%–37.5% (Table 1).

4 Discussion

SNPs are usually discovered in the coding and non-coding regions of genes. Coding regions are always related to the structure and function of the translated proteins, whereas non-coding regions are usually related to the gene regulation and pre-mRNA splicing to form mature mRNA (Ng et al., 2001; Sunyaev et al., 2001). For the candidate genes, a coding region can be obtained by reverse-transcription of mRNA, but a non-coding region cannot be obtained directly through reverse-transcription. Chromosome walking provides an effective method for a non-coding sequence of candidate genes. Chromosome walking of the PCR itself has two steps with a high number of reaction cycles, so its PCR results cannot be used for SNPs detection directly. However, primers can be designed from the gene sequence derived from chromosome walking for discovery of SNPs in a sample population. During chromosome walking and the SNP discovery process, attention should be paid to the following points.

4.1 Primer design

Of the two PCR primers used in chromosome walking, only one was the gene-specific primer with its length from 25 to 27 bases and GC content from 40% to 60%. In the 3′ end of the six bases of the primer, the total number of G and C was not more than three. Because the secondary PCR was a further specification of the primary PCR, the results will be better if GSP1 and GSP2 for primary and secondary PCR do not overlap. If the primer design is based on the cDNA sequence, the primer location should be in the exon region.

4.2 PCR reaction

Generally, primary PCR results are simple and some libraries may have several PCR bands. The reason for this

Table 1 SNP detection results for *actin* gene in *Pinus radiata*

number	SNP marker	region	position	frequency/%
1	TTCCCCTTGTT[G/A]TAGTTGCATT	promoter	−125	12.5
2	TTTGTACATTT[C/T]GATTGTTGTA	promoter	−104	10.0
3	TCATTACGTT[G/A]CTTAGCTTGC	promoter	−39	7.5
4	GGAACTGGAA[C/T]GGTTAAGGTA	exon	53	20.0
5	GTTCTTGGAT[G/A]TATTGTATTG	exon	79	7.5
6	TACTGGATAA[T/C]ATTGTGCATA	exon	132	7.5
7	TCAGTCAAAA[T/A]GAGGTATCCT	exon	364	7.5
8	GAAGAACATC[T/C]TGTACTTCTT	exon	485	20.0
9	GCCATTCAG[T/G]CAGTTCTGTC	exon	591	10.0
10	GTACAACTGG[A/T]GAGCATGTGA	exon	629	37.5
11	TTGCTAAGGG[C/T]TTATATTAAC	intron	703	15.0
12	TTAGTCACAC[G/A]GTGCCAATTT	exon	774	17.5
13	TCCATCATGA[T/A]GTGTGATGTG	exon	1139	7.5
14	CAGCAGCATG[A/G]AAATCAAGGT	exon	1265	12.5
15	TTATGATTTC[T/C]TATTACCTTC	intron	1396	17.5
16	TCACTTCTCA[T/C]TGTTCTTATT	intron	1449	17.5
17	GATGATGGTG[G/A]CAATAGTGCT	3′UTR	1626	17.5
18	CAGTTTATTT[T/C]TTATGCAAGG	3′UTR	1659	20.0
19	AGCATAGTTT[T/A]GTGGCTCTGT	3′UTR	1715	7.5
20	TTGCTAGATG[T/C]GACCAGAAAC	3′UTR	1758	7.5
21	TTTGAACCCT[G/A]TAATGTTATG	3′UTR	1803	7.5

phenomenon can be explained by the fact that the adaptor, ligated to the digested genomic DNA fragments, was formed with a long DNA chain and a short complement DNA chain with ammonificated 3′ end base. The AP1 primer designed from the long DNA chain, but not a short complement DNA, can match the AP1. Therefore, the first PCR cycle of the primary PCR can only start from a gene-specific primer and consequently result in a faint PCR. Sometimes shortening the preheating time can improve the quality of primary PCR results. If there were no clear bands in the secondary PCR, while the primary PCR produced simple bands, the secondary PCR can be carried out and may also produce a clear, single band. To improve PCR specificity, the anneal temperature is usually set at 67°C. If more than one band comes out, the anneal temperature can be increased proportionately.

4.3 Library construction

Library construction is critical to the chromosome walking technology. The size of PCR products is usually related to the selected restriction endonuclease. The library construction process involves cutting, cleaning up and ligation for genomic DNA. Certain fragments may be lost during PCR and thus do not produce PCR results in both primary and secondary PCR. In addition, genomic sequences for genes always produce various different segments by one restriction endonuclease. Therefore, it is rare to obtain the full gene sequence by constructing one library alone. To clone a full genomic gene sequence, using gene walking techniques, several gene walking libraries are required using different restriction endonucleases. With several libraries, each chromosome walking PCR can provide more than one band for selection. Generally speaking, the more libraries are constructed, the higher the probability to obtain the targeted fragment. But with more libraries constructed, there will be more costs associated with primary and secondary PCR. For our experiment with radiata pine, four libraries seemed adequate to obtain satisfactory chromosome walking results.

Acknowledgements The experiment was conducted while the senior author was at the CSIRO Forest Bioscience, a joint CSIRO/China Administration of the State Forest Post-Doctoral Program with partial funding from the Juvenile Wood Initiative to HXW (PN03.1916). The EST of radiata pine *actin* gene was derived from a microarray experiment by Dr Xinguo LI. Assistance by Dr Shannon DILLON on chromosome walking is greatly acknowledged.

References

Chasman D, Adams R M (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure based assessment of amino acid variation. JMB, 307: 683–706

David W, Meinke J, Michael C (1998). *Arabidopsis thaliana*: A model plant for genome analysis. Science, 282: 662–682

Doyle J J, Doyle J L(1990). Isolation of plant DNA from fresh tissue. Phys Rev Focus, 12: 13–15

Feltus F A, Singh H P, Lohithaswa H C (2006) A comparative genomics strategy for targeted discovery of single nucleotide polymorphisms and conserved-noncoding sequences in orphan crops. Plant Physiol, 140: 1183–1191

Goff S A, Ricke D, Lan T H (2002). A draft sequence of the rice genome (*Oryza* L. ssp. *japonica*). Science, 296: 92–100

Gupta P K, Roy J K, Prasad M (2001). Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Current Sci, 80: 524–535

Kumar S, Tamure K, Nei M (2004). MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinfor, 5: 2

Landegren U, Nilsson M, Kwok P Y (1998). Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. Geome Res, 8: 769–776

Ng P C, Henikoff S (2001). Predicting deleterious amino acid substitutions. Genom Res, 11: 863–874

Rozas J, Sanchez-delbarrio J C, Messegure X (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics, 19: 2496–2497

Sanger F, Nicklen S, Coulson A R (1997). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci, 74: 5463–5467

Sunyaev S, Ramensky V, Koch I, Lathe W, Kondrashov A S, Bork P (2001). Prediction of deleterious human alleles. Human Mol Genet, 10: 591–597

Tuskan G A, Difazio S, Jansson S (2006). The genome of black cottonwood *Populus trichocarpa* (Torr. & Gray). Science, 313: 1596–1604