INTERNATIONAL JOURNAL OF ADVANCED BIOLOGICAL RESEARCH

www.scienceandnature.org

COMPARATIVE SEQUENCE ANALYSES REVEAL A SPECIES-SPECIFIC NUCLEOTIDE SEGMENT IN *MATK* REGION OF *SANTALUM ALBUM* L.

Tresa Hamalton
Silviculture and Forest Management Division,
Institute of Wood Science and Technology, Bangalore, India
Corresponding Author E-mail: d.tresa@gmail.com

ABS**TRACT**
Santalum album L., the most valued East Indian sandalwood, is highly prone to theft and adulteration which is difficult to monitor. Genetic methods such as phylogenetic analysis of rDNA sequences and DNA barcoding are promising for detection of *S. album*, but require a reference database for sequence comparison without which these methods cannot be practically utilised. Though DNA barcoding was documented to be useful for distinguishing *S. album* from adulterant species, the traditional method is not precise as the existing databases of standard DNA barcodes are not comprehensive. When a sequence identical to the DNA barcode of the species investigated is not available in the database, it will be predicted as the species presenting the closest match. This triggered the search for short nucleotide fragments within the *S. album* genome, which can be used for exactly ascertaining species identity. In order to identify sequences in the genome of *S. album*, which possess no variation within the species but differ from homologous regions of other species, the highly conserved *matK* region was selected, and nucleotide sequences for *S. album* and its substituent species were retrieved. Their comparative sequence analyses revealed a 357nt segment specific to *S. album*, and the species-specific nucleotide segment (SSNS) was verified using NCBI-BLASTn and BOLDIDENGINE. The SSNS of *S. album* identified in this study is a potential candidate for the development of marker systems for discriminating it from its substituent species.

KEYWORDS: Sandalwood, adulterants, timber identification, species specific nucleotide segment, *in silico* analysis

INTRODUCTION

Santalum album L. is the most valuable tropical hardwood species from India, known for its fragrant essential oil and valuable carving wood, and also possesses medicinal value (Mahesh *et al.*, 2018). In commercial markets, it is highly prone to adulteration, tracing of which is difficult due to the lack of technical tools for accurate identification of *S. album* or discrimination from the other species (Dev *et al.*, 2014). The fragrant-oil yielding heartwood of *S. album* is the most traded, and is substituted with many look-alike species including *S. acuminatum, S. austrocaledonicum, S. lanceolatum, S. murrayanum, S. spicatum* and *S. yasi* within the genus, and also *Buxus sempervirens, Chukrasia tabularis, Erythroxylum monogynum, Osyris lanceolata, O. wightiana* and *Ximenia americana* from other genera. These species are scented woods or have similar wood anatomical features as *S. album*, rendering them undistinguishable (Jiao *et al.*, 2018; Chembath *et al.*, 2012). Also, in the case of powdered wood samples and mixtures of the same used in medicinal formulations, the lack of morphological features and the non-availability of suitable chemical methods make it impossible to recognise the species of origin, thereby depending solely on genetic approaches for species discrimination (Hamalton, 2018).

Chembath *et al.* (2012) sequenced the 18S and 26S rDNA sequences of *S. album, E. monogynum* and *O. wightiana*, and sequences of 4 other substituent species were recovered from the NCBI nucleotide library. Though phylogenetic analysis of these sequences could distinguish the 6 species, rDNA sequences belong to the nuclear genome and are not easily recovered from aged or degraded tissues, as compared to plastid DNA. Dev *et al.* (2014) sequenced the standard DNA barcode regions (*rbcL, matK* and *trnH-psbA*) of *S. album, E. monogynum* and *O. wightiana*. Comparison of *rbcL* sequences of these species has revealed the presence of species-specific SNPs, which can be used to discriminate between them. Jiao *et al.* (2018) screened different DNA barcode regions of 5 *Santalum* spp. including *S. album*, and demonstrated that combination of *psbA-trnH + trnK* can be used for phylogenetic analysis among the 5 species to identify each of them. Sequencing and characterisation of the nuclear, chloroplast and mitochondrial genomes of *S. album* individuals has also been performed (Mahesh *et al.*, 2018; Dasgupta *et al.*, 2019; Yang *et al.*, 2020), which can be used for comparison with genomes of the look-alike species to devise mechanisms for species discrimination. The current study was aimed at identifying species-specific nucleotide segments (SSNS) within a standard barcode region present in the plastid genome, which can be easily retrieved from aged/degraded tissues, for developing a molecular marker system for distinguishing *S. album* from its commercial substituents without the need for a comparative database.

MATERIALS AND METHODS

The *matK* nucleotide sequences of 5 *Santalum* species and 6 substituent species from other genera, were retrieved from NCBI nucleotide library and BOLDSYSTEMS. Comparative sequence analysis was performed for different sets of sequences *viz.*, intraspecific, intrageneric and interspecific. 'KALIGN' from EMBL-EBI with default parameters (https://www.ebi.ac.uk/ Tools/ msa/ kalign), was used for performing multiple sequence alignment (MSA), and BIOEDIT 7.2.5 (https://bioedit. software. informer.com /7.2) was used for viewing alignments and editing sequences.

Intraspecific MSA was performed individually for each of the 11 species, and the alignment was screened for locating nucleotide polymorphisms. By trimming the ends at all positions presenting end gaps in the alignment, and by replacing nucleotides at positions presenting variations (with R/Y for transitions, and N for transversions), a single representative *matK* nucleotide sequence was derived for each of the species. To study the intrageneric variation in the *matK* sequences within *Santalum* genus, MSA was performed for the derived representative *matK* sequences of the 5 *Santalum* species. The interspecific variation between *S. album* and its substituent species was studied by MSA with derived single *matK* sequences of 6 substituent species of other genera, and also with all the 10 substituent species. The species-specific nucleotide segment (SSNS) for *S. album* was thus identified. It was then validated by comparison with all the sequences in NCBI nucleotide library using BLASTn, and by submission in BOLDIDENGINE.

RESULTS AND DISCUSSION

Being a highly conserved region in the chloroplast genome, *matK* region was used in this study to locate species-specific nucleotide fragments which can be used to discriminate *S. album* from its substituent species. During the intraspecific alignment of *matK* sequences of all 11 species, the length of the final derived partial sequences for each of the species varied. The details of the species-wise final partial *matK* sequences are listed in Table 1.

TABLE 1: Details of derived single sequences of *matK* for each species

Species name	Final partial sequence length (nt)	No. of substitutions (nt)
S. album	357	0
S. acuminatum	357	2
S. lanceolatum	357	2
S. murrayanum	357	0
S. spicatum	316	1
B. sempervirens	662	15
C. tabularis	777	0
E. monogynum	679	0
O.lanceolata	713	0
O. wightiana	674	0
X. americana	706	1

The 357nt region designated as the derived *matK* sequence of *S. album* was found to be exactly similar in all the 30 sequences compared, and is therefore determined to be highly conserved within the species. Hence, this single partial *matK* sequence is henceforth designated as species-specific nucleotide segment (SSNS) of *S. album.* The same region in the *matK* sequences of other *Santalum* spp. presented variations in the intrageneric alignment, which were in the form of nucleotide substitutions without any in-dels. The species-specific SNPs encountered during intrageneric comparison of *Santalum matK* sequences are listed in Table 2. SNPs detected during interspecific comparison of homologous sequences have also been proposed earlier as markers for species identification of *S. album* (Dev *et al.,* 2014).

During interspecific comparison of the 357nt long derived *matK* sequence of *S. album* with the other 10 species, the homologous region in the *matK* sequences presented variation except with *E. monogynum*. This is evident in the phylogenetic tree for interspecific comparison in which derived partial *matK* sequences of *S. album* and *E. monogynum* are clustered together with NJ distance=0 (Fig. 1). A 357nt DNA sequence was also found to be best aligned for *matK* within *Santalum* genus in the study by Jiao *et al.* (2018). However, the unique and consistent nature of the *S. album* segment is reported herein for the first time. The percent identity for pairwise alignment between the *S. album* sequence and the other species is tabulated in Table 3, which showed that the partial *matK* sequences of *E. monogynum* and *S. album* are similar. The identical nature of *E. monogynum* and *S. album matK* sequences observed in the sequence alignment, phylogenetic tree and percent identity matrix, has occurred because the *matK* sequence of *E. monogynum* closest to that of *S. album* was initially chosen in this study (KC503282), which had 17.1% variation with the other *matK* sequence of *E. monogynum* (MG737440).

TABLE 2: Intrageneric SNPs in *Santalum* partial *matK* sequences

Species	Position	Substitution identified	
S. acuminatum	19	Transversion	A→N
	95	Transversion	G→C
	146	Transversion	T→G
	242	Transversion	A→C
	312	Transition	T→C
	335	Transversion	A→N
S. lanceolatum	167	Transversion	A→N
	173	Transversion	A→C
	280	Transition	A→R
S. murrayanum	95	Transversion	G→C
	115	Transition	C→T
	146	Transversion	T→G
	350	Transition	C→T
S. spicatum	83	Transversion	T→G
	95	Transversion	G→C
	146	Transversion	T→G
	270	Transition	G→A
	279	Transition	C→T
	320	Transition	C→T
	332	Transition	T→Y

TABLE 3: Percent identity with *S. album* partial *matK*

Species	Percent identity
S. acuminatum	98.32
S. lanceolatum	99.16
S. murrayanum	98.88
S. spicatum	97.78
B. sempervirens	82.91
C. tabularis	80.39
E. monogynum	100.00
O. lanceolata	94.96
O. wightiana	95.52
X. americana	85.63

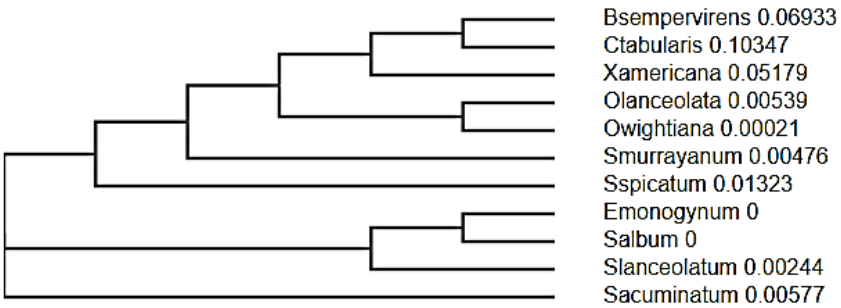

FIGURE 1: Phylogenetic (NJ) tree for interspecific comparison with all substituent species

The species-specific partial *matK* sequence of *S. album* was then validated by comparison with all the sequences in NCBI nucleotide library using BLASTn. Out of all the comparisons, 100% similarity was obtained with 30 sequences. 28 among the 30 sequences belonged to *S. album*, one sequence belonged to *S. boninense* and one sequence (KC503282) belonged to *E. monogynum*. The partial sequence of KC503282, was already known to be identical with *matK* sequence of *S. album*, and also used in this study. From the BLASTn results, it is evident that the partial *matK* sequence of *S. album* can be used at 93% accuracy for identification of *S. album* samples. Since it has the ability to resolve species identification, it is designated as a species-specific nucleotide segment (SSNS), and this is the first report of a unique species identifier (Hollingsworth *et al.*, 2011).

The SSNS of *S. album* was then submitted to BOLDIDENGINE. Among the top matches, 100% similarity was obtained with 19 sequences. These 19 sequences also included two *S. boninense* sequences, one *Zingiber officinale* sequence, one *E. monogynum* sequence and two terrestrial samples. Though 13 out of the 19 sequences belonged to *S. album*, the best match displayed was *S. boninense*. Since the BOLDIDENGINE results

show that the SSNS identified in this study can be used with only 68% accuracy for species identification, it is proposed that a longer fragment of the *matK* region needs to be identified for use as SSNS of *S. album* with 100% accuracy.

CONCLUSION

The SSNS identified for *S. album* needs to be validated with diverse samples of *S. album*, and cross verified with samples of other species to circumvent a mismatch. After confirming the individuality of the SSNS by comparison with all available datasets, it can be used for simple, fast and accurate species authentication. A simple or nested PCR for DNA extracted from the sample with primers specific to the SSNS can be used to detect *S. album* by visualising the presence or absence of bands after electrophoresis. Alternatively, DNA hybridisation chips containing strands complementary to the SSNS of either single or multiple species can be used to infer identity of species in case of substitution for wood samples, and to identify adulterants in mixtures of powdered wood (like medicinal preparations). Upcoming research areas for discriminating *S. album* from its commercial substituents should also focus on (i) standardising a uniform method for DNA extraction from all tissues of *S. album* and its substituent species, (ii) sequencing and compiling the DNA barcodes of *S. album* and its substituents to enrich existing databases, (iii) sequencing whole plastid genomes to identify other potential SSNS which can be used for species identification complementing the one reported here, and most importantly (iv) developing a simple, reliable and cost-effective method for species discrimination.

In the era of DNA barcoding, which involves extracting the total DNA, amplifying the standard barcode region, sequencing it and then searching comparative databases for establishing species identity, this manuscript reports the discovery of a 'species-specific segment for *Santalum album* L.', making it a potential candidate for 'simple and rapid identification of sandalwood' especially in forensic discrimination. Similar sequence comparisons of the same loci from different species can be used to identify SSNS for other species as well. Many researchers have debated on the accuracy and limitations of DNA barcoding for species identification. In contrast to conventional DNA barcoding, the techniques employing the SSNS possess higher discriminatory power as they do not require a reference database to obtain a high similarity score for species identification, but the inference of species is explicitly based on the occurrence of the SSNS in the sample. It is expected that this study will open new avenues in the research on techniques for precise species identification.

REFERENCES

Chembath, A., Balasundaran, M. and Sujanapal, P. (2012) Phylogenetic relationships of *Santalum album* and its adulterants as inferred from nuclear DNA sequences. *Int. J. Agric. Biol.* 2(4), 150-156.

Dasgupta, M.G., Ulaganathan, K., Dev, S.A. and Balakrishnan, S. (2019) Draft genome of *Santalum album* L. provides genomic resources for accelerated trait improvement. *Tree Genet. Genomes*. 15, 34.

Dev, S.A., Muralidharan, E., Sujanapal, P. and Balasundaran, M. (2014) Identification of market adulterants in East Indian sandalwood using DNA barcoding. *Ann. For. Sci*. 71(6), 517-522.

Hamalton, T. (2018) Genetic tools for timber forensics. *Van Sangyan* 5(2), 17-20.

Hollingsworth, P.M., Graham, S.W. and Little, D.P. (2011) Choosing and using a plant DNA barcode. *PLoS ONE* 6 (5), e19254.

Jiao, L., He, T., Dormontt, E.E., Zhang, Y., Lowe, A.J. and Yin, Y. (2018) Applicability of chloroplast DNA barcodes for wood identification between *Santalum album* and its adulterants. *Holzforschung* 73(2), 209-218.

Mahesh, H.B., Subba, P., Advani, J. (2018) Multi-omics driven assembly and annotation of the sandalwood (*Santalum album*) genome. *Plant Physiol*. 176(4), 2772-2788.

Yang, D., Qui, Q., Xu, L., Xu, Y. and Wang, Y. (2020) The complete chloroplast genome sequence of *Santalum album*. *Mitochondrial DNA Part B* 5(1), 406-407.