Indian Journal of Biotechnology
Vol 13, October 2014, pp 464-472

Utilization of EST-derived SSRs in the genetic characterization of *Artemisia annua* L. genotypes from Ladakh, India

Jitendra Kumar[1], Prasad Bajaj[2], Gyan P Mishra[1], Shashi Bala Singh[1], Harvinder Singh[2] and Pradeep K Naik[1]*

[1]Defence Institute of High Altitude Research, Defence R & D Organization, Leh 194 101, Jammu & Kashmir, India
[2]Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, Solan 173 215, India

Received 15 November 2012; revised 7 May 2013; accepted 19 June 2013

Artemisia annua L. is an important medicinal plant that produces substantial quantity of artemisinin, an antimalarial agent. In India, it grows wild in the Ladhak region and has harboured considerable variability over the years. EST-derived SSR markers were used to measure the genetic diversity among the *A. annua* germplasm collected from geographically separated Leh (11,500 ft) and Nubra (9600 ft) valleys (Ladakh, India). After analysing 68,974 non-redundant (of 3,60,906 available) ESTs of *A. annua*, 4,342 SSR markers were developed. On an average, one SSR was found per 8.9 kb of EST sequence with dinucleotide motifs in highest frequency (52.2%), followed by tri (42.4%), tetra (3.6%), hexa (1.2%) and pentanucleotide (0.6%) repeat types. A set of 16 primer pairs were designed by considering only the SSR-containing ESTs from the artemisinin biosynthetic pathway. In total, 38 alleles were identified from 13 polymorphic SSR loci, ranging from 1-7 alleles per locus and displayed moderate genetic diversity with an average of 0.24. It was found that the genetic diversity among individual from Nubra valley was narrower than that of Leh valley, suggesting the importance and feasibility of introducing elite genotypes from different origins for *Artemisia* germplasm conservation and breeding programmes.

Keywords: *Artemisia annua*, cluster analysis, expressed sequence tag (EST), genetic diversity, simple sequence repeat (SSR) marker

Introduction

Artemisia annua L. is an annual herb, native to Asia, and commonly found in many countries throughout Europe, North America, Central and South America. It is one of the most important medicinal plants from cold arid regions of India, especially in the Ladakh region. It is well adapted to high altitudes of 9600-11500 ft above mean sea level (MSL) and grows well under low temperature, nutrient deficiency and environmental stress to which they are exposed. Over the years, the *Artemisia* populations in the Ladakh region have developed considerable variability, necessitating the study of genetic diversity and characterization. This plant synthesizes and accumulates substantial quantities of many derivatives of a cadiene skeleton, including artemisinin (endoperoxide seco sesquiterpene lactones). Artemisinin is currently the most effective agent against multidrug resistant strains of *Plasmodium* species, the malarial parasites[1,2]. This property of *A. annua* has placed it among the top ten industrial medicinal plants of the modern world.

Chemical synthesis of the artemisinin is commercially nonviable, and efforts to produce artemisinin in engineered yeast cells as well as in tissue cultured cells have so far not been very fruitful[3]. Thus, the sole source of the drug is from the widely grown or cultivated plants[4]. Since the production of artemisinin varies among the genotypes, it has generated worldwide interest in studying the genetic diversity of *A. annua* populations, cloned variants, chemotypes, ecotypes and in the development of pure-line cultivars. For effective utilization and protection of plant genetic resources, the analysis of genetic diversity and relatedness between or within different genotypes is a prerequisite[5]. It also helps in developing DNA based molecular markers for identification of genotypes with better traits. However, so far reports are not available regarding the genetic characterization of this plant from Ladakh region and hence a detailed investigation is required.

Traditionally, the genetic diversity of the plant is assessed by various DNA fingerprinting methods using specific and non-specific DNA-based markers. Microsatellites or simple sequence repeats (SSRs) based DNA markers are short tandem repeats of 1 to 6 bp in length, evenly distributed in the genomes and

*Author for correspondence:
Tel: +91-1792-239227; Fax: +91-1792-245362
E-mail: pknaik1973@gmail.com

have small locus size[6]. Flanking sequences to specific SSR loci are generally conserved within a particular species, within a genus and sometimes even across related genera, which makes ease to design primers for individual SSR loci. The multiallelic nature, highly polymorphic, codominant inheritance and relative abundance of SSRs in the genomes facilitate high reproducible genotyping using polymerase chain reaction (PCR)[7]. The SSRs reside within the transcribed sequences can be identified from ESTs[8]. Recently, EST-derived SSRs (EST-SSRs) have received much attention due to the high availability and ease in mining of EST sequences using data mining techniques[9,10]. Aadvantages of EST-SSRs have been reported for a number of plant species, such as, *Medicago* species[11], maize[12], rye[13] and wheat[14], indicating that EST-SSR markers have potential for use in *A. annua* genetic diversity studies.

The objective of the present study was to determine the genetic relationships among several genotypes of *A. annua* from Ladhak region (the trans-Himalayan region of India). This information would be helpful in enriching the available genetic linkage map of *A. annua* and could help further in the development of high yielding genotypes.

Materials and Methods

Plant Materials

Ten plants each (~3-wk-old) of *A. annua* were selected from Leh (11,500 ft) and Nubra (9,600 ft) valleys of Ladakh region, which are separated from each other through natural mountain barrier (Fig. 1). Fresh leaf samples were collected from each selected plants of both the valleys and stored in laboratory at –20°C until further analysis. The average distance interval between samples was 100-200 m and the pair wise distance between valleys was 50-250 Km.

Fig. 1—Collection sites of 20 *A. annua* genotypes from two valleys (Leh and Nubra) and the two collection sites (Leh and Partapur) located in Ladakh (Jammu and Kashmir, India).

DNA Extraction

DNA was extracted from the leaf tissues based on a modified cetyltrimethylammonium bromide (CTAB) method described by Doyle and Doyle[15]. The yield of the extracted DNA and purity was checked by running the samples on 0.8% agarose gels along with standard uncut λ DNA marker (Biogene, USA).

Data Mining for SSR Marker

The available EST sequences (3,60,906) of *A. annua* were downloaded as FASTA format from the NCBI database (dbEST). The redundant ESTs were removed using BLASTCLUST (http://www.ncbi.nlm.nih.gov/BLAST) with sequence similarity 60 and 100% sequence coverage. After removing the redundant ESTs, the total numbers of ESTs left were 68,974. A Perl script, known as MIcroSAtellite (MISA; http://pgrc.ipk-gatersleben.de/misa/), was used to mine microsatellites. In the present study, SSRs were considered to contain motif one to six nucleotide in size. Frequency of SSRs refers to kilo-base pair of EST sequences containing one SSR.

PCR Primer Design and PCR Amplification

A set of 16 pairs of primers (Table 1) were designed by considering only the ESTs from the gene encoding enzymes for the artemisinin biosynthetic pathway. Primer-3 program (Whitehead Institute for Biomedical Research, Cambridge, MA, USA) was used for designing of primers based on following core criteria: (1) melting temperature (Tm) between 57°C and 65°C, with 60°C as optimum; (2) product size ranging from 90 bp to 600 bp; (3) primer length ranging from 18 bp to 24 bp with amplification rate larger than 80%; and (4) GC% content between 30 and 70%. All primer pairs were custome synthesized from GBiosciences (Geno Technology Inc., USA). PCRs were performed in a thermocycler (MyCycler™, Bio-Rad Laboratories, USA). The volume of PCR solution was 15 µL, containing 75 ng of template DNA, 1× PCR buffer (Mg^{2+} free), 0.5 U of *Taq* DNA polymerase, 300 µmol/L deoxynucleotide triphosphates (dNTPs), 2.25 mmol/L Mg^{2+}, and 0.75 µmol/L forward and reverse primers. The optimized PCR amplifying conditions used were: denaturation, 1 cycle of 3 min at 94°C, an

Table 1—Details of primers used for SSR amplification

Primer	Primer Sequence (5′ – 3′)	GC (%)	Tm (^{0}C)	Allele number	Product size range (bp)	PIC
1 F	ATAATACGCATGAGCTGGTTAG	40.9	56.5	2	90-380	0.245
1 R	CCACTACCAATCACAATAACAG	40.9	56.5			
2 F	GAATTGAGATTGTGGTCCTTAG	40.9	56.5	2	200-400	0.444
2 R	GGTTGCTAAGAATGTCGATTG	42.9	55.9			
4 F	ATCGTATTACCTTGGTCACATC	40.9	56.5	2	300-330	0.245
4 R	TGTCATACTGACTTACACAGGG	45.5	58.4			
5 F	TAAGCCAAAGGCTCAAGTAAAC	40.9	56.5	3	180-220	0.198
5 R	GGATTGCTCATCTAGTGCTTAT	40.9	56.5			
6 F	GCATGCATTTATGTTGGATCAC	40.9	56.5	2	100-400	0.565
6 R	CAGCAGCAACAACAACAACAG	47.6	57.9			
7 F	GGAACAGATGATCTATATGCCT	40.9	56.5	2	420-450	0.340
7 R	GCATACTATGTGCAAGGTCTAGT	43.5	58.9			
8 F	TGGTAGAACTCCACCTACTAACT	43.5	58.9	2	500-550	0.426
8 R	TATAATAGTTGGGTGGTTCCTC	40.9	56.5			
9 F	GAGAAAGAGAAAGCCAAACAC	42.9	55.9	1	220	0
9 R	TAGCTCCATAGATCTCAAACCT	40.9	56.5			
10 F	GGATCATTAAGTTACGCTCCT	42.9	55.9	2	300-350	0.495
10 R	CCATGCTTTATGTTGTAGAGTG	40.9	56.5			
13 F	GTAAGTTATACCTGGTTTCCAGC	43.5	58.9	4	300-600	0.595
13 R	ACCACTACACCTTGCATTCTA	42.9	55.9			
14 F	CTCTCTTCCTCTTTGTGTGTCT	45.5	58.4	4	150-500	0.503
14 R	CAAGATGGTACGAATACTGTTG	40.9	56.5			
15 F	CGAGCAATCGGAGAGTTAGC	55.0	59.4	7	250-550	0.802
15 R	ATGCATCTCGCGAATCTTCT	45.0	55.3			
16 F	GTGTGAGGCCTCTGCTCTG	63.2	61.0	5	120-480	0.716
16 R	ACCGCCATGTCTTCTCCATA	50.0	57.3			
Total				38		

annealing temperature of 55^{o}C for 35 cycles (45 sec at 94^{o}C, 30 sec at 55^{o}C, 1 min at 72^{o}C) and an additional cycle of 7 min at 72^{o}C. Each primer pair was screened twice to confirm the repeataility of the observed bands in each genotype. The amplified fragments were seperated on 6% metaphore IV agarose gels. The gels were electrophoresed at 100 V for 2 h in 1× TBE [Tris-borate-ethylenediaminetetraacettic acid (EDTA)] buffer. The banding patterns were scanned, scored and data were collected from reproducible bands. Band sizes were estimated by comparing with the molecular mass standards (100 bp λ ladder) included in each gel.

Data Collection and Analysis

The numbers of alleles detected and amplified by SSR markers were scored as present (1) or absent (0), each of which was treated as an independent character. The resulting matrix was used to estimate genetic similarity (GS) among all genotypes by Jaccard's similarity coefficient. The similarity matrix was subjected to cluster analysis by unweighted pair group method with arithmetic means (UPGMA) and a dendrogram was generated using the program NTSYSpc (numerical taxonomy and multivariate systems) version 2.01[16]. POPGENE software was used to calculate Nei's unbiased genetic distance among different genotypes with all markers. Data for observed number of alleles (Na), effective number of alleles (Ne), Nei's genetic diversity (H), Shannon's information index (I), number of polymorphic loci (NPL) and percentage polymorphic loci (PPL) were also analyzed[17]. Total genetic diversity (Ht)[18] was calculated within the species and within two major groups (as per their collection site) using POPGENE software. The data were subjected to a hierarchical analysis of molecular variance (AMOVA)[19], using two hierarchical levels; among valleys and among genotypes within each valley. GenAlEx software was used to calculate a principal coordinates analysis (PCA) that plots the relationship between distance

matrix elements based on their first two principal coordinates[20]. The allelic and genotypic frequencies were calculated for the samples analyzed. The genetic diversity of the samples as a whole was estimated based on the number of alleles per locus (total number of alleles/number of loci), the percentage of polymorphic loci (number of polymorphic loci/total number of loci analyzed) and polymorphism information content (PIC). The polymorphism was determined according to the presence or absence of the SSR locus. The value of PIC was calculated using the formula:

$$PIC = 1 - \sum_{i=1}^{n} p_i^2$$

Where p_i is the frequency of an individual genotype generated by a given EST-SSR primer pair and summation extends over *n* alleles.

Results

ESTs Containing Microsatellites

A total of 68,974 non-redundant ESTs from *A. annua* were used to evaluate for the presence of SSR motifs. A complete search of dinucleotide, trinucleotide, tetranuleotide, pentanucleotide and hexanucleotide microsatellite revealed the identification of 4,342 (6.3%) unique ESTs containing microsatellites. Of these, 499 (11.49%) ESTs contained more than one SSR and 40 (0.92%) were compound SSRs that have more than one repeat type (Table 2).

Frequency of SSRs Based on Motif Sizes

We examined 68,974 *A. annua* ESTs comprising of 44.12 Mb DNA sequences. The frequency of occurrence for SSRs was one SSR in every 8.943 kb of EST sequences. Analysis of SSRs motifs revealed that the proportions of SSR motif sizes were not

Table 2—Occurrence and number of repeats of 77 SSR motifs in *A. annua*

Repeats	Number of repeat unit										Total repeat	%
	5	6	7	8	9	10	11	12	13	Above		
AC/GT	617	195	103	107	47	22	12	13	4	26	1146	23.23
AG/CT	276	83	57	33	8	3	4	4	1	16	485	9.83
AT/AT	663	115	66	26	13	5	2	1	3	18	912	18.48
CG/CG	27	2	1	1	1					0	32	0.65
AAC/GTT	220	51	28	13	15	5	2		1	5	340	6.89
AAG/CTT	226	53	21	8	1		3			3	315	6.38
AAT/ATT	165	62	20	14	2	2	1	2		11	279	5.65
ACC/GGT	269	137	27	7	1	2				0	443	8.98
ACG/CTG	56	11	6	4	2			2		0	81	1.64
ACT/ATG	124	31	6	7	1	1	1	2		1	174	3.53
AGC/CGT	69	15	11	3	1					0	99	2.01
AGG/CCT	42	17	4		1	2				0	66	1.34
AGT/ATC	188	55	15	8	2	3	1	2		1	275	5.57
CCG/CGG	15	4	1							0	20	0.41
AAAC/GTTT	11	6							1	0	18	0.36
AAAG/CTTT	8	1	1							0	10	0.20
AAAT/ATTT	44	26	12	6	7			1		0	96	1.95
AACC/GGTT	13									0	13	0.26
AACT/ATTG	1	1								0	2	0.04
AAGG/CCTT	2									0	2	0.04
AATC/AGTT	8	2								0	10	0.20
AATT/AATT	9	3								0	12	0.24
ACAT/ATGT	3									0	3	0.06
ACCC/GGGT	2									0	2	0.04
AGAT/ATCT	3									2	5	0.10
AGGG/CCCT	1									0	1	0.02
AGGT/ATCC	1									0	1	0.02
AGTC/AGTC			1							0	1	0.02
CCCG/CGGG	2									0	2	0.04

(*Contd.*)

Table 2—Occurrence and number of repeats of 77 SSR motifs in *A. annua*—(*Contd*.)

Repeats	Number of repeat unit										Total repeat	%
	5	6	7	8	9	10	11	12	13	Above		
AAAAC/GTTTT		3		2						0	5	0.10
AAAAG/CTTTT	4	3								0	7	0.14
AAAAT/ATTTT	2	3	2							0	7	0.14
AAAGT/ATTTC	1									0	1	0.02
AAATC/AGTTT	2									0	2	0.04
AAATG/ACTTT	1									0	1	0.02
AACAG/CTTGT	1									0	1	0.02
AAGAC/CTGTT	1									0	1	0.02
AAGGT/ATTCC	1									0	1	0.02
ACACT/ATGTG			1							0	1	0.02
ACATC/AGTGT			1							0	1	0.02
AGCGT/ATCGC	1									0	1	0.02
AGGCC/CCGGT	1									0	1	0.02
AAAAAC/GTTTTT	1									0	1	0.02
AAAAAG/CTTTTT	1									0	1	0.02
AAAAAT/ATTTTT	1	1								0	2	0.04
AAAACC/GGTTTT	1									0	1	0.02
AAACAC/GTGTTT	1									0	1	0.02
AAACCC/GGGTTT	1									0	1	0.02
AAACTG/ACTTTG	2									0	2	0.04
AAAGAT/ATTTCT	2									0	2	0.04
AACAAG/CTTGTT	1									0	1	0.02
AACACC/GGTTGT	1									0	1	0.02
AACACT/ATTGTG	1									0	1	0.02
AACCAC/GGTGTT	2									0	2	0.04
AACCAG/CTTGGT	2									0	2	0.04
AACCCC/GGGGTT	1									0	1	0.02
AACCCT/ATTGGG	4		1							0	5	0.10
AAGACC/CTGGTT	4		1							0	5	0.10
AAGCCC/CGGGTT	3									0	3	0.06
AAGGTT/AATTCC		2								0	2	0.04
AAGTGG/ACCTTC			1							0	1	0.02
AATAGT/ATCATT	1									0	1	0.02
AATGAT/ACTATT	1									0	1	0.02
AATGGG/ACCCTT	1									0	1	0.02
AATGGT/ACCATT	3									0	3	0.06
ACCACT/ATGGTG	1									0	1	0.02
ACCATC/AGTGGT	1		1							0	2	0.04
ACCGCC/CGGTGG	2									0	2	0.04
ACCTCG/AGCTGG	1									0	1	0.02
ACGAGC/CGTGCT		3								0	3	0.06
ACGATG/ACTGCT	1									0	1	0.02
ACGGGG/CCCCTG	1									0	1	0.02
ACGGTG/ACTGCC	1									0	1	0.02
ACGTCC/AGGTGC		1								0	1	0.02
ACTCGG/AGCCTG	2									0	2	0.04
AGCAGT/ATCGTC	1									0	1	0.02
AGGCTC/AGTCCG	2									0	2	0.04
Total	3128	886	388	239	102	45	26	27	10	83	4934	

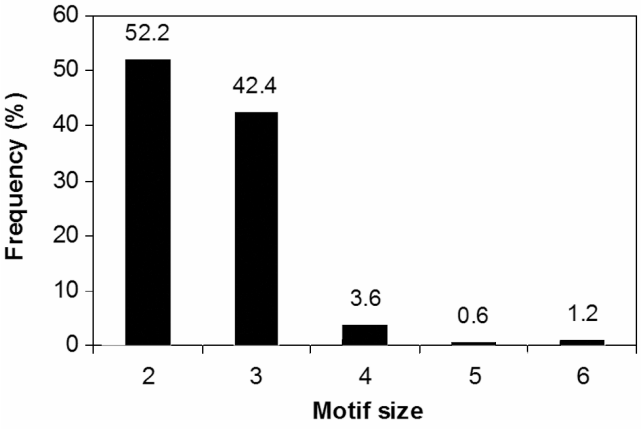

Fig. 2—The distribution pattern of SSRs in terms of motif size.

evenly distributed. Among the SSRs detected, the dinucleotides are the most frequent with a frequency of 52.19%, followed by trinucleotides (42.39%). The tetranucleotide SSRs have a much lower frequency (3.6%). Further, the frequency of SSRs with motif sizes of five and six are 0.60% and 1.20%, respectively (Fig. 2). The mean SSR length of each motif varied between 10 and 78 bp. The overall average SSR length was 20 bp with a maximum of 78 bp (motif types: AT/AT and AAT/ATT).

Distribution of SSRs Based on Motif Types

The SSRs identified in the present study were characterised by 77 types of motifs (Table 2). In general, the SSRs were found to be unevenly distributed across motif types. The motif AC/GT had the highest frequency of 23.23%, followed by AT/AT (18.48%), AG/CT (9.8%), ACC/GGT (8.89%), AAC/GTT (6.89%), AAG/CTT (6.38%), AAT/ATT (5.5%) and AGT/ATC (5.57%). The other types of motifs possessed a frequency of < 5%.

EST-SSR Polymorphism

Using DNA samples isolated from 20 genotypes of *A. annua* as templates, polymorphic DNA fragments were amplified from 13 out of 16 SSR primer pairs selected in the study. The sizes of these fragments ranged from 90 to 600 bp. A total of 38 alleles, with the average alleles per locus of 2.92, were detected at 13 loci. More than one allele was detected at 12 out of 13 loci, with the polymorphic markers ratio of 0.92. A maximum of 3-7 alleles were detected with 5 markers, followed by 2 alleles detected by 7 markers and 1 allele by only 1 marker (Table 1). Fig. 3 shows the representative amplified products from 3 different SSR markers. PIC refers to the value of a marker for

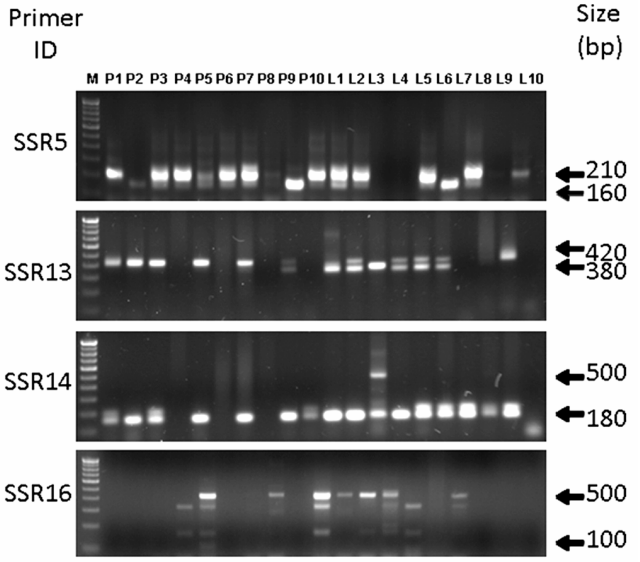

Fig. 3—SSR amplification products obtained from the 20 genotypes of *A. annua* studied: M, λ DNA as mol wt marker; P1 to P10, Genotypes collected from Nubra (Partapur) valley; & L1 to L10, Genotypes collected form Leh valley.

detecting polymorphism within a population or set of genotypes by taking into account not only the number of expressed alleles but also the relative frequencies of alleles per locus. PIC was calculated for the markers generated 2 or more alleles. As evident, SSR marker '15' showed the highest level of polymorphism with PIC value of 0.802, followed by SSR marker '16' (0.716), SSR marker '13' (0.595) and SSR marker '6' (0.565). The PIC values for rest of the SSR markers were in the range of 0.198-0.495. In case of non-amplifying SSR primers, different annealing temperatures (±5°C of Tm) in combination with different PCR reactions were tried but no amplification was observed.

Cluster Analysis

Significant genetic variation was found among *Artemisia* genotypes with Jaccard's genetic similarity coefficient ranged from 0.24 to 0.87. All 20 genotypes could be discriminated successfully by SSR markers (Fig. 4). The genotypes collected from Leh valley were distributed among 2 clusters (I and IIb), except only one genotype. Similarly, all the genotypes collected from Nubra valley were distributed within one cluster (IIa). The results of PCA analysis were comparable to the cluster analysis (Fig. 5).

Genetic Diversity Analysis

A relatively high genetic variation was detected among the *A. annua* genotypes. Genetic diversity

analysis in terms of Na, Ne, H, I, Ht, NPL and PPL among both the valleys revealed higher values for Leh valley, indicating more variability among the genotypes in comparison to Nubra valley (Table 3). The respective values for overall genetic variability for Na, Ne, H, I, Ht, Hs, Gst, NPL, PPL and Gene flow (Nm) across all the 20 genotypes are also given in Table 4. Gst value of 0.1808 indicated that 81.92% of the genetic diversity resided within the population.

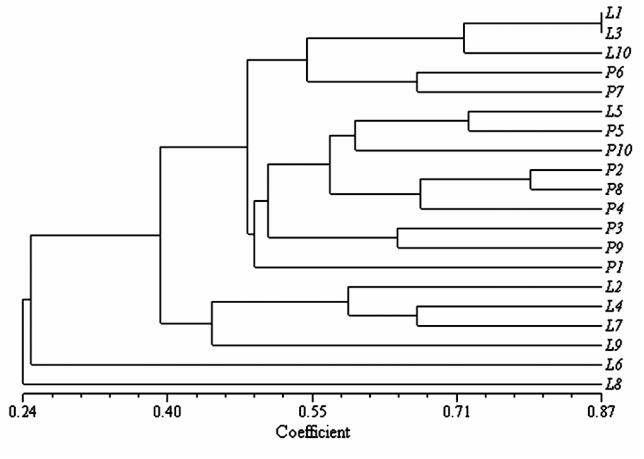

Fig. 4—Dendrograms generated using UPGMA analysis showing relationships between 20 *A. annua* genotypes using SSRs data.

AMOVA helps in partitioning of the overall genetic variations among groups and populations within the group. Molecular variance among valley is 23% and among genotypes within valley is 77% (Table 5), which reveals higher variation within the population in *A. annua*. All the components of molecular variation were significant ($P < 0.001$).

Discussion

Microsatellites have been used as a resource for random candidate markers in population genetics studies[21]. To better understand the natural diversity of *A. annua* genotypes and to develop strategies for its

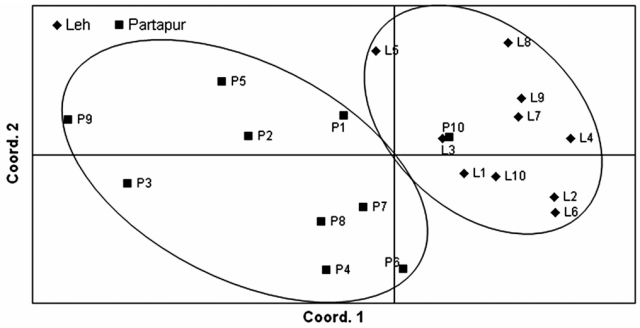

Fig. 5—Two-dimensional plot of PCA of 20 *A. annua* genotypes using SSRs analysis.

Table 3—Summary of genetic variation statistics for all loci of SSRs among the *A. annua* populations with respect to their distributions among two valleys

Valley	Sample size	Na	Ne	H	I	Ht	NPL	PPL
Nubra	10	1.5676 (0.5022	1.3544 (0.3645)	0.2097 (0.2008)	0.3127 (0.2904)	0.2097 (0.0403)	21	56.76
Leh	10	1.7297 (0.4502)	1.4539 (0.3773)	0.2643 (0.1929)	0.3945 (0.2707)	0.2643 (0.0372)	27	72.97
Mean		1.649	1.404	0.237	0.3536	0.237	-	-

Na = Observed number of alleles; Ne = Effective number of alleles; H = Nei's gene diversity; I = Shannon's Information index; Ht = Total genetic diversity; NPL = Number of Polymorphic Loci; PPL = percentage of Polymorphic Loci

Table 4—Overall genetic variability across all the 20 genotypes of *A. annua* based on SSR analysis

Na	Ne	H	I	Ht	Hs	Gst	NPL	PPL	Nm
1.9730 (0.1644)	1.4646 (0.3026)	0.2893 (0.1409)	0.4499 (0.1409)	0.2893 (0.0198)	0.2370 (0.0169)	0.1808	36	97.30	1.1325

Hs = Genetic diversity in population; Gst = Genetic diversity between population; Nm= Estimate of gene flow from Gst [Nm = 0.25 (1-Gst)/Gst]

Table 5—Summary of nested analysis of molecular variance (AMOVA) based on SSR marker among the populations of *A.annua* (Levels of significance are based on 1000 iteration steps)

Source of variation	Df	SSD	Variance component	Percentage	P-value
Among valley	1	19.350	1.448	23	<0.001
Among genotypes/valley	18	87.70	4.872	77	<0.001

Where Df = Degree of freedom; SSD = Sum of square deviation; P-value = Probability of null distribution

sustainable utilization, we identified SSR motifs in a dataset of 68,974 unique EST sequences (about 44.12 Mb). From a total of 4,934 putative SSR motifs, 4,342 (6.3%) unique ESTs were identified. Some of the ESTs des not have any SSRs motifs, whereas some ESTs consistsof more then one SSR motifs. The incidence of SSRs (6.3%) within ESTs was lower in comparison to other plant, such as, apple (20 %)[22], but it was similar to that of some dicotyledonous species (ranging from 2.65 to 16.82%)[23]. In the present study, the frequency of occurrence for EST-derived SSRs was one in every 8.94 kb. This is in accordance to the earlier findings with many species, such as, one EST-SSR occurs every 13.8 kb in *Arabidopsis thaliana*, 3.4 kb in rice, 8.1 kb in maize, 7.4 kb in soybean, 11.1 kb in tomato, 20.0 kb in cotton and 14.0 kb in poplar[24]. Among the SSRs, dinucleotides (2,344) were the most abundant repeat units, followed by trinucleotides (1,872), tetranucleotides (172), hexanucleotides (56) and pentanucleotides (30). Tri-nucleotide repeats have been found to be common feature in EST-derived SSRs. High frequency of these repeats in coding regions could be due to mutation and selection pressure for specific amino acids[25]. The abundance of trinucleotide repeats EST-SSR is likely due to suppression of other kind of repeats in the coding region, which reduces the frame-shift mutations in the coding regions[26].

The frequency of occurrence of SSRs types in *A. annua* is different from that of other medicinal plant *Epimedium sagittatum*[27] and cereal species[10], where trinucleotide repeat units are the most dominant SSR, followed by di and tetranucleotide repeat units[10,22]. The relative frequency of repeats with different dinucleotide compositions was also biased towards one of four possible repeat classes (Table 3). Among the dinucleotide repeat classes, (AC/GT)n repeats were the most common dimer motif (23.23%), followed by (AG/CT)n, (AT/AT)n and (CG/CG)n with a frequency of 9.83, 18.48 and 0.65%, respectively. This is in the agreement with studies in cultivated peanut (*Arachis hypogaea* L.)[28] and wild *Arachis* species[29]. Among the trinucleotide repeats, (ACC/GGT)n, (AAC/GTT)n, (AAG/CTT)n, (AAT/ATT)n and (AGT/ATC)n were the largest repeat class, followed by (ACT/ATG)n, (AGC/CGT)n, (ACG/CTG)n, (AGG/CCT)n and (CCG/CGG)n (Table 2). In other plant species, the most frequent trinucleotide repeat motifs were (AAC/TTG)n in wheat, (AGG/TCC)n in rice, (CCG/GGC)n in maize, (AAG/TTC)n in soybean, and (CCG/GGC)n in barley and sorghum[10,30,31]. These unique putative transcript-derived SSR markers that were generated in the present study provide a valuable genetic resource for future studies of *A. annua* and other related species.

Some of the EST-SSRs may tightly link with functional genes that control synthesis of bioactive compound, artemisinin. Therefore, we considered ESTs derived from the genes belonging to artemisinin biosynthetic pathway for experimental analysis and polymorphism study among 20 individuals of *A. annua.* The SSRs confirmed in this analysis may be valuable for screening of high yielding genotypes. In the present study, 38 alleles were detected with 13 SSR loci with an average of 2.92 alleles per locus. This relatively small number is probably due to the limited number of genotypes studied and also relatively high genetic similarity that exists within the investigated group of *Artemisia* germplasm. Genetic relationships were found to be very close among the genotypes from the Nubra valley in comparison to Leh valley. The fact that eight genotypes from the Nubra valley clustered into the same sub-group gave the strong indications of the narrow genetic background in *Artemisia* germplasm. The genotypes from Leh valley investigated in this study, on the other hand, showed a broader genetic diversity. Therefore, in order to avoid the potential risks associated with too little genetic diversity, the adoption of elite genotypes from different origins used as parental lines is highly recommended for any genetic conservation and breeding programme in *A. annua*. AMOVA analysis revealed higher genetic variation within the population of *A. annua*. This is helpful in making strategy for germplasm collection and evaluation. The estimated gene flow was 1.1325. In population genetics, a value of a gene flow (Nm) <1.0 (less than one migrant per generation into a population) or, equivalently, a value of gene differentiation (Gst) > 0.25 is generally regarded as the threshold quantity beyond which significant population differentiation occurs[32]. Overall the study indicates that *A. annua* populations in the trans-Himalayan region of Ladakh are genetically highly diverse.

In summary, the successful experimental validation of majority of the computationally predicted SSR motifs confirms the utility of mining 68,974-ESTs for genetic markers. The development of EST-SSR markers in *A. annua* should greatly facilitate marker-

assisted selection, germplasm breeding, adulterated species identification and genetic diversity studies in this medicinal species.

Acknowledgement

The authors are thankful to the Defence Research and Development Organization, New Delhi for its support and funding the project.

References

1 Phillipson J D & Wright C W, Antiprotozoal agents from plant sources, *Planta Med*, 57 Suppl (1991) S53-S59.

2 Klayman D L, *Artemisia annua*: From weed to respectable antimalarial plant, in *The human medicinal agents from plants*, edited by A D Kinghorn & M F Balandrk (American Chemical Society Symp Ser, Washington, DC) 1993, 242-255.

3 Ro D K, Paradise E M, Ouellet M, Fisher K J, Newman K L, *et al*, Production of the antimaralial drug precursor artemisinic acid in engineered yeast, *Nature (Lond)*, 440 (2006) 940-943.

4 Sangwan R S, Agarwal K, Luthra R, Thakur R S & Sangwan S M, Biotransformation of arteannuic acid into artennuin-b and artemisinin in *Artemisia annua*, *Phytochemistry*, 34 (1993) 1301-1302.

5 Weising K, Atkinson R G & Gardner R C, Genomic fingerprinting by microsatellite-primed PCR: A critical evaluation, *PCR Methods Appl*, 4 (1995) 249-255.

6 Tautz D & Schlotterer C, Simple sequences, *Curr Opin Genet Dev*, 4 (1994) 832-837.

7 Liu Z J, Tan G, Kucktas H, Li P, Karsi A *et al*, High Levels of conservation at microsatellite loci among ictalurid catfishes, *J Hered*, 90 (1999) 307-312.

8 Liu Z, Tan G, Li P & Dunham R A, Transcribed dinucleotide microsatellites and their associated genes from channel catfish *Ictalurus punctatus*, *Biochem Biophys Res Commun*, 259 (1999) 190-194.

9 Scott K D, Eggler P, Seaton G, Rossetto M, Ablett E M *et al*, Analysis of SSRs derived from grape ESTs, *Theor Appl Genet*, 100 (2000) 723-726.

10 Varshney R K, Thiel T, Stein N, Langridge P & Graner A, *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species, *Cell Mol Biol Lett*, 7 (2002) 537-546.

11 Eujayl I, Sledge M K, Wang L, May G D, Chekhovskiy K *et al*, *Medicago truncatula* EST-SSRs reveal cross-species genetic markers for *Medicago* spp., *Theor Appl Genet*, 108 (2004) 414-422.

12 Varshney R K, Graner A & Sorrells M E, Genic microsatellite markers in plants: Features and applications, *Trends Biotechnol*, 23 (2005) 48-55.

13 Kuleung C, Baenziger P S & Dweikat I, Transferability of SSR markers among wheat, rye, and triticale, *Theor Appl Genet*, 108 (2004) 1147-1150.

14 Gadaleta A, Mangini G, Mulè G & Blanco A, Characterization of dinucleotide and trinucleotide EST-derived microsatellites in the wheat genome, *Euphytica*, 153 (2007) 73-85.

15 Doyle J J & Doyle J L, Isolation of fresh DNA from fresh tissue, *Focus*, 12 (1990) 13-15.

16 Rohlf F J, NTSYS-pc numerical taxonomy and multivariate analysis system, version 2.1 manual (Applied Biostatistics, Inc., New York) 1992.

17 Zhao W G, Zhang J Q, Wangi Y H, Chen T T, Yin Y L *et al*, Analysis of genetic diversity in wild populations of mulberry from western part of Northeast China determined by ISSR markers, *J Genet Mol Biol*, 7 (2006) 196-203.

18 Nei M, Estimation of average heterozygosity and genetic distance from a small number of individuals, *Genetics*, 89 (1978) 583-590.

19 Excoffier L, Smouse P E & Quattro J M, Analyses of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data, *Genetics*, 131 (1992) 479-491.

20 Peakall R & Smouse P E, GenAlEx V5: Genetic analysis in Excel. Population genetic software for teaching and research (Australian National University, Canberra, Australia) 2001.

21 Kashi Y & King D G, Simple sequence repeats as advantageous mutators in evolution, *Trends Genet*, 22 (2006) 253-259.

22 Newcomb R D, Crowhurst R N, Gleave A P, Rikkerink E H A, Allan A C *et al*, Analyses of expressed sequence tags from apple, *Plant Physiol*, 141 (2006)147-166.

23 Kumpatla S P & Mukhopadhyay S, Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species, *Genome*, 48 (2005) 985-998.

24 Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D *et al*, Computational and experimental characterization of physically clustered simple sequence repeats in plants, *Genetics*, 156 (2000) 847-854.

25 Morgante M, Hanafey M & Powell W, Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes, *Nat Genet*, 30 (2002) 194-200.

26 Metzgar D, Bytof J & Wills C, Selection against frameshift mutations limits microsatellite expansion in coding DNA, *Genome Res*, 10 (2000) 72-80.

27 Zeng S, Xiao G, Guo J, Fei Z, Xu Y *et al*, Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. et Zucc.) Maxim, *BMC Genomics*, 11 (2010) 94.

28 Luo M, Dang P, Guo B Z, He G, Holbrook C C *et al*, Generation of expressed sequence tags (ESTs) for gene discovery and marker development in cultivated peanut, *Crop Sci*, 45 (2005) 346-353.

29 Proite K, Leal-Bertioli S C M, Bertioli D J, Moretzsohn M C, da Silva F R *et al*, ESTs from a wild *Arachis* species for gene discovery and marker development, *BMC Plant Biol*, 7 (2007) 7.

30 Peakall R, Gilmore S, Keys W, Morgante M & Rafalski A, 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 (1998) 1275-1287.

31 Thiel T, Michalek W, Varshney R & Graner A, Exploiting EST databases for the development and characterization of gene derived SSR-markers in barley (*Hordeum vulgare* L.), *Theor Appl Genet*, 106 (2003) 411-422.

32 Slatkin M, Gene flow and geographic structure of natural populations, *Science*, 236 (1987) 787-792.