Development of PCR based markers in Terpene synthase genes for marker assisted selection of high resin yielders in *Pinus roxburghii* Sarg.

S. Allen*, H.S. Ginwal and S. Barthwal

Division of Genetics and Tree propagation, Forest Research Institute, Dehradun 248006 India

(Received 27th August 2015)

**Abstract**

Pine oleoresin is a complex mixture of volatile and nonvolatile terpenes and is exploited for commercial production in India. Pines have long reproductive cycle vegetative phase extending up to many years. Therefore, there is an urgent need to identify high resin pine yielders at an early stage. Due to the strong influence of genetic factors on resin yield in pine species, marker assisted selection may serve as a potential tool for early identification of the genotype of interest. This study was planned to identify high resin pine yielders at early stage to avoid exploitation of natural germplasm of chir pine. Protein sequences of different terpene synthases were downloaded from the NCBI database and were multiply aligned to identify conserved and variable regions across the sequences. Primers were designed based on the related coding sequences to target these regions using Primer 3 software to amplify and isolate related genomic loci in chir pine genotypes. These primers were characterized for specificity using Primer Blast and the presence of hairpin loop formation using OligoCalc software. DNA was extracted from different resin yielding pine genotypes and primers were run on them for molecular studies. With the help of bioinformatics tools, we were able to associate some markers with resin yield. This study holds a key promise for the conservation of pine germplasm in natural habitat.

**Key words:** *Pinus roxburghii* Sarg, Terpene synthase, PCR-marker, Bioinformatics, association analysis.

**Introduction**

*Pinus roxburghii* Sarg., commonly known as Chir pine, belongs to family Pinaceae, and is one of the most extensively distributed pine species in India which is mainly exploited for resin. The forests of *P. roxburghii* are found ranging from longitudes 70°E to 93°E and latitudes 26°N to 36°N in subtropical and warm temperate monsoon belts, between 450 to 2300 m altitudes in Siwaliks and Himalayan main river valleys, from Kashmir to Bhutan. 'Naval Stores' is a term used to denote the products obtained from the oleoresin or resin of pine trees (Coppens and Hone, 1995). Oleoresin is a complex mixture of terpenoids (Keeling and
BOHLMANN, 2006), consisting of a turpentine (monoterpene C10, and sesquiterpene C15) and rosin (diterpene C20) fraction. The α- and β-pinenes of turpentine are the starting materials for a wide range of fragrance, flavours, vitamins and polyterpene resins and form the basis of a substantial and growing chemical industry.

As a result of the advancement in high sequencing technologies, genomics has resulted in the deposition of a huge quantity of DNA sequence data from terpene synthase gene regions in NCBI database (www.ncbi.nlm.nih.gov). Rapid accumulation of sequence and expression data of terpene coding gene regions would enable the development of specific genomic loci markers. Till date, no sequence coding for terpene synthase in chir pine has been submitted in the NCBI database. Resin yield is a highly heritable trait and important genetic gains can be obtained from the selection of the high resin yielders (MERGEN et al., 1955; SQUILLACE and DORMAN, 1959; GODDARD and PETERS, 1965).

In India, much less work has been done in understanding the natural variation in resin yield in Pinus roxburghii which leads to wide exploitation of natural forest of chir pine in Himalayan region. To identify the allelic variation of relevant traits within the genetic resource collections, marker assisted selection can be used as a potential approach for the same. Marker based selection helps in identification of favorable allele at a gene of phenotypic relevance than individuals taken at random. Based on the literature survey and problem of over exploitation of pine over the years, this study was aimed to identify high resin pine yielders at early stage to minimize damage in natural germplasm. Once a genetic marker has been demonstrated to be associated with a phenotypic trait of interest, it can be used as a selection target to obtain an indirect response in the trait. This study was initiated with a broad objective of identifying different genomic loci specific markers associated with the resin production by studying polymorphism in phenotypically varying populations of P. roxburghii.

Materials and Methods

Primer design

Protein sequences and related coding DNA sequences were downloaded from NCBI in the FASTA format. In order to increase the chance of specific primer pairs for a specific primer pair, at least one primer should be located in regions where the PCR template does not share high similarity to unintended targets if possible. A primer pair is deemed to be specific only if has no amplicons on any targets other than the submitted template within the specificity checking threshold specific else considered non-specific. The combined sets of sequences were multiply aligned using MEGA 5.05 software and primers were designed using conserved and variable regions. Primers were designed to amplify terpenes specific regions eg pinene synthase, geranyl diphosphatase, etc. EST/coding sequences for the enzymes which play a key role in terpene metabolic pathway were downloaded and primers were designed which were then amplified on the genomic DNA isolated from pine samples collected from Uttarakhand region.

While designing primers, PCR amplicon size was limited between 100–800 bp. With the help of BLAST, available pine genomes available on NCBI were checked for matches and for unique locations for each of the amplicons. These primers were then cross checked on OligoCalc software for secondary hairpin loop like formation if any. Primer-BLAST resulted in similar product size as expected.

Selection of germplasm

In the previous study (RAWAT et al., 2014), highly variable fifty three genotypes were selected out which were collected from different aspect i.e. northern and Southern aspect and also from different sites having maximum resin yield variation from as low as 0.80 kg/year to a maximum of 6.40kg/year. In the present study, these genotypes were screened by PCR amplification of genomic fragments in different terpene synthase genes using newly designed primer pairs.

Extraction of Genomic DNA from needles

The present protocol was modified from DOYLE and DOYLE (1990) and STANGE et al. (1998). Young needles (500 mg) were ground in liquid nitrogen and immediately transferred to the preheated (60°C) extraction buffer (1.0 ml CTAB buffer+3 µl β-mercaptoethanol). The ground tissue was then incubated at 60°C in water bath for 45 min. and then kept at room
temperature for 5–10 min. 500 µl of chloroform: isoamyl alcohol (24:1) was added and centrifuged at 13,000 rpm for 5 min. at room temperature. The upper aqueous layer was transferred to sterile 1.5 ml micro centrifuge tubes and an equal volume of chilled isopropanol (–20°C) was added to the supernatant to precipitate the DNA. Tubes were then centrifuged at 13,000 rpm for 15 min. at 4°C, and the supernatant was discarded. DNA pellet was dried in vacuum concentrator for 10 min. at 37°C and re-suspended in 100 µl of TE buffer.

PCR – genotyping of different individuals

Sterile ultrapure water was added to obtain a primer concentration of 50 pM/µl based on the concentration of stock. 40 µl of this solution was distributed into a number of 0.6 ml micro centrifuge tubes. The tubes were then placed in DNA vacuum concentrator to dry the primer aliquots and stored at –20°C. Finally, 100 µl of sterile ultrapure water was added to the tube, vortexed, spinned down and allowed to dissolve at room temperature for 15 min. to give a 20 µM working solution.

Primers were tested on random subsets of six individuals before using them for PCR on all samples. Tests of amplicon were around the expected size and single band producing primers were further used. Multiple bands producing primers were redesigned.

Genetic Structure analysis and AMOVA analysis

For the analysis of population structure, Evanno method (Earl et al., 2012) of cluster analysis was performed based on loci. This analysis was implemented in the software STRUCTURE HARVESTER (Earl et al., 2012) which identify subgroups of accessions with distinct allele frequencies within the germplasm using Structure ver 2.2 (Pritchard et al., 2000a and 2000b) outputs. STRUCTURE computes a Q matrix defined as an n x p population structure incidence matrix where n is the number of individuals assayed and p is the number of sub-populations assumed, Q is inferred from Pritchard’s STRUCTURE estimates with p (Pritchard’s K) sub-populations. The model based cluster analysis was used to test the hypothesis of one to ten sub-populations (K=1 to K=10) assuming admixture and correlated allele frequencies in different subpopulations. 100,000 iterations and a burn-in period of 100,000 were carried out for each run. Ten independent STRUCTURE runs were performed separately for each K. The results folder created by STRUCTURE was then loaded in zipped for-

![Figure 1. – Multiple sequence alignment of protein coding sequences from Pinus genus using Clustal W in MEGA 5.05 software.](image)
mat in Structure Harvester software and value of delta K was calculated using Evanno method (Earl et al., 2012) and plots the median value of ln Pr(K) for each K.

$$\Delta K = m (\frac{1}{2} L(K+1) - \frac{1}{2} L(K) + L(K) - L(K-1)) / s[L(K)]$$

Where, L(K) is Ln P(D), the posterior probability of the data for a given K, Pr(X|K) in STRUCTURE output, s[L(K)] is the standard deviation of L(K), and m is mean in the parenthesis. \(\Delta K\) shows a clear peak at the true value of K. The population genetic structure was inferred by an analysis of variance framework (AMOVA) according to Excoffier et al. (1992) using the Arlequin software version 3.11 (Excoffier et al., 2005). The type of hierarchial AMOVA implemented here was with genotypic data, one group of populations and number within individual level. This technique treats genetic distances as deviations from a group mean position, and uses squared deviations as variances. The total sum of squares of genetic distances can then be partitioned into components that represent the within population and among population mean squares.

AMOVA was carried out in two ways i.e. by collection site (genotypes were grouped on the basis of site of collection) and by resin yield (genotypes were grouped on the basis of resin yield).

**Results**

**Bioinformatics analysis**

In order to target variations in resin yield among the selected genotypes of chir pine, we
aimed to develop PCR-based markers that detect specific fragment size differences in terpene synthase genes. For primer designing, protein sequences and the related coding DNA sequences of different terpene synthases were downloaded from NCBI and were aligned with the help of Multiple Sequence Alignment (MSA) software MEGA. Partial sequences were removed for further analysis to avoid false results. (Figure 1: Section of a multiple sequence alignment of protein sequences of different Pinus terpene synthases).

Highly conserved aspartate rich motif (Bohlmann et al., 1998) namely DDXXD which is found in all terpenoids synthases and RRX₃W which is the characteristic of most of monoterpenes synthases was also found in MSA. These region help in regulation of biosynthesis pathways as single base pair mutation in these regions can alter the final product. PCR specific primers for terpene synthase genes were designed around these regions to study possible variation at coding sequences which may lead to low and high resin production (Table 1).

The protocol by Vendramin et al. (1996) with modifications successfully amplified the different genomic loci in *P. roxburghii*. Different concentrations of MgCl₂ (2.00 mM, 2.5 mM, 3.00 mM), 0.2 mM of dNTPs, 0.2 µM of forward and reverse primer, 0.1 unit of Taq DNA polymerase and 15 ng/µl of template DNA was found to be optimal for the amplification. Figure 2 shows gel image of genomic DNA extracted from needles.

**Molecular characterization of *P. roxburghii* genotypes using PCR-markers in different terpene synthase genes**

Ten genomic regions were found to be polymorphic when considering the related PCR-fragment length and were used to evaluate the level of genetic variation among the high and low resin yielding genotypes of *P. roxburghii*. These loci demonstrated a comparable degree of polymorphism across the genotypes. Figure 3 show the gel image of 8% PAGE of CAMP1.

PCR amplification using different genomic loci produced a total of 22 bands, out of which 14 were found to be polymorphic. The total number of bands amplified per marker varied from a minimum of 1 (DS7, FDP2, GGPP1, GGPP2, GPP1, PS6, PS9) to a maximum of 2 (CAMP1), with an average of 1.4 per marker. Genetic divergence in terms of per cent polymorphism ranged from 50 to 100% with a mean of 63.3% per marker (Table 2).
Genomic loci specific attributes

Polymorphism Information Content

The polymorphism information content (PIC) ranged from 0.171 (PS9) to 0.491 (DS7) with an average of 0.281 per marker.

Marker Index

The marker index (MI) ranged from 0.357 (DS9) to 1.706 (DS7) with an average of 0.75 per marker. There was a strong correlation between PIC and MI ($r^2=0.83$).

Resolving power

Resolving power (RP) ranged from 0.226 (DS9) to 1.566 (DS7) with an average of 0.74. There was a strong correlation between RP and PIC ($r^2=0.86$).

Genetic structure analysis using STRUCTURE software

The model-based clustering algorithm was applied to identify subgroups with distinct allele frequencies (PRITCHARD et al., 2000) using the software STRUCTURE ver 2.3.2. and Structure harvester. Structure harvester processes Structure results, generates in -files for use with CLUMPP, and when possible, executes the “Evanno method” (EVANNO et al., 2005). According to the Structure harvester results, the log likelihood steadily improved from K= 2 to maximum peak at K=5, and then sharply decreased at K=6 and then leveled off. The Evanno table output (Table 3) also highlights at K=5.

With five as the optimum population structure, inferred ancestries (Q matrix) of individuals were determined. All the genotypes were represented in five clusters represented by red, green, blue, yellow and pink coloured bars in the plot differentiating the genotypes (Fig. 4).

Linkage disequilibrium

The squared allele frequency correlations, $r^2$ and the p value representing linkage disequilibrium (LD) were assessed for these markers. The marker DS9 and DS7 were detected to be in significant linkage disequilibrium with $r^2$ values higher than 0.05. These regions code for Diterpene synthase which plays a key role in resin biosynthesis pathway. FDP 2 and GGPP2 were also found to be in significant linkage disequilibrium with DS9 showing the probability of this marker to be associated with resin yield (Figure 5).
Analysis of molecular variance revealed that the percentage of variation within the groups was higher than among the groups. AMOVA by collection site revealed only 3.91% of the total variation to be among the groups with an FST value of 0.005. On the other hand, in AMOVA by resin yield, the percentage of variation among the groups was 22.3% with an FST value of 0.01 which was highly significant (p<0.001).

Discussion

Forests are the reservoir of wild alleles that have commercial significance. Commercially important traits like timber, resin etc can be exploited for plantation sector to avoid exploitation of natural germplasm. Till date, no terpene
coding sequences from *P. roxburghii* were found in NCBI database, thus highlighting the less attention paid to this commercially important species. Protein sequences were downloaded and were used for multiple alignment studies. Partial, incomplete and duplicate sequences were later eliminated for further analysis to prevent the false results. With the help of MEGA 5.05 software, all sequences were aligned for identification of conserved and variable regions, motif search etc. Duran et al., 2009 also used MEGA software for alignment and Primer 3 software for primer designing to develop species specific primer against pine pathogen. Highly conserved aspartate rich motif (DDXXD) was present which is found in all terpenoids synthases and is recognised as binding site for the divalent metal ion of the substrate diphosphate-metal ion complex (Tarshis et al., 1994). The conserved motif RXX₃W, was also present which is the characteristics of most of monoterpenic synthases (Aubourg et al., 2002).

The present study was carried out with the aim to develop different genomic loci specific primers which can be used for marker assisted selection. Therefore we focused on identification of sequence variation manually as well as with the help of bioinformatics software for introgression of the identified superior alleles. This holds a key step for allele mining studies. The primers developed in this study show potential for other pinus species and so, are a toolkit for comparative analysis. The main problem which we encountered with the primers is that at times they produce multiple bands on agarose gel which was only solved by redesigning the primers. This work is supported by Kelleher et al. (2012) who carried out the similar work in *Populus tremuloides* and targeted genes coding for lignin biosynthesis pathway. Here, sequenced based allele mining was performed while in the former study, Ecotilling was used which is less cost effective but excellent for studying single nucleotide polymorphism based molecular markers.

Genomic loci specific primers have been used as marker assisted selection for many agricultural crops mainly in rice where thirty rice varieties were exploited for allelic variation for the development of specific primers which were used in marker assisted selection to improve grain quality in rice (Kumar et al., 2010). Out of these ten primers, primers coding for monoterpenic and diterpenic were found to be maximum polymorphic which constitute the major portions of resin. PCR amplification produced a total of 22 bands, out of which 14 were found to be polymorphic (The total number of bands amplified per marker varied from a minimum of 1 (DS7, FDP2, GGPP1, GGPP2, GPP1, PS6, PS9) to a maximum of 3 (CAMP1), with an average of 1.4 per marker. This is in agreement with Rawat (2013) where the total number of polymorphic bands amplified per marker varied from as low as 1 to a maximum of 4 (Pt30204 and Pt45002), with an average of 2.105 per marker. These primers also showed polymorphism which were used to calculate Linkage Disequilibrium using software TASSEL, which revealed that DS7 and DS9 are found to be in significant association with resin yield. CAMP1 was able to explain 13.8% of the total variation in resin yield when association was tested on account of the marker alone whereas when full model was used (including population structure) it could explain 30.72% of the total variation in resin yield (p<0.01).

Analysis of molecular variance also supported the genetic basis of resin regulation. The percentage of variation was significantly higher than the percentage of variation among the groups. No significant variation was observed among the groups when analysis of molecular variance was carried out by collection site whereas in AMOVA by resin yield, percentage of variation among population was 22.3% with an Fst value 0.01 which was significant (p<0.01). This supports the genetic basis of the trait in Pinus genotypes which are not correlated on the basis of site collection.

**References**


Association mapping for identification of ISSR and AFLP markers linked with resin yield in Himalayan Chir Pine (*Pinus roxburghii* Sarg.)

By A. Rawat¹, *) S. Barthwal¹, H. S. Ginwal¹ and S. B. Tripathi²

(Received 4th September 2015)

### Abstract

In *Pinus roxburghii*, marker-trait associations were studied for resin yield using a set of 238 DNA based markers derived from 9 ISSR and 5 AFLP primer pairs. A total of 240 genotypes of *P. roxburghii* from a natural population in Chakrata division (Tiunee range), Uttarakhand (India) were evaluated for resin yield. Based on stable resin production in consecutive years, 53 genotypes that were best representatives of the variation in resin yield (0.25 to 8.0 kg/tree/year) were used for genotyping and association analysis. For carrying out the association studies, two approaches were used. First, using simple linear regression, resin yield was regressed on all 238 available polymorphic markers (148 ISSR markers and 90 AFLP markers) and second, evidence for association was examined applying population structure information using STRUCTURE and STRAT to avoid any spurious associations between a candidate marker and phenotype. In regression analysis, 16 ISSR and 12 AFLP markers showed significant association with resin yield. In the second approach, 18 ISSR and 18 AFLP loci were found to be associated with resin yield in the absence of population structure out of which 2 ISSR and 3 AFLP loci were common with those associated in regression analysis. After accounting for population structure, 38 ISSR, and 3 AFLP loci revealed association out of which one ISSR and one AFLP locus was found to be common with both simple linear regression and test of association in the absence of population stratification. There was no obvious correlation between population structure and resin yield.

**Key words:** regression, population structure, structured association test, ISSR, AFLP, *Pinus roxburghii*.

### Introduction

*Pinus roxburghii* (Sarg.) commonly called as long leaf pine or Himalayan chir pine is one of the most important conifers found in the lower Himalayan region between latitudes 26°N and 36°N and longitudes 71°E and 93°E in subtropical and warm temperate monsoon belts, between 450 to 2300 m altitudes in Siwaliks and Himalayan main river valleys, from Kashmir to Bhutan (GHILDIYAL et al., 2009). Trees of *P. roxburghii* yield the highest amount of oleoresin in India (COPPEN and HONE, 1995), which is a commercially important product, having huge export potential. The annual production of resins in India as estimated by the state forest departments was found to be 27,961 tons in...