Industrial Crops & Products 140 (2019) 111600

Contents lists available at ScienceDirect

Industrial Crops & Products

journal homepage: www.elsevier.com/locate/indcrop

Transcriptome characterization and development of functional polymorphic SSR marker resource for Himalayan endangered species, *Taxus contorta* (Griff)

Aasim Majeed, Amandeep Singh, Shruti Choudhary, Pankaj Bhardwaj*

Molecular Genetics Laboratory, Department of Plant Sciences, Central University of Punjab, Bathinda, India

ARTICLE INFO

Keywords:
Paclitaxel
SSR
Transcriptome
RNA sequencing

ABSTRACT

Taxus contorta is an important medicinal plant currently listed as endangered in IUCN Red Data List. It produces an anticancer drug, paclitaxel which is well known in the industrial sector. Due to habitat destruction and overexploitation, it is at the verge of extinction. Genomic and transcriptomic data for this species is scarce which has hampered its genomic studies. Moreover, large scale polymorphic informative codominant marker resource is also scarce which hinders its population and landscape genetic analysis. Here, we generated a reference transcriptome for this species which would facilitate the understanding of the functional elements and promote genomic research in this species. Also, a robust polymorphic SSR marker resource was characterized which can be used in conservation of this species. More than 100 million paired end raw reads were obtained through Illumina sequencing. A total of 129,869 unigenes with mean sequence length of 1244 nt were obtained from 209,860 *de novo* assembled transcripts. Of these, 35,752 transcripts were assigned 5971 unique GO terms. Around 40,386 transcripts were found to have 2163 unique Pfam Ids. Pathway analysis against KEGG database yielded 3721 unique enzyme numbers. Screening of the transcripts for microsatellite regions yielded 7041 SSRs. Among the 100 SSRs selected for characterization on 30 genotypes, 37 polymorphic markers showed a total of 214 alleles with mean of 5.78 alleles per locus. Mean effective number of alleles (Ne) was found to be 3.64 and average PIC value of 0.64 was observed. Observed heterozygosity (0.57) was found to be lower than expected (0.69). This effective polymorphic SSR marker resource will act as valuable tool for deciphering its genetic diversity.

1. Introduction

Taxus contorta (earlier *Taxus wallichiana*) is currently categorized as endangered by the IUCN Red Data List (https://www.iucnredlist.org/species/46171879/9730085). In contrast to the common view that *T. wallichiana* is the only yew species distributed along the entire Himalayan range, recent morphological, molecular and climatic analysis distinguishes two different species in the Himalayan range viz *T. contorta* (Synonym *Taxus fuana*) which is distributed westwards from central Nepal upto Pakistan and Afghanistan and *T. wallichiana* distributed eastwards from central Nepal (Moeller et al., 2007; Shah et al., 2008; Poudel et al., 2012). But bulk of the literature still mentions the species of Taxus distributed along the Himalayan regions as *T. wallichiana* (sometimes also referred to as *Taxus baccata* ssp *wallichiana*) without any distinction between east and west Himalayan Taxus. So, much of the literature mentioning *T. wallichiana* actually refers to the newly identified *T. contorta*. The distinction has recently been acknowledged by IUCN Red Data List and Gymnosperm Database (https://www.conifers.org/ta/Taxus_contorta.php) and here we followed the same classification of *T. wallichiana* and *T. contorta* being the separate species.

T. contorta is an evergreen medium sized tree which is famous for its anticancer drug, paclitaxel. The tree grows scattered as an understory in the coniferous forests between the altitudes of 1800–3300 m asl (above mean sea level) (Juyal et al., 2014). It is known by different names locally in Indian Himalayas like Postul (Kashmir), Rakhal/Thuner (Uttrakhand) and Barmi (Himachal Pradesh. The genus Taxus has gained enough reputation in literature as well as in the industrial sector due to its anticancer drug synthesizing potential. Besides this major role, the plant has other utilities like anticonvulsant, analgesic,

* Corresponding author at: Molecular Genetics Laboratory, Department of Plant Sciences, Central University of Punjab, City Campus, Mansa Road, Bathinda, 151001, India.

E-mail addresses: pankajihbt@gmail.com, pankajbhardwaj@cup.edu.in (P. Bhardwaj).

https://doi.org/10.1016/j.indcrop.2019.111600
Received 12 November 2018; Received in revised form 22 July 2019; Accepted 23 July 2019

antipyretic, antibacterial, antifungal, anti-tuberculosis and hypoglycaemic properties (Ahmed et al., 2004; Banskota et al., 2006; Nisar et al., 2008a; Nisar et al., 2008b) that catch the attention of pharmaceutical industries. It is also used in treatment of other ailments like common cold, fever, cough, pain, respiratory infections, epilepsy, indigestion, snake bite, bronchitis, asthma, headache, diarrhoea and biliousness (Juyal et al., 2014; Khan et al., 2006; Purohit et al., 2001).

Populations of *T. contorta* have been left fragmented and isolated due to human overexploitation. The distribution of this species is continuously decreasing and is likely to be extinct if the current conditions prevail further (Molur and Walker, 1998). So, it is imperative to take robust steps for its immediate conservation. Before planning the conservation strategies for any species, it is necessary to decipher its genetic diversity and structure, which requires a robust marker resource. The informative marker resource for elucidating the population and landscape genetics of *T. contorta* is lacking. Although different studies have addressed this issue (Cheng et al., 2015; Gajurel et al., 2013; Yang et al., 2009) but have not been able to produce polymorphic SSR marker resource sufficient for genetic studies in this species. RAPD and AFLP markers were also utilized for diversity analysis (Mohapatra et al., 2009; Saikia et al., 2000; Zhang et al., 2009) but due to their dominant nature, they are not favoured. SSR markers have also been developed for different species of the genus Taxus and tested for cross species transferability (Liu et al., 2011).

One of the main limitations of studying non model plants at the genomic and transcriptomic level is the lack of structurally and functionally annotated reference genomes and transcriptomes. In order to get in-depth insights about the genomics of non-model species, focus needs to be turned to explore their genomes and transcriptomes. This would facilitate the discovery and exploration of novel genomic elements and functions and also decipher the unassessed genetic variation in nature. Nowadays, massive parallel sequencing platforms have greatly facilitated the genomic research because of the decreasing cost of input and relatively much decreased time for the generation of complete genomes (Koboldt et al., 2013). Through massively parallel sequencing technology, a bulk of the data at rapid pace has been generated both for model and non-model species (Ellegren, 2014). Trees in general and gymnosperms in particular remain uncovered at the genomic level. *T. contorta* being an important medicinal and industrial plant for decades is still unexplored at the genomic level. Large scale informative polymorphic marker resource for conservation genetic studies is also scarce for this species. In view of this, here we generated a functionally annotated reference transcriptome of *T. contorta* through RNAseq approach. The immediate application of the reference transcriptome would identify the SSR markers in *T. contorta*. The development and characterization of a robust polymorphic SSR marker resource for this species would prove useful in studying its underlying genetic diversity and structure, which is a pre-requisite for planning any conservation strategies of this industrially and medicinally important endangered plant.

2. Material and methods

2.1. Sample collection and nucleic acid isolation

Samples for RNA isolation were collected from the Salooni, Himachal Pradesh (N 32°44′12″, E 76°00′06″, altitude = 2154 m asl) in liquid nitrogen and transported to the lab. RNA was isolated using the protocol described by Kejani et al (2010) with some modifications. Quality and quantity of the isolated RNA was assessed through Qubit Flurometer and Nanodrop Spectrophotometer respectively. The integrity of the RNA was checked using Agilent 2100 Bioanalyzer. Leaf samples for DNA isolation were collected from different locations in Indian Himalayan regions (details of sampling points are given in supplementary file S1). DNA was isolated through CTAB method developed by Doyle and Doyle (1990) with some modifications. The quality and quantity was assessed through Nanodrop spectrophotometer ND-1000 and agarose gel electrophoresis. The isolated DNA was stored at -20^0C for subsequent use.

2.2. Library prepration and sequencing

RNA sequencing libraries were prepared with Illumina-compatible NEBNext® Ultra™ Directional RNA Library Prep Kit (New England BioLabs, MA, USA). 1 µg of total RNA was taken for mRNA isolation, fragmentation and priming followed by synthesis of first and second strands. The double stranded cDNA was then ligated with Illumina Universal Primers as per NEBNext® Ultra™ Directional RNA Library Prep Kit protocol. The adaptor ligated fragments were enriched to produce a sequencing library. The libraries from two samples were then sequenced on Illumina HiSeq 2000 platform using paired end approach.

2.3. De novo assembly

Cleaning of raw reads was done by removing low quality bases and adaptors through Trim Galore version 0.4.1. FastQC was used to check the quality of the raw reads. Concatenation of the quality trimmed reads from individual samples was done prior to the construction of assembly. *De novo* assembly from the concatenated and cleaned raw reads was constructed through Trinity version 1.6 (Grabherr et al., 2011). We used Read Representation, Blast against SwissProt and BUSCO for quality assessment of the assembly. Read representation was achieved by mapping the raw reads back to the assembly using Bowtie2 version2.3.0 (Langmead and Salzberg, 2012). Blast was done for counting the number of full length transcripts and BUSCO v2 (Simão et al., 2015) was used for assessing the completeness of the assembly. Removal of the redundant sequences and generation of unigenes was done through CD-HIT-EST version 4.6 (Li and Godzik, 2006) at 95% sequence identity threshold. The schematic representation of the pipeline used is shown in supplementary figure S2.

2.4. Functional annotation of unigenes

The non-redundant assembly was then functionally annotated using the annotation pipeline, Annocript (Musacchia et al., 2015). We used customised homology search against viridiplantae at an e-value of 0.00001. KEGG IDs were assigned to the transcripts by executing Blast through KAAS (Moriya et al., 2007). Further, the assignment of transcripts to gene families was done using the pipeline TRAPID (Van Bel et al., 2013) with PLAZA2.5 as a reference database at an e-value of 10e-5. Protein domains were identified from Pfam database. For the identification of transcription factors, the transcripts were aligned against the PlantTFDB v4.0 (Jin et al., 2016) at an e-value of 0.00001.

2.5. SSR screening and characterization

The assembled transcripts were screened for microsatellite markers using MISA (Beier et al., 2017). Positional distribution of the SSRs in the transcripts was analysed by predicting the ORFs from the SSR containing sequences using orfPredictor (Min et al., 2005) followed by correlating the SSR start and end positions with the start and stop positions of the predicted ORFs using an in-house python script. The primers for the SSR containing sequences were designed using BatchPrimer3 (You et al., 2008). The characterization of the synthesized primers was performed on 30 samples, representing three populations, through non denaturing PAGE followed by silver staining using the protocol developed by Huang et al (2018). The details of the populations are given in supplementary information S1. Bands were scored and the raw data was analyzed for the calculation of marker parameters like PIC using Cervus (Marshall et al., 1998), Observed and expected heterozygosity, Shannon's information Index and Diversity statistics using Popgene v3.2 (Yeh et al., 1999) and GenAlex v6.5 (Peakall and

Table 1
: Raw reads, transcripts and unigenes statistics.

Raw Reads	
Total bases (nt)	161600676
Raw reads	101,071,384
Assembled transcripts	
Number	209,860
Median length (nt)	848
Average contig	1294.39
N50 statistics (nt)	1712
GC%	39.30
Unigenes	
Number	129,869
Mean sequence length (nt)	1244
N50 statistics (nt)	1606
Longest/shortest contig	17362/500
Number of sequences > 1 K (nt)	51,352 (39.5%)
Number of sequences > 10 K (nt	57 (0.0%)
Unigenes annotation	
Full length	11587 (8.9%)
Quasi full length	10660 (8.2%)
Partial	81,224 (62%)
ORF with start codon	93,033
ORF with stop codon	114,956
GO hits	41,647
SwissProt hits	35,199
TrEMBL hits	56,946
Pfam	40,386
TFs	20,064
KEGG hits	13,625

Smouse, 2012).

2.6. Cross species transferability

For cross species transferability analysis, SSRs were also screened through MISA from *T. baccata* genome assembly retrieved from oneKP (https://sites.google.com/a/ualberta.ca/onekp/) which is a consortium responsible for sequencing of over 1000 plants. The identified SSRs were then characterized on the selected genotypes of *T. contorta* using non denaturing PAGE.

3. Results

3.1. Assembly

101,071,384 raw reads were assembled into 209,860 transcripts using Trinity. Following removal of the transcripts less than 500 nt length and as well as redundant sequences, 129,869 unigenes were retained for downstream analysis. Mean sequence length was found to be 1244. The N50 value of 1606, L50 of 29,069 and GC content of 39.22% was observed (Table 1). Read representation using Bowtie2 showed that 92.09% of the reads successfully mapped back to the assembly. The Number of proteins representing nearly full length transcripts and having an alignment coverage of ≥80% using blast against SwissProt database was found to be 6854. Completeness assessment using BUSCO showed that out of 1440 queried genes, the number of core genes detected were 1214 (84.31%).

3.2. Annotations

Using Annocript, out of 129,869 transcripts, the total number of sequences having at least one blast hit against viridiplantae were found to be 59,354. Thus only these sequences were available for annotation by Annocript. The transcripts without any blast hits thereby remaining unannotated may be considered as novel to *T. contorta*. Using blastx, the Annocript yielded a total of 35,199 hits in SwissProt corresponding to 342 organisms, 56,946 hits in TrEMBL corresponding to 511 organisms. 6219 sequences were observed to be long non-coding RNAs under 0.95 probability and 100 nucleotides of maximum length of ORF. Using TRAPID pipeline, out of the 129,869 transcripts in the assembly, 26,398 (20.3%) were Meta annotated as full length, 11,588 (8.9%) as quasi full length, 10,659 (8.2%) as partial and 81,224 (62.5%) with no Meta annotation information (supplementary file S3). Further, 93,033 (71.6%) of our transcripts have the open reading frame with start codon while 114,956 (88.5%) transcripts bear stop codon. Also, using Plaza Database, TRAPID assigned the transcripts into gene families. A total of 50,628 (39%) transcripts were assigned to 6854 gene families of which HOM000056 constitutes the largest family accounting about 2316 transcripts. The GO enrichment analysis of this family of transcripts showed that there are 24 categories enriched in Molecular function, 24 in biological processes and 12 in cellular components. In protein domain enrichment analysis, around 22 categories were found as enriched in this family. 1714 transcripts were found as single copy gene families. Overall, we retrieved around 41,647 (32.1%) transcripts with 5831 associated unique GO terms (supplementary file S4). GO:0005488 (binding), GO:0008152 (metabolic processes), GO:0009987 (cellular processes), GO:0003824 (catalytic activity) and GO:0044238 (primary metabolic process) are the most frequent GO terms associated with our transcripts. There were 3067 GO terms in Biological Processes, 559 in Cellular components and 2205 in Molecular functions (Fig. 1).

Further, alignment against KEGG database through KASS server yielded 13,625 transcripts with 3721 uniquely assigned KO numbers (supplementary file S5). Pathway mapping showed that most of the KO numbers belong to metabolic pathways followed by biosynthesis of secondary metabolites while isoflavonoid biosynthesis was found to be least frequent. A search in Pfam database yielded 40,368 transcripts with corresponding 2163 unique Pfam Ids (Supplementary file S6). About 325 domains were found as single transcript domains, thus they represent the least abundant domains in *T. contorta*. PF00078 (rvt), PF00069 (Pkinase), PF00665 (rve), PF01535 (PPR) and PF00067 (p450) were the largest families observed in our transcripts (supplementary file S9). About 20,064 sequences were identified as transcription factors (supplementary file S7). These TFs belonged to about 57 unique TF families. Among them bHLH, MYB_related, NAC, ERF and ARF were the most prominent TF families (Supplementary figure S8).

3.3. SSR screening and identification

Screening of the transcripts for microsatellite containing loci showed that about 6507 sequences contain 7041 SSRs out of which 534 were in compound form and 800 sequences contained > 1 SSR. An overall density of 1 SSR/22.95 kb of the sequences was determined. Distribution of different repeat classes and abundant identified motifs is shown in Fig. 3. Among the dinucleotide repeats, TA/AT repeats were found in greater frequency (23.83%) followed by AG/GA (9.06%), CT/TC (8.3%), GT/TG (8.90%) and AC/CA (6.65%). GC/CG repeats were found to be in a very low frequency of 0.14%. We observed 60 types of different trinucleotide repeat motifs contained in 2618 SSRs among which TTC repeat comprise the largest fraction (2.98%) followed by TCT (2.20%) and ATT (2.07%). A low percentage (5.4%) of tetra, penta and hexa nucleotide repeat motifs were observed in overall identified SSRs. CTGC, GGGAC and TTCCTC were among the dominant motifs in tetra, penta and hexa nucleotide repeat motifs respectively. 7.5% of the SSRs were present in compound formation with a maximum of 100 bases between two SSRs. The largest SSR observed was in compound formation with AGG, TCC and CCT repeated 7 times with interrupting sequences between them.

Positional distribution of the SSRs in the transcripts was analysed by predicting the ORFs from the SSR containing sequences, then the start and end positions of SSRs were correlated with the start and stop positions of the predicted ORFs. Among the 6496 sequences with predicted ORFs, about 1781 (27.41%) sequences have SSRs in their CDS region, 1711 (26.33%) sequences have SSR in 5`UTR while 2919 (44.93%) sequences have SSR in 3`UTR region and 85 (1.305%) have

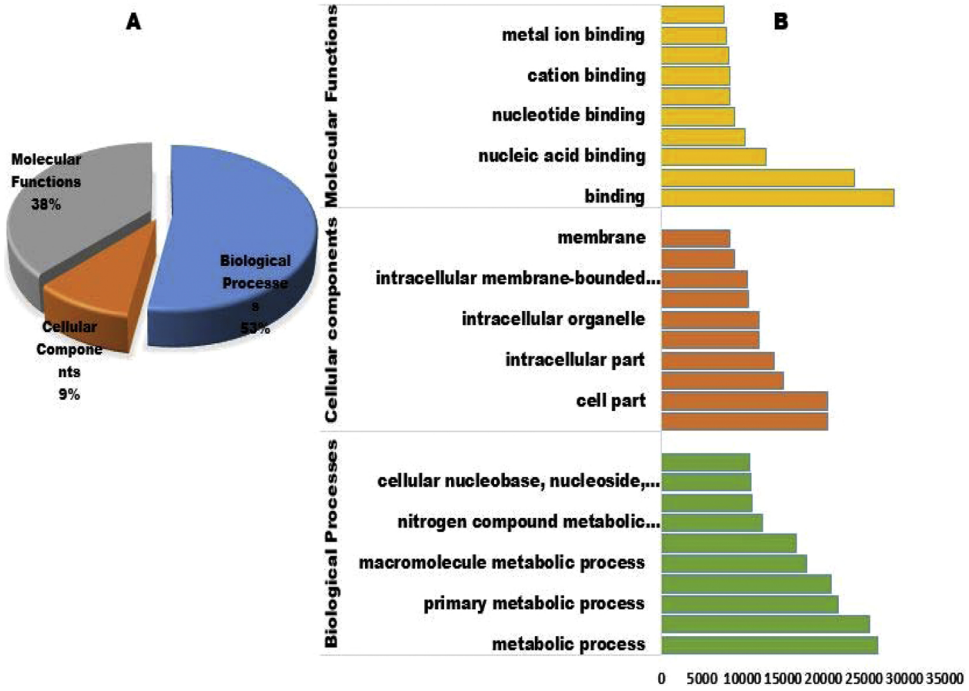

Fig. 1. GO annotation of transcripts. A) Distribution of different GO classes. B) Top GO terms in Biological Processes, Cellular components and Molecular Functions.

SSRs partially in CDS and partially in UTR regions. 11 sequences were found without ORFs.

We further examined the sequences containing SSRs in the CDS region. CDS region contains triplet codons which are translated into amino acids. Among the 1781 sequences having SSRs in the CDS region, we identified 934 (62%) sequences containing 1118 trinucleotide repeat containing SSRs. Among them, 139 sequences contain more than one SSR and 52 were in compound formation. A search for triplet codons in these SSRs showed that GAA is the most frequent codon followed by GGA and AGA while ACC and ACG are least frequent. Most of the amino acids specified by the identified triplet codons were of polar nature with neutral amino acids being dominant. No aromatic or acidic amino acids were observed. Of the overall sequences having SSR in CDS, the amino acid arginine appears to be dominant followed by glycine and glutamine while histidine is least dominant.

For cross species transferability, the genome assembly of *T. baccata* was screened for the identification of SSR regions. Out of 45,495 examined sequences, a total of 1099 sequences were found to contain SSRs. Overall, 1235 SSRs were identified. About 116 sequences were found to contain more than one SSR and 62 SSRs were found to be in compound formation.

3.4. SSR characterization

A total of 100 SSR markers (80 form *T. contorta* and 20 from *T. baccata*) were tested for characterization on 30 genotypes of *T. contorta* comprising three populations (supplementary file 1). Around 49 SSRs (61.25%) developed from transcriptomic sequences of *T. contorta* showed successful amplification among which 27 markers showed polymorphism while 22 revealed monomorphism. 15 out of 20 markers designed from genomic sequences of *T. baccata* showed successful amplification across *T. contorta* genotypes with 10 as polymorphic and 5 as monomorphic. Thus a total of 75% cross transferability rate was achieved between the *T. baccata* and *T. contorta*.

Of all the characterized loci, 37 (27 and 10 from *T. contorta* and *T. baccata* respectively) were polymorphic with ≥ 3 alleles, 4 comprised of 2 alleles whereas 22 were monomorphic. Only 37 polymorphic loci were analysed for downstream analysis. A total of 214 alleles were observed in the 30 genotypes. The number of alleles (Na) ranged from 3 to 11 with mean value of 5.78 and effective number of alleles (Ne) ranges from 1.84 to 8.73 with mean value of 3.64. Ho (observed heterozygosity) and He (expected heterozygosity) ranged from 0.066 - 0.96 and 0.46 - 0.90 with mean of 0.57 and 0.69 respectively. Average Shannon's Information Index (I) was found to be 1.38. Polymorphic Information Content (PIC) was found to be 0.64 with 0.41 and 0.87 as upper and lower limits. 25 loci showed significant deviations from HWE. Marker characters are summarized in Table 2.

3.5. Diversity characterization

Since the SSR markers were mostly designed from the RNA sequences, therefore they also have the potential to reveal the underlying functional diversity because RNA represents the functional elements of the genome. Biological functional diversity of these markers is presented in Fig. 2. Population diversity analysis reveal that pop 1 (Uttrakhand) has greater genetic diversity, Ho = 0.6 than pop 2 (Himachal Pradesh), Ho = 0.55 and pop 3 (J&K), Ho = 0.57. Shannon Informational Diversity Statistics (Fig. 4) partitioned by populations and total, showed that there lies 87% diversity within the populations in contrast to 13% among the populations. There appears an overlap of 25% within populations coupled to 75% overlap among the populations. An indication of slight inbreeding among the individuals of populations is revealed by positive value of inbreeding coefficient (Fis = 0.09). Mean allelic pattern across the populations reveal that pop 3 has highest number of effective alleles (Ne = 3.2) followed by pop 2 (Ne = 3.1) and pop 1 (Ne = 2.9). From a total of 50 population specific alleles, pop 3 exhibits highest number of mean private alleles (0.48) followed by pop 2 (0.45) and pop 1(0.40).

4. Discussion

T. contorta is an endangered species lacking reference genome. At the genomic and transcriptomic level, the species remains uncovered which appeal to turn our attention towards its exploration at the genomic and transcriptomic level. This would facilitate the discovery and exploration of novel genomic elements and functions and also

Table 2
Characteristics of 37 Polymorphic markers of *Taxus contorta*.

	ID	Sequence	Motif	Obs. Size range	Ta	PIC value	Na	Ne	Ho	He
1	Tx32	F:GGTCGGTGGGTATGTGTGTT R:CTGGGTCTGCACTTGATTTC	(AG)17	148-156	52	0.775	6	5.0992	0.7	0.8175
2	Tx35	F:CAAGATGGCGTTTACAGAGC R:GCTTCTTACTGTGCCCCATAG	(AG)9	154-160	54	0.649	4	3.3835	0.7667**	0.7164
3	Tx41	F:TGTCGTCGGATGGTCTGTAA R:CACGAGGGCACACCTATGTT	(AG)9	150-154	52	0.5	4	2.236	0.3**	0.5621
4	Tx43	F:CCGCTTGTCTCAGAATGGAC R:GAGGAGGAAGAAGAAGAGGAAG	(CTC)9	133-145	52	0.487	6	2.3622	0.9333**	0.5864
5	Tx44	F:TCAAGAGCCTTGTTCTGCTG R:CCCGTGGAAAGATACCAAAC	(CTG)7	150-165	52	0.789	7	5.3892	0.8667	0.8282
6	Tx45	F:CCTGCAAATGAAAGCCTGAG R:CGCTTGCCAACAAGTCTAC	(AG)9	154-162	52	0.592	6	2.8257	0.9333*	0.6571
7	Tx47	F:CCCCAGTTCATTGCAACTAC R:TGCGTGTTCACATCTGTGTC	(GA)8	178-180	47	0.545	3	2.6667	0.8333	0.6356
8	Tx52	F:GGTTATGCATACAGGCAAGTGG R:AGGGTCCCTTAACCCAGGTA	(GA)17	184-190	49	0.712	5	4.1002	0.5	0.7689
9	Tx56	F:AGACCACTCATTCGCTAC R:GGGAGTTATGGCTCTTTG	(GA)11	144-150	47	0.689	5	3.7815	0.6667**	0.748
10	Tx58	F:AGGGTCCAGCGTTACAAATG R:GTGGCTGATTTCCCTCTTCC	(GA)15	94-108	47	0.765	7	4.8387	0.6	0.8068
11	Tx63	F:GGTCATATGGCAAGGGAATG R:GCTTAACATCTAGCCGCTGGT	(GGC)5	118-130	49	0.742	6	4.4888	0.6*	0.7904
12	Tx071	F:TGAAGAGGCCTTGCAGATG R:TCATTGAGCCGTTTGTAGGC	(AG)9	176-184	46	0.688	6	3.666	0.8667	0.7395
13	Tx72	F:CACACAAGGACCCAACTTGA R:GGGAAACCCATTGGCTAAGT	(AC)9	144-150	48	0.592	5	2.8662	0.9667*	0.6621
14	Tx73	F:GGCATCAGTGTCAAGACGAA R:CGTTGGAGACTGTTGCTTCA	(GA)24	150-156	48	0.748	5	4.6036	0.5*	0.796
15	Tx74	F:GAAAAGTATGGGCGGCTA R:GGGGTCTTGGTAAGAATC	(GT)24	184-194	43	0.749	7	4.534	0.6667	0.7927
16	Tx79	F:GCCTAGTGCCCCTCATATGTT R:GGCGGGTTTAATGTCTCTGA	(TCT)11	146-150	48	0.504	4	2.3715	0.6667*	0.5881
17	Tx83	F:CAACCATGATACACCGCTTG R:GAGCAAGAATGATGCTGAGG	(CTG)12	169-282	48	0.585	6	2.651	0.4*	0.6333
18	Tx86	F:GAGCAATTGGGTCAGATTCC R:CAGGGCCACATTTCTCTTCT	(GAG)8	141-153	48	0.682	6	3.5503	0.4667*	0.7305
19	Tx91	F:TTCACCCTAAGCCCAGATTG R:GAGAAGGGAAGCTCGGAAAT	(GA)12	112-116	48	0.443	4	2.1102	0.0667**	0.535
20	Tx105	F:ATAACGGGCTCCCAAGTAGG R:CACCGTGGAGACTTTGTTGA	(AG)10	178-184	48	0.6	5	2.8846	0.3667	0.6644
21	Tx109	F:ACCCGGGACCTCTTGTTAAT R:CTTGCATTTCCACTCCCACT	(GA)12	132-140	48	0.608	6	2.8846	0.5333**	0.6644
22	Tx110	F:CTCCACAGCAAATCCGTA R:ATGATGCATTGTCCCTAGCC	(GT)17	130-174	46	0.674	6	3.5785	0.2333	0.7328
22	Tx117	F:GCCTCAAAAGTCCTCTCAAC R:CTTTCAGGGCCAATCGAA	(TG)15	182-196	46	0.804	9	5.7325	0.6667	0.8395
24	Tx121	F:GCACCCATAGATTACCAGCA R:GGAAGGAGAGTCAGCTAGTGG	(GA)18	160-182	50	0.81	8	5.8824	0.7	0.8441
25	Tx122	F:TACCTGCCTATGGGAAGCAC R:TGCAGGAGAGTCAGCTAGTGG	(GA)11	170-184	50	0.737	8	4.2857	0.6	0.7797
26	Tx127	F:ACTGGCATTCGCTCCATTAG R:AGGATGCTGGACTGAGAACG	(AG)15	150-158	48	0.665	6	3.3708	0.4333	0.7153
27	Tx152	F:CGTCATACCACGGCCTTTAC R:AAAGCTCTGGTTCTGTCTCC	(GA)20	178-196	48	0.875	11	8.7379	0.8	0.9006
28	Tx2G	F:AATTGTTCATATGGTTCATGC R:GGTTTGGCTTTTTATAGGTTT	(AC)9	200-210	43	0.803	7	5.7692	0.6667**	0.8407
29	Tx5G	F:CTTGTCTGAAAATCAAGCACT R:CACGGTAATATAGAGCAGGAA	(AG)9	162-164	43	0.565	3	2.7607	0.4667*	0.6486
30	Tx6G	F:GCCAAGCTTGACTAGAATACA R:AGCAGATAACACCAAATTCAA	(AC)11	156-160	43	0.699	4	3.9474	0.4333**	0.7593
31	Tx8G	F:AGAAACACGAAACATGTCAGT R:GCCTGAACATCATGAGTAGAC	(AG)10	156-160	46	0.416	4	1.8614	0.2*	0.4706
32	Tx10G	F:ATATTTATATGGCCACACAGC R:CTTGTTGTCTCTCCCTCTCTC	(GA)8	142-152	46	0.727	7	4.2553	0.5333	0.778
33	Tx11G	F:TGTGACAAACTAAGAAGCAAA R:TTGCCTGTAATCTAATCTCCA	(CT)11	158-164	39	0.431	5	1.8499	0.1667*	0.4672
34	Tx12G	F:TCTCAACATCCCAAACTTAGA R:GAAAAGACCAAGGATCAAATC	(TC)12	142-150	43	0.539	6	2.5937	0.2**	0.6249
35	Tx15G	F:CTAGTTCCAAACCACCATTTT R:TGTTGTAGTAATGATCCACGA	(AG)22	158-162	46	0.48	4	2.1429	0.5667	0.5424
36	Tx16G	F:CGCGTCGTACTTATTTGTTAG R:GTATGGTCAAATCACTCCAAA	(AG)18	150-156	46	0.447	4	2.0619	0.7333	0.5237

(*continued on next page*)

Table 2 (*continued*)

	ID	Sequence	Motif	Obs. Size range	Ta	PIC value	Na	Ne	Ho	He
37	Tx20G	F:TAATTGCAAGAGATTCGGATA R:TTCTTTTTCTCCTTCTCCTTC	(AAG)12	160-181	43	0.627	9	2.8391	0.7667	0.6588

Ta = annealing temperature, PIC = polymorphic information content, Na = number of alleles, Ne = effective number of alleles, Ho = observed heterozygosity, He = Expected heterozygosity, *, ** and *** depict significant deviations from HWE at 0.5, 0.01 and 0.001 probability respectively.

decipher the unassessed genetic variation in nature. Next generation sequencing (NGS) has revolutionized the field of genomics and made it possible to generate large scale sequence data at a brisk pace and relatively much decreased cost (Koboldt et al., 2013; Ellegren, 2014). Taking the advantage of this technology, genomic and transcriptomic studies are now possible for non-model species as well. Using massively parallel sequencing, the transcriptome of *T. contorta* was sequenced. Through *de novo* assembly, we provided a reference transcriptome for this species. More than one lakh unigenes were generated with ≥ 500 nt length. Further assessment revealed that the assembly had a very high quality owing to 92.09% read mapping with Bowtie2, 84.81% of queried genes recovered from completeness assessment by BUSCO and recovery of 6854 full length transcripts from SwissProt which is better than achieved for *Rhododendron arboreum* (Choudhary et al., 2018). Further, 63% success rate of the SSR primers also compliments the quality of the assembly. N50 value of 1712 obtained in this study is much higher than achieved in a comparative transcriptomic analysis of different Taxus species by Saxena and Karvadi (2015) wherein an N50 value of 318, 868 and 1079 was observed for *T. x media*, *T. cuspidata* and *T. baccata* respectively. The high quality transcriptome generated here will facilitate further genetic studies in this important species.

It is imperative to assign biological functions to the *T. contorta* unigenes owing to very scarce genomic data available for this species. A total of 45.7% transcripts were annotated during the functional annotation by Annocript. 31.08% and 32% of the unigenes were assigned Pfam IDs and GO terms respectively while 15.44% were identified as TFs. Since no reference transcriptome or draft genome is available for *T. contorta,* we believe that the functionally annotated reference transcriptome generated here would suffice the gap to some extent if not completely. The most frequent KEGG hit was found to be K04733 (supplementary file S10) corresponding to IRAK4 (interleukin-1 receptor-associated kinase 4) which is involved in signalling processes like MAP kinase and plays critical role in host defence mechanism in animals (Li et al., 2002). They have been found structurally and functionally similar to Receptor-like Kinases in plants which are remarkably expanded in plants. More than 400 such kinase genes were observed in *Arabidopsis* (Klaus-Heisen et al., 2011). Till now, there occurs no information about the TFs of *T. contorta* in the PlantTFDB. To the best of our knowledge this is the first report of the identification of the TFs in this species. A diverse array of TFs were identified here, which represent about 57 families. The identification of these TFs may promote the studies involving the regulation of gene expression. In view of lack of draft genome, it was needed to generate the transcriptomic data that could bridge the gap in exploring the genomic elements of this species and understand its functional and regulatory aspects.

Owing to the extreme threat to this species, it was urgently needed to elucidate its underlying population and landscape genetics. However, a robust polymorphic marker resource is a pre requisite for such analysis. In view of this, here we generated an efficient polymorphic SSR marker resource for this species which would prove useful to bridge the gap in its conservation planning and management. We identified 7014 SSR loci in *T. contorta* which is much greater than identified in *T. baccata* by Olsson et al (2018). To the best of our knowledge, the number of SSRs identified in this study is the largest number identified till date for any species of the genus, Taxus. 27.41% sequences were found to have SSRs in their CDS region in contrast to 71.26% in UTRs. This observation is corroborated by Qu and Liu (2013) who observed the density of SSRs to be highest in UTRs in maize which gradually diminishes towards coding regions. This lower ratio of SSRs in CDS to UTRs is a result of selection pressure against ORF change in CDS region that could cause potential frameshift mutations (Zhang et al., 2004).

Characterization of *T. contorta* SSR markers showed 61.25% success rate. A comparatively high proportion (27.5%) of SSRs showed monomorphism which is natural in case of genic markers as they lie within the coding regions, so are expected to have comparatively less variability than genomic markers (Olsson et al., 2018; Postolache et al., 2014). The greater proportion of monomorphism can also be attributed to the geographical isolation of the fragmented small patchy populations of *T. contorta*.

A moderate level of genetic diversity was observed in our study using a set of 37 markers. This is corroborated by different studies

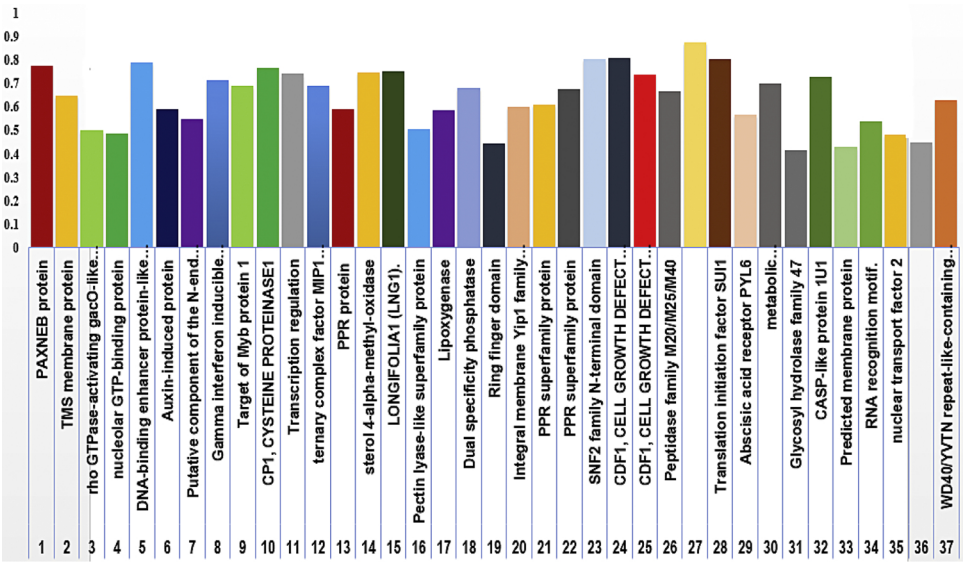

Fig. 2. Representation of functional diversity of the characterized polymorphic markers. Height of the vertical bars is proportional to their informativeness (0–1) based on PIC values. Horizontal numbers (1–37) represent the markers numbers in the same order as given in Table 2. The text against the vertical bars shows their putative biological function.

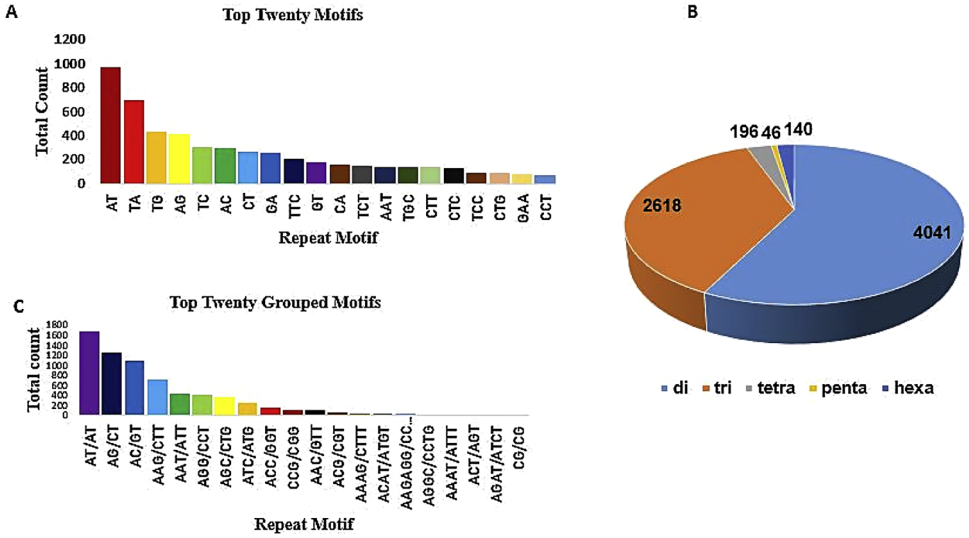

Fig. 3. Statistics of identified SSRs. A and C represent Top observed motifs and grouped motifs respectively while B represents the frequency of different repeat classes in identified SSRs.

which reveal inbreeding and low to moderate level of genetic diversity in various Taxus species (Wen et al., 2018; Litkowiec et al., 2018; Miao et al., 2016). However, observed heterozygosity (0.57) was found slightly lower than expected (0.69). This shows that there is less gene diversity as expected from HWE and some inbreeding may be involved which is revealed by positive value of Fis. The indication of inbreeding in Taxus species is revealed in different studies (Poudel et al., 2014; Myking et al, 2009; Dubreuil et al., 2010; Senneville et al., 2001; El-Kassaby and Yanchuk, 1994; Miao et al., 2014). The understory habitat of Taxus is considered as a factor promoting inbreeding as it severely restricts pollen and seed dispersal thereby favouring mating between relatives (Chybicki et al., 2012; Litkowiec et al., 2018). The populations of *T. contorta* have been rendered fragmented and isolated. Prolonged fragmentation and isolation of populations can lead to enhanced genetic erosion through drift (Ellstrand and Elam, 1993). Further, gene flow and immigration are greatly reduced (Couvet, 2002) and inbreeding is also accelerated in fragmented populations (Keller and Waller, 2002). Also, during our field observations we noticed very few individuals in many population. All these factors may explain the low heterozygosity observed in the present study.

We observed a substantial PIC value of 0.64 which reveals the high informativeness of the characterized markers. Out of the total allele diversity observed (I = 1.38), most of the diversity lies within the populations (87%) indicating a high gene flow between the members of individual populations. On the other hand populations show only 13% diversity among them indicating that there is a diversity overlap between populations which is revealed by scaled diversity overlap analysis that shows 75% overlap. Trees are often outcrossing which along with their comparatively greater power of gene flow and larger population sizes (Petit and Hampe, 2006) than the herbaceous species, tend to show greater within population diversity as compared to among population differentiation (Hamrick et al., 1992; Hamrick and Godt, 1996; Nybom, 2004). A higher rate (75%) of cross transferability between *T. contorta* and *T. baccata* was observed. Such higher transferability rate is common in congeners or closely related genera. A similar higher transferability (65–85%) rate was also achieved between the species of Glycine. However, transferability was lowered down to 3–13% outside the genus. This is in contrast to animas where the transferability achieved is very low (Peakall et al., 1998). Transferability success is directly proportional to the evolutionary distance

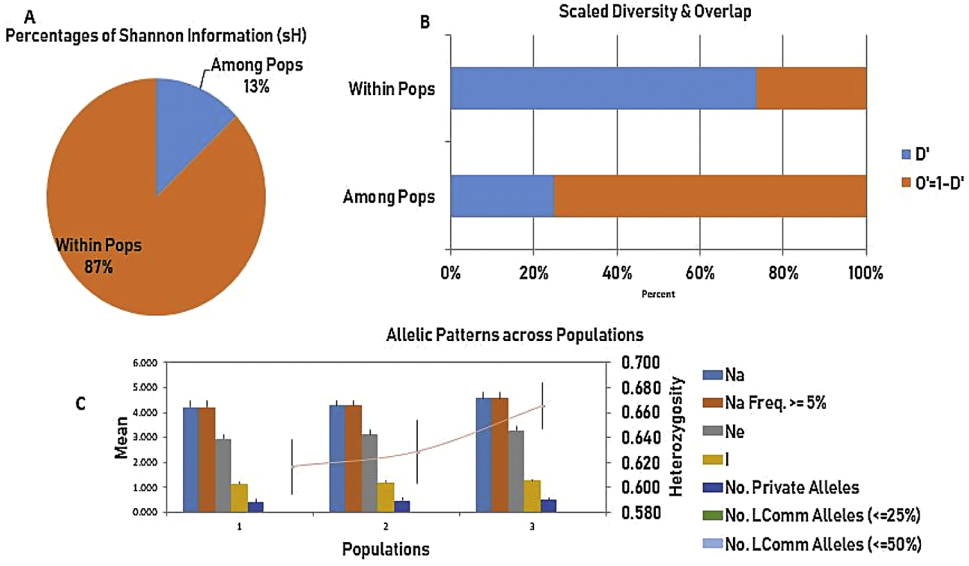

Fig. 4. Graphical representation of Shannon's Information (A), Scaled diversity and overlap (B) and Allelic pattern across populations (C).

(Liewlaksaneeyanawin et al., 2004). A complete transferability between the species of rubber was achieved which decreased to 73.3% in cross genera analysis (Feng et al., 2009).

The SSR markers characterized here represent the functional elements of the genome. The markers linked to functional elements of a genome are important as they can give a direct estimate of functional diversity among the genotypes. The characterized markers, represent the important proteins or enzymes playing their part in specific biological processes, so these markers not only have the potential to reveal the underlying genetic diversity but also show the variation at the functional level among genotypes. The polymorphic markers were assigned putative biological functions. Their biological role ranges from signalling to gene expression control (Fig. 2). Tx45 and Tx5G are involved in signalling pathways of auxin and abscisic acid respectively. Tx32, Tx56, Tx63, Tx2G play their part in complex gene expression processes. Other markers showing important biological functions include Tx73 and Tx6G which have role in lipid biosynthesis and metabolic processes respectively. However, the functional diversity was found to be lower than expected. This is quite obvious because the functional elements are generally highly conserved and do not show much variation.

5. Conclusion

We provided a high quality reference transcriptome for *T. contorta* that can prove useful for further genetic studies of the species. We also generated a robust polymorphic and functionally diverse SSR marker resource for this species. These markers are highly informative and their number is sufficient for any population and landscape genetic analysis of this species. Such marker resource can act as a valuable tool for deciphering the pattern of genetic variation, genetic structure and other parameters which are required for planning conservation strategies for an endangered species. Since *T. contorta* is an endangered species with isolated and fragmented populations, immediate need is to focus our attention towards its conservation for which our study will prove very useful.

Data archiving statement

The raw sequencing data was submitted to NCBI under the SRA accession numbers of SRR8130348 and SRR8130349.

Author contribution

PB conceived and organized the study. AM carried out sampling, all the wet lab experiments and bioinformatic analysis and wrote the manuscript. AS and SC participated in sampling and nucleic acid isolation. PB further edited and coordinated in finalizing the manuscript. All authors have carefully read and approved the manuscript. The authors declare no conflict of interest.

Acknowledgements

This study was financially supported by MoEF&CC (Ministry of Environment, Forests and Climate Change), India under the grant NMHS/SG-2016/011. AM acknowledges CSIR New Delhi for their financial assistance. The authors also acknowledge Mr. Amit Singh for his valuable support in writing customized python scripts.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.indcrop.2019.111600.

References

Ahmed, E., Arshad, M., Ahmad, M., Saeed, M., Ishaque, M., 2004. Ethno pharmacological survey of some medicinally important plants of Galliyat Areas of NWFP, Pakistan. Asian J. Plant Sci. 3 (4), 410–415. https://doi.org/10.3923/ajps.2004.410.415.

Banskota, A.H., Nguyen, N.T., Tezuka, Y., Nobukawa, T., Kadota, S., 2006. Hypoglycemic effects of the wood of *Taxus yunnanensis* on streptozotocin-induced diabetic rats and its active components. Phytomedicine 13 (1-2), 109–114. https://doi.org/10.1016/j.phymed.2004.01.015.

Beier, S., Thiel, T., Münch, T., Scholz, U., Mascher, M., 2017.). MISA-web: a web server for microsatellite prediction. Bioinformatics 33 (16), 2583–2585. https://doi.org/10.1093/bioinformatics/btx198.

Cheng, B.B., Sun, Q.W., Zheng, Y.Q., 2015. Development of microsatellite loci for *Taxus wallichiana var. Wallichiana* (Taxaceae) and cross-amplification in Taxaceae. Genet. Med. Res. 14 (4), 16018–16023. https://doi.org/10.4238/2015.December.7.15.

Chybicki, I.J., Oleksa, A., Kowalkowska, K., 2012.). Variable rates of random genetic drift in protected populations of English yew: implications for gene poolconservation. Conserv. Genet. 13 (4), 899–911.

Choudhary, S., Thakur, S., Najar, R.A., Majeed, A., Singh, A., Bhardwaj, P., 2018. Transcriptome characterization and screening of molecular markers in ecologically important Himalayan species (*Rhododendron arboreum*). Genome 61 (6), 417–428. https://doi.org/10.1139/gen-2017-0143.

Couvet, D., 2002. Deleterious effects of restricted gene flow in fragmented populations. Conserv. Biol. 16 (2), 369–376. https://doi.org/10.1046/j.1523 1739.2002.99518.x.

Doyle, J.J., Doyle, J.L., 1990. Isolation of plant DNA from fresh tissue. Focus 12 (13), 39–40.

Dubreuil, M., Riba, M., González-Martínez, S.C., Vendramin, G.G., Sebastiani, F., Mayol, M., 2010. Genetic effects of chronic habitat fragmentation revisited: strong genetic structure in a temperate tree, *Taxus baccata* (Taxaceae), with great dispersal capability. Am. J. Bot. 97 (2), 303–310.

El-Kassaby, Y.A., Yanchuk, A.D., 1994. Genetic diversity, differentiation, and inbreeding in Pacific yew from British Columbia. J. Hered. 85 (2), 112–117.

Ellegren, H., 2014. Genome sequencing and population genomics in non-model organisms. Trends Ecol. Evol. (Amst.) 29, 51–63. https://doi.org/10.1016/j.tree.2013.09.008.

Ellstrand, N.C., Elam, D.R., 1993.). Population genetic consequences of small population size: implications for plant conservation. Annu. Rev. Ecol. Syst. 24 (1), 217–242. https://doi.org/10.1146/annurev.es.24.110193.001245.

Feng, S.P., Li, W.G., Huang, H.S., Wang, J.Y., Wu, Y.T., 2009. Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (*Hevea brasiliensis*). Mol. Breed. 23 (1), 85–97. https://doi.org/10.1007/s11032-008-9216-0.

Gajurel, J.P., Cornejo, C., Werth, S., Shrestha, K.K., Scheidegger, C., 2013. Development and characterization of microsatellite loci in the endangered species *Taxus wallichiana* (Taxaceae). Appl. Plant Sci. 1 (3), 1200281. https://doi.org/10.3732/apps.1200281.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Chen, Z., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29 (7), 644–652. https://doi.org/10.1038/nbt.1883.

Hamrick, J.L., Godt, M.J.W., Sherman-Broyles, S.L., 1992. Factors influencing levels of genetic diversity in woody plant species. Population Genetics of Forest Trees. Springer, Dordrecht, pp. 95–124. https://doi.org/10.1007/BF00120641.

Hamrick, J.L., Godt, M.J.W., 1996. Effects of life history traits on genetic diversity in plant species. Philos. Trans. R. Soc. Lond. B Biol. Sci. 351 (1345), 1291–1298. https://doi.org/10.1098/rstb.1996.0112.

Huang, L., Deng, X., Li, R., Xia, Y., Bai, G., Siddique, K.H., Guo, P., 2018. A fast silver staining protocol enabling simple and efficient detection of SSR markers using a non-denaturing polyacrylamide gel. J. Visual. Exp.: JoVE 134. https://doi.org/10.3791/57192. e57192.

Jin, J., Tian, F., Yang, D.C., Meng, Y.Q., Kong, L., Luo, J., Gao, G., 2016. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucl. Acids Res. https://doi.org/10.1093/nar/gkw982. gkw982.

Juyal, D., Thawani, V., Thaledi, S., Joshi, M., 2014. Ethnomedical properties of *Taxus wallichiana* Zucc.(Himalayan yew). J. Tradit. Complement. Med. 4 (3), 159–161. https://doi.org/10.4103/2225-4110.136544.

Kejani, A.A., Tafreshi, S.A.H., Nekouei, S.M.K., Mofid, M.R., 2010. Efficient isolation of high quality nucleic acids from different tissues of *Taxus baccata* L. Mol. Biol. Rep. 37 (2), 797. https://doi.org/10.1007/s11033-009-9607-2.

Keller, L.F., Waller, D.M., 2002. Inbreeding effects in wild populations. Trends Ecol. Evol. (Amst.) 17 (5), 230–241. https://doi.org/10.1016/S0169-5347(02)02489-8.

Khan, M., Verma, S.C., Srivastava, S.K., Shawl, A.S., Syamsundar, K.V., Khanuja, S.P.S., Kumar, T., 2006. Essential oil composition of *Taxus wallichiana*Zucc. from the Northern Himalayan region of India. Flavour Fragr. J. 21 (5), 772–775. https://doi.org/10.1002/ffj.1682.

Klaus-Heisen, D., Nurisso, A., Pietraszewska-Bogiel, A., Mbengue, M., Camut, S., Timmers, T., Lefebvre, B., 2011. Structure-function similarities between a plant receptor-like kinase and the human interleukin-1 receptor-associated kinase-4. J. Biol. Chem. https://doi.org/10.1074/jbc.M110.186171. jbc-M110.

Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., Mardis, E.R., 2013. The next-generation sequencing revolution and its impact on genomics. Cell 155 (1), 27–38. https://doi.org/10.1016/j.cell.2013.09.006.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9 (4), 357–359. https://doi.org/10.1038/nmeth.1923.

Liewlaksaneeyanawin, C., Ritland, C.E., El-Kassaby, Y.A., Ritland, K., 2004. Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. Theor.

Appl. Genet. 109 (2), 361–369. https://doi.org/10.1007/s00122-004-1635-7.
Li, S., Strelow, A., Fontana, E.J., Wesche, H., 2002. IRAK-4: a novel member of the IRAK family with the properties of an IRAK-kinase. Proc. Natl. Acad. Sci. 99 (8), 5567–5572. https://doi.org/10.1073/pnas.082100399.
Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22 (13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.
Litkowiec, M., Lewandowski, A., Wachowiak, W., 2018. Genetic variation in *Taxus baccata* L.: a case study supporting Poland's protection and restoration program. For. Ecol. Manage. 409, 148–160.
Liu, J., Gao, L.M., Li, D.Z., Zhang, D.Q., Möller, M., 2011. Cross-species amplification and development of new microsatellite loci for *Taxus wallichiana* (Taxaceae). Am. J. Bot. 98 (4), e70–e73. https://doi.org/10.3732/ajb.1000445.
Marshall, T.C., Slate, J., Kruuk, L.E.B., Pemberton, J.M., 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7, 639–655. https://doi.org/10.1046/j.1365-294x.1998.00374.x.
Miao, Y.C., Lang, X.D., Zhang, Z.Z., Su, J.R., 2014. Phylogeography and genetic effects of habitat fragmentation on endangered *Taxus yunnanensis* in southwest China as revealed by microsatellite data. Plant Biol. 16 (2), 365–374.
Miao, Y.C., Zhang, Z.J., Su, J.R., 2016. Low genetic diversity in the endangered *Taxus yunnanensis* following a population bottleneck, a low effective population size and increased inbreeding. Silvae Genet. 65 (1), 59–66.
Min, X.J., Butler, G., Storms, R., Tsang, A., 2005. OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Res. 33 (suppl_2), W677–W680. https://doi.org/10.1093/nar/gki394.
Mohapatra, K.P., Sehgal, R.N., Sharma, R.K., Mohapatra, T., 2009. Genetic analysis and conservation of endangered medicinal tree species *Taxus wallichiana* in the Himalayan region. New For. 37 (2), 109–121. https://doi.org/10.1007/s11056-008-9112-9.
Moeller, M., Gao, L.M., Mill, R.R., Li, D.Z., Hollingsworth, M.L., Gibby, M., 2007. Morphometric analysis of the *Taxus wallichiana* complex (Taxaceae) based on herbarium material. Bot. J. Linn. Soc. 155 (3), 307–335. https://doi.org/10.1111/j.1095-8339.2007.00697.x.
Molur, S., Walker, S., 1998. Conservation Assessment Management Plan Workshop Report for Selected Medicinal Plants of Northern, North-eastern and Central India. (CAMP). Organised by Zoo Outreach Organisation CBS, Coimbatore and NBFGR Lucknow, 22–26th Sep. 1997, p 156.
Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M., 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35 (suppl_2), W182–W185. https://doi.org/10.1093/nar/gkm321.
Musacchia, F., Basu, S., Petrosino, G., Salvemini, M., Sanges, R., 2015. Annocript: a flexible pipeline for the annotation of transcriptomes able to identify putative long non coding RNAs. Bioinformatics 31 (13), 2199–2201. https://doi.org/10.1093/bioinformatics/btv106.
Nisar, M., Khan, I., Simjee, S.U., Gilani, A.H., Perveen, H., 2008a. Anticonvulsant, analgesic and antipyretic activities of *Taxus wallichiana* Zucc. J. Ethnopharmacol. 116 (3), 490–494. https://doi.org/10.1016/j.jep.2007.12.021.
Nisar, M., Khan, I., Ahmad, B., Ali, I., Ahmad, W., Choudhary, M.I., 2008b. Antifungal and antibacterial activities of *Taxus wallichiana* Zucc. J. Enzyme Inhib. Med. Chem. 23 (2), 256–260. https://doi.org/10.1080/14756360701505336.
Nybom, H., 2004. Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. Mol. Ecol. 13 (5), 1143–1155. https://doi.org/10.1111/j.1365-294X.2004.02141.x.
Olsson, S., Pinosio, S., González-Martínez, S.C., Abascal, F., Mayol, M., Grivet, D., Vendramin, G.G., 2018. De novo assembly of English yew (*Taxus baccata*) transcriptome and its applications for intra-and inter-specific analyses. Plant Mol. Biol. 97 (4-5), 337–345. https://doi.org/10.1007/s11103-018-0742-9.
Peakall, R., Gilmore, S., Keys, W., Morgante, M., Rafalski, A., 1998. Cross-species amplification of soybean (*Glycine max*) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. Mol. Biol. Evol. 15 (10), 1275–1287. https://doi.org/10.1093/oxfordjournals.molbev.a025856.
Peakall, R., Smouse, P.E., 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28, 2537–2539. https://doi.org/10.1093/bioinformatics/bts460.
Petit, R.J., Hampe, A., 2006. Some evolutionary consequences of being a tree. Annu. Rev. Ecol. Evol. Syst. 37, 187–214. https://doi.org/10.1146/annurev.ecolsys.37.091305.110215.
Postolache, D., Leonarduzzi, C., Piotti, A., Spanu, I., Roig, A., Fady, B., Roschanski, A., Liepelt, S., Vendramin, G.G., 2014. Transcriptome versus genomic microsatellite markers: highly informative multiplexes for genotyping Abies alba Mill. And congeneric species. Plant Mol. Biol. Rep. 32, 750–760. https://doi.org/10.1007/s11105-013-0688-7.
Poudel, R.C., Moeller, M., Gao, L.M., Ahrends, A., Baral, S.R., Liu, J., Li, D.Z., 2012.). Using morphological, molecular and climatic data to delimitate yews along the Hindu Kush-Himalaya and adjacent regions. PLoS One 7 (10), e46873. https://doi.org/10.1371/journal.pone.0046873.
Poudel, R.C., Möller, M., Liu, J., Gao, L.M., Baral, S.R., Li, D.Z., 2014. Low genetic diversity and high inbreeding of the endangered yews in Central Himalaya: implications for conservation of their highly fragmented populations. Divers. Distrib. 20 (11), 1270–1284.
Purohit, A., Maikhuri, R.K., Rao, K.S., Nautiyal, S., 2001. Impact of bark removal on survival of *Taxus baccata* L.(Himalayan yew) in Nanda Devi biosphere reserve, Garhwal Himalaya, India. Curr. Sci. 586–590.
Qu, J., Liu, J., 2013. A genome-wide analysis of simple sequence repeats in maize and the development of polymorphism markers from next-generation sequence data. BMC Res. Notes 6 (1), 403. https://doi.org/10.1186/1756-0500-6-403.
Saikia, D., Khanuja, S.P.S., Shasany, A.K., Darokar, M.P., Kukreja, A.K., Kumar, S., 2000.). Assessment of diversity among *Taxus wallichiana* accessions from northeast India using RAPD analysis. Plant Genet. Resour. Newsl. 27–31.
Saxena, P., Karvadi, B., 2015. Comparative transcriptome analysis of non-model Taxus species for taxol biosynthesis. J. Chem. Pharm. Res. 7 (3), 800–805.
Senneville, S., Beaulieu, J., Daoust, G., Deslauriers, M., Bousquet, J., 2001. Evidence for low genetic diversity and metapopulation structure in Canada yew (*Taxus canadensis*): considerations for conservation. Can. J. For. Res. 31 (1), 110–116.
Shah, A., Li, D.Z., Möller, M., Gao, L.M., Hollingsworth, M.L., Gibby, M., 2008. Delimitation of *Taxus fuana* Nan Li & RR Mill (Taxaceae) based on morphological and molecular data. Taxon 57 (1), 211–222. https://www.jstor.org/stable/25065961.
Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31 (19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y., Vandepoele, K., 2013. TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. Genome Biol. 14 (12), R134.
Wen, Y., Uchiyama, K., Ueno, S., Han, W., Xie, W., Tsumura, Y., 2018. Assessment of the genetic diversity and population structure of Maire yew (*Taxus chinensis var. mairei*) for conservation purposes. Can. J. For. Res. 48 (5), 589–598. https://doi.org/10.1186/gb-2013-14-12-r134.
Yang, J.B., Li, H.T., Li, D.Z., Liu, J., Gao, L.M., 2009.). Isolation and characterization of microsatellite markers in the endangered species *Taxus wallichiana* using the FIASCO method. Hort Science 44 (7), 2043–2045.
Yeh, F.C., Yang, R.C., Boyle, T., Ye, Z.H., Mao, J.X., 1999. POPGENE, Version 1.32: the User Friendly Software for Population Genetic Analysis. Molecular Biology and Biotechnology Centre, University of Alberta, Edmonton, AB, Canada.
You, F.M., Huo, N., Gu, Y.Q., Luo, M.C., Ma, Y., Hane, D., Anderson, O.D., 2008. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinform. 9 (1), 253. https://doi.org/10.1186/1471-2105-9-253.
Zhang, L., Yuan, D., Yu, S., Li, Z., Cao, Y., Miao, Z., Tang, K., 2004.). Preference of simple sequence repeats in coding and non-coding regions of *Arabidopsis thaliana*. Bioinformatics 20 (7), 1081–1086. https://doi.org/10.1093/bioinformatics/bth043.
Zhang, X.M., Gao, L.M., Möller, M., Li, D.Z., 2009. Molecular evidence for fragmentation among populations of *Taxus wallichiana var. mairei*, a highly endangered conifer in China. Can. J. For. Res. 39 (4), 755–764. https://doi.org/10.1139/X09-003.