INTERNATIONAL JOURNAL OF AGRICULTURE & BIOLOGY
ISSN Print: 1560–8530; ISSN Online: 1814–9596
18–1705/2019/21–6–1293–1300
DOI: 10.17957/IJAB/15.1024
http://www.fspublishers.org

Full Length Article

In-Silico Identification and Characterization of EST-SSR Markers from *Catharanthus roseus* for Genetic Diversity Analysis

Manish Kapoor, Jyoti Rani* and Navneet Kaur
Department of Botany, Punjabi University, Patiala-147002, Punjab, India
*For correspondence: jdmanishkapoor@yahoo.com

Abstract

Catharanthus roseus is one of the most significant and important perennial plant of Apocynaceae family, globally known for its ornamental and medicinal values. Molecular markers have great importance in order to investigate composition, genetic variation and evolutionary relationships. In this study, accessions collected from five north Indian state were cultivated at Plant Conservatory, Punjabi University, Patiala and their morphological variation were analyzed. In total 22, 867 ESTs (expressed sequence tags) were downloaded from NCBI database and assembled into 2,853 sequences with an average size of 659 bp by using bioinformatic tools. Among assembled dataset a total of 427 SSRs (simple sequence repeats) were successfully detected in 358 sequences. For functional annotation all SSR containing sequences were subjected to Blast-x with E-value of 10^{-5} against NCBI-nr (non-redundant) database and Swiss-Prot database that resulted functional hits for 279 (75%) sequences. Furthermore, GO (Gene ontology) was predicted by comparing the sequences with Arabidopsis database (TAIR), total 268 GO ids were assigned to 259 sequences. Additionally, 141 different pathways were predicted by using KEGG database with highest enrichment of metabolic pathways category. Finally, to validate the identified SSR containing sequences, total 260 primers targeting repeat regions were designed. Further, 28 primers of class I category were selected for validation among 25 accessions of *C. roseus* and resulted that 17 primers amplified the corresponding repeated region. Additionally, among validated primers, 5 primers were found to be polymorphic further used for analysis of variation among accessions. These identified markers can be further used to investigate the variation among the species and can be helpful for detection of genes responsible for the survival of plant for particular stress condition.

Keywords: *Catharanthus roseus*; Simple sequence repeats (SSRs); Expressed sequence tags (ESTs); KEGG pathways

Introduction

Simple Sequence Repeats (SSRs) can be developed either through development of genomic libraries or from expressed sequence tag (EST) database; however, SSR marker development from customary genomic library is time consuming and requires huge infrastructure lab facilities. With the continuous advancement and improvement in computational tools, it has become quite easier to extract and analyze enormous data of organisms present in various databases. The ESTs are derived through the reactions of single sequencing means by selecting random clones from the libraries of cDNA (Adams *et al.*, 1991). Over the last 10–15 years, *In-silico* based EST based data mining and analysis has played a great role in various laboratories dealing with several molecular biology projects (Ewing *et al.*, 1999; Ronning *et al*., 2003). Any young newly born tissue which is in active stage of growth can be selected for the isolation of mRNA which is then reverse transcribed to cDNA and further selected for sequencing in order to form EST sequences. The formed ESTs again assembled to form non-redundant sequences (singletons and contigs). In the last decade, various researches (Wang *et al.*, 2001) have exploited ESTs as a great source of co-dominant SSRs that can reveal genetic diversity and polymorphism. Extensive studies have been undertaken to identify SSR microsatellites in several crop species including grapes (Scott *et al.*, 2000), cereals (Varshney *et al.*, 2002), bread wheat (Gupta *et al.*, 2003), sugarcane (Pinto *et al.*, 2004), citrus (Chen *et al.*, 2006), watermelon (Verma and Arya, 2008), walnut (Zhu *et al.*, 2009) and cauliflower (Vaidya *et al.*, 2012). SSRs or microsatellites are short repeat motif of 1–6 base pair long. SSRs are more important markers compared to other molecular markers and are utilized for an assortment of various conditions as these are co-dominant markers and are present in abundance, these have high reproducibility rate and large genome coverage, also they show hyper variability due to multi-allelic nature (Kantety *et al.*, 2002). Markers derived from EST has been proved to play important roles in marker assisted selections which is helpful in genetic mapping, comparative genomics, genetic diversity and identification of quantitative trait locus (QTLs)

To cite this paper: Kapoor, M., J. Rani and N. Kaur, 2019. *In-silico* identification and characterization of EST-SSR markers from *Catharanthus roseus* for genetic diversity analysis. *Intl. J. Agric. Biol*., 21: 1293-1300

of important traits (Durand *et al.,* 2010; Vaidya *et al.,* 2012).

Catharanthus roseus (L.) G. Don. [*Vinca rosea* L.; *Lochnera rosea* (L.) Rchb. and *Ammocallis rosea* (L.) Small] belongs to family Apocynaceae, has 16 somatic chromosome number, native to Madagascar Island and spread throughout tropical and subtropical regions. It is an important ornamental plant with a high medicinal value, having various profitable uses in ayurvedic medicine. *C. roseus* possesses anti-tumor, antioxidant, anti-inflammatory, anti-aging, immune modulatory and rejuvenating properties (Heijden *et al.,* 2004). Initially, the plant was grown in gardens because of its beautiful flowers and various colours, such as white, pink, white with red and yellow eye, deep rose, lavender blue, dark purple and scarlet red (Shaw *et al.,* 2009). Though considerable variation can be observed in gardens around the world; however, attempts have not been made so far to study the genomic relationship among different accessions of *C. roseus*. As of now, various microsatellite markers have been created and conveyed for the investigation of intraspecific and interspecific and in addition intrageneric and intergeneric genetic diversity analysis, in *C. roseus* (Joshi *et al.,* 2011). Due to the continuous emerging sequencing technologies (Genomic and Transcriptomic) large scale ESTs have been generated and submitted to publicly accessible databases offering a chance to create EST determined SSR markers by data mining and concreate analysis. The present investigation tends to survey the appropriateness of existing publicly available EST sequences at NCBI for the mining and creation of SSR markers in order to identify the transcribed region in genome of *C. roseus*. After functional characterization these markers were utilized to study the genetic diversity among 25 different accessions of *C. roseus*.

Materials and Methods

Plant Material

The plant materials of several accessions of *C. roseus* (seeds/plants) were collected from various localities of five north Indian state (Chandigarh, Delhi, Haryana, Punjab and Rajasthan), depending upon morphological variations. Thereafter, the seeds of each accession were raised in the nursery trays with potting media of cocopeat and vermicompost on February 21, 2014. One seed was sown per cell in the tray. The germination started after 10–21 days. Seedlings (about 45 days old) were used as planting material. The seedlings were transplanted to experimental field and cultivated at Plant Conservatory of Punjabi University, Patiala on April 15, 2014.

Experimental Design and Field Layout

The experiment was laid out in Randomized Block Design (RBD) replicated thrice. The seedlings were planted in plots (150 cm x 120 cm) at 50 cm between rows and 30 cm within plants in a row. Each accession was planted in three rows with four plants per row providing a total of 12 plants per accession. Data on the morphological characters was collected from 10 randomly selected plants and their mean was recorded for all observations. 25 accessions of *C. roseus*, differing in flower petal color, eye and flower center color and arrangement of petals as well as growth habits (Fig. 1; Table 1).

Retrieval of EST Sequences and SSR Mining

Total 22,867 EST sequences of *C. roseus* retrieved from EST database at NCBI (www.ncbi.nlm.nih.gov/nucest) were downloaded followed by *de novo* assembly using CLC Genomics Workbench version 6.5 (CLC Inc., Aarhus, Denmark), the minimum length for the generation of transcripts was set as 200. Thereafter, MISA tool (http://pgrc.ipk-gatersleben.de/misa/) was used to identify of SSR containing sequences with the motifs ranging from 2 to 6 nucleotides in length. The repeating length criteria for different repeated units was: repeat numbers $\geq$ 6 for dinucleotide, at least $\geq$ 5 for trinucleotide and $\geq$ 3 was considered for tetranucleotides, pentanucleotides and hexanucleotides with maximum 100bp differences between two SSRs in a sequence.

Functional Characterization of Sequences with SSRs

Assembled sequences with SSR regions were selected for the functional identification by comparative means. Three publicly available databases (NCBI-NR, Swiss-Prot and TAIR) were selected to perform similarity search through Blast-X by selecting e-value less than 0.00001. Gene ontology (GO) annotation of the unigene was done by using AgriGO (DU *et al.,* 2010). GO functional classification of all SSR unigene was done by WEGO and classification of enrichment analysis was examined based on cellular, biological and molecular functions. Additionally, KEGG database was used for pathway predictions to study the crucial role of SSRs in several regulating or metabolic pathways.

Primer Designing and Validation

Primers were designed from the flanking region of the microsatellite marker loci, using Batch Primer3 program. The length of amplicon was set to 100–250 bp, various parameter for primer designing were selected for optimum length of 20 bp with probable range of 18–22 bp, optimum GC content was 55% (40–60% range) and optimum temperature was 60°C set as a range of 57–65°C. Functionally annotated Class-I primers were selected for further validation among 25 accessions of *C. roseus*. CTAB method was used to isolate the genomic DNA from newly formed young leaves (Doyle, 1990) from plants grow at

Table 1: Accession codes, flower petal colors, flower eye colors and photograph of flowers of the 25 different accessions of *C. roseus*

Accession No.	Petal color (RHSCC)	Petal arrangement	Eye color (RHSCC)
Cr00PFRE	RHS-57C	Slightly overlapping	RHS-154D
Cr00LPFLPE	RHS-155B	Free	RHS-66C
Cr00DPF	RHS-66B	Strongly overlapping	RHS-64A
Cr00WFYE	RHS-155C	Free	RHS-66C
Cr00WFRE	RHS-155D	Touching	RHS-67C
Cr00WFRE2	RHS-155D	Strongly overlapping	RHS-57A
Cr00WFSRE	RHS-155C	Strongly overlapping	RHS-58D
Cr00WFYE2	RHS-155B	Strongly overlapping	1C
Cr00DP	RHS-68B	Touching	RHS-155C
Cr00LP	RHS-64D	Strongly overlapping	RHS-155D
Cr00BPF	RHS-72D	Free	RHS-155B
Cr00SFP	RHS-69C	Free	RHS-57B
Cr00CAF	RHS-179D	Strongly overlapping	RHS-57B
Cr00LPNF	RHS-75D	Free	RHS-64B
Cr00SBRF	RHS-67B	Free	1D
Cr00PLMF	RHS-61D	Touching	50B
Cr00CHEF	46C	Slightly overlapping	RHS-62B
Cr00SNFF	RHS-67B	Slightly overlapping	RHS-155C
Cr00CLF	RHS-80C	Free	RHS-155D
Cr00SAF	39D	Strongly Overlapping	RHS-61C
Cr00TDRF	RHS-67A	Strongly Overlapping	-
Cr00DRYE	RHS-57C	Free	RHS-154D
Cr00FRFF	RHS-66B	Free	RHS-57B
Cr00PFWE	RHS-60C	Slightly overlapping	RHS-155D
Cr00PFYE	RHS-67D	Free	RHS-115B

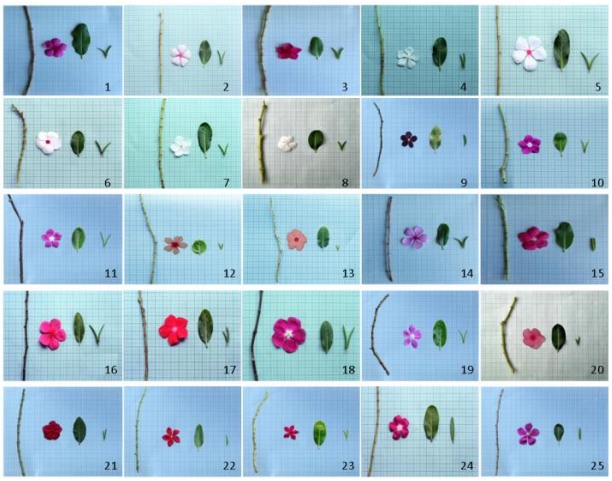

Fig. 1: different accessions f *C. roseus* with code: 1 (Cr00PFRE), 2 (Cr00LPFLPE), 3 (Cr00DPF), 4 (Cr00WFYE), 5 (Cr00WFRE), 6 (Cr00WFRE2),7 (Cr00WFSRE), 8 (Cr00WFYE2), 9 (Cr00DP), 10 (Cr00LP), 11 (Cr00BPF), 12 (Cr00SFP),13 (Cr00CAF), 14 (Cr00LPNF), 15 (Cr00SBRF), 16 (Cr00PLMF) 17 (Cr00CHEF), 18 (Cr00SNFF) 19 (Cr00CLF), 20 (Cr00SAF), 21 (Cr00TDRF), 22 (Cr00DRYE), 23 (Cr00FRFF) 24 (Cr00PFWE), 25 (Cr00PFYE)

Plant conservatory. To check the quality and concentration of DNA electrophoresis was used with 0.8% agarose gel. PCR amplification for each SSR was performed in a total volume of 25 μL reaction having genomic DNA of 10 ng, Taq polymerase (Invitrogen) quantity was1.0 U, every primer was 0.5 µM, content of $MgCl_2$ was 1.5 m*M*, each dNTP (Invitrogen) with 2.5 mM and 10x buffer used was 2.5 μL. ICycler (Bio-Rad Laboratories, Hercules, C.A., U.S.A.) performed the PCR reactions by 4 min at 94°C followed by 36 cycles of 1 min at 94°C. Primer annealing temperature was set at 56°C for 45 seconds and a last augmentation at 72°C for 5 min was done. The PCR products were evaluated by using gel electrophoresis in 2% agarose and finally visualized by staining ethidium bromide dye. Finally, amplified products expected sized bands were further separated on 8% metaphor agrose gel.

Data Analysis

To calculate the population genetics parameters among various accessions of *C. roseus* POPGENE v1.31 was used (Yeh, 1999), it involves in estimation of various types of population genetics data that includes the number of allele (N_A) counting, examination of expected and observed heterozygosity (H_E and H_O), the values of polymorphic information content (PIC) and the Shannon diversity index (I). To calculate the genetic similarity a simple coefficient matching was used based on which dendrogram was constructed using UPGMA (Unweighted Pair Group Method with Arithmetic Mean) method (Sokal, 1958). Genes software was considered to analyze heterozygosity and bootstrapping (Cruz, 1998). Finally, PIC was calculated by the formula $PIC=1-\sum p_i^2$, where p_i represents the recurrence of i^{th} allele from each SSR marker (Anderson *et al.*, 1993).

Results

Assembly of EST Sequences

ESTs are significant for identification of sequence features and gene discovery as these are partial and redundant cDNA sequences. The publicly available ESTs of *C. roseus* were downloaded assembled by CLC Genomic workbench version 6.5 program to grep longer length sequences with low redundancy. CLC Genomic workbench locates the overlapping region of sequences in order to generate the final consensus sequences for further analysis. A total of 2,853 contigs were formed with average length of 626 and N_{50} value of 681. Distribution of matched read length is shown in Fig. 2.

Frequency and Distribution of EST-SSRs

All assembled transcripts (2,853) were considered to identify and mining of SSRs (without considering mono-nucleotides) and a total 358 sequences were detected to carry 427 (10.22%) SSRs with the criteria of unit size of (2/6), (3/5), (4/3), (5/3) and (6/3) (Table 2). Among these SSRs, di-repeats were found to be most abundant

Table 2: Summary of the *in-silico* search for SSRs in *C. roseus*

Parameter used for screening	Data generated by MISA
Total number of sequences examined	2853
Total size of examined sequences (bp)	1785878
Total number of identified SSRs	427
Number of SSR containing sequences	358
Number of sequences containing more than 1 SSR	57
Number of SSRs present in compound formation	28
Dinucleotide	126
Trinucleotide	109
Tetranucleotide	122
Pentanucleotide	34
Hexanucleotide	36

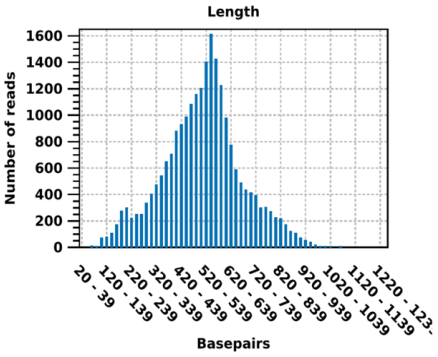

Fig. 2: Read length distribution

(29.5%) followed by tetra-repeats (28.57%), other repeat types like tri repeats, penta-repeats and hexa-repeats were found to be 25.52, 8.4 and 7.9% respectively. Further analysis revealed that among di-repeat nucleotides AG/CT repeat types were 27.7%, followed by AT/AT repeat types, which were found to be 22.2%. Furthermore, in tri-repeat types motif AAG/CTT were found to be 29.5% followed by AGC/CTG (11.4%) and AAT/ATT (10.6%). Thereafter, motifs AAAG/CTTT (17.6%) followed by AATT/AATT (12.7%) were found to be most common among tetra-repeat types. There was no significant contribution of penta-repeat and hexa-repeat nucleotides was observed. Repeat type frequency of identified SSRs is shown in Fig. 3.

Annotation of Sequences Containing SSRs

For the functional analysis of SSRs containing sequences comparative similarity search was conducted by using the Blast-X by selecting E-value of 0.00001 against NCBI-nr database followed by Swiss-Prot and TAIR database. We found that 311 SSR containing sequences have top hits from non-redundant database of NCBI further showed that, a total of 111 functional hits were found from *Coffea canephora* followed by *Solanum tuberosum* with total 24 similar hits. Top ten species from which top hits were obtained are shown in Fig. 4. Comparative analysis with TAIR resulted in 268 functional hits and Swiss-Prot database yielded 233

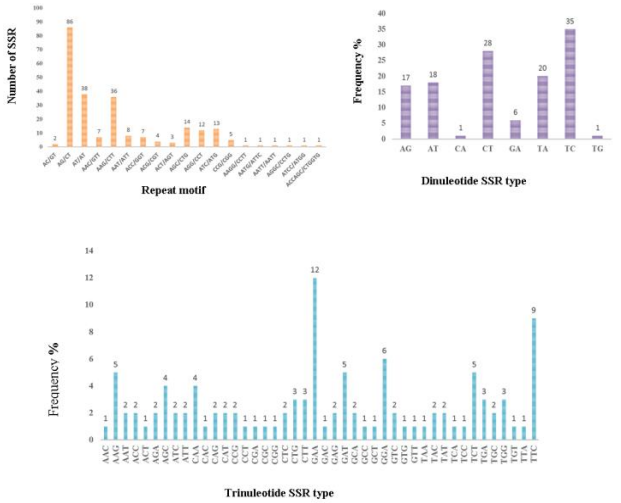

Fig. 3: Frequency of different nucleotide repeats in identified SSR sequences

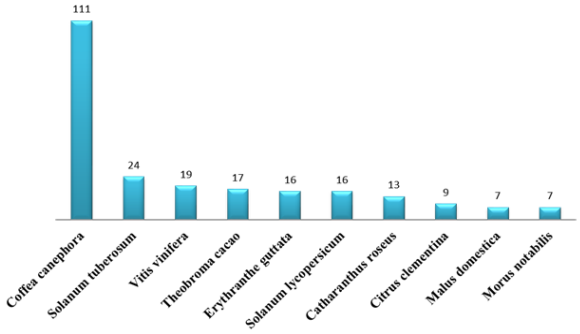

Fig. 4: Top ten nr-BLAST hits

top hits for functional characterization. Annotation results showed 87% homology with NCBI-nr database, 74% homology with TAIR database and 65% with Swiss-Prot database. Additionally, gene ontology (GO) annotation revealed that SSRs containing sequences belonged to the three major functional classes such as biological processes, molecular functions and cellular components. Further analysis revealed that these have further sub-categorized in 40 classes including apoptosis, reproduction, development, growth and metabolism. Among these, biological process (392), were more prominent, followed by molecular function (309) and cellular component (273). Other significant sub-categories of biological processes included cellular process (105), metabolic process (94), response to stimulus (73), pigmentation (37), biological regulation (41) and developmental process (25). Under molecular function, contigs were categorized for binding (95), catalytic activity (83) and transcription regulation (22). Cell (147) and organelle (101) constituted major portion under cellular component category (Fig. 5). The functional

Table 3: Characteristics of 28 EST-derived SSRs for *C. roseus*

Code	Sequence of primer pairs	Target motif	Annealing temperature (°C)	GC %	Product size (bp)
CR164	F: GAA GAT TCT CTG GCC TTC TTT G R: CAC CAC CAG ACA TCT TGG AA	CT	54 55	45 50	165
CR266	F: AAC AAC ACC GAG GTT GAA GC R: AAG GCA TCA GCC AAT TCT TG	AGC	56 54	50 45	152
CR508	F: CGT CAG TCT CTT CTC CCA TCA R: AGG GAG CCG GAA TCA TTA AG	ACC	56 54	52 50	163
CR556	F: CCG ACT TTC TTCTTC TGG ACA R: TTT CCC TCA TCT TGT AGC CTT C	TC	54 54	47 45	151
CR651	F: CTA AAG CGG CTG CAA AGA CT R: GTA GGG GCA GGA AGA TGT GA	GAA	56 56	50 55	142
CR678	F: CAA CAG AGG CAA GGA AAG AA R: TCC CAA TGA CCT TTC CTC TT	AG	53 53	45 45	143
CR708	F: TGT TGG CCT TGA TGA GAA GA R: GGT GCA TGC TGA AGT TTT GA	TA	54 54	45 45	164
CR870	F: AGC TTG CAC GCG ATA AAA TC R: CAT TGT GTT GGG CAT TGG TA	AT	54 54	45 45	152
CR1127	F: CCT CAC CTT CTC GTG GCT TA R: GCC GGT TGT TCG TTG TTT AG	AG	56 55	55 50	207
CR1144	F:GCC TCC GTA GGT CTT GTT CA R:CTC ATC CGA TTG AAC ATC CA	AT	57 52	55 45	187
CR1356	F:CCC TCC TGA AGC TGC TTA TC R:GAC CAA GAC GAC CAA GGT TC	TTC	55 56	55 55	152
CR1396	F:TTA GCC TCA TTC CGT CCT TG R:CTC AGA ATG GGG GTG ATG AT	TTC	55 54	50 50	146
CR1438	F:GAA GAA CAG CAA CAA CAG AG R:CAC TCC ACC TCT GGT ATA AA	CT	51 51	45 45	254
CR1538	F:CCT TTG ACC TGT TTT TGT CC R:GGA GGA AAT ATG CAG ACG AA	ATT	55 55	45 45	149
CR1543	F:ACC GAT GAA GAT CGG TGT C R:GGT CGA CCA GTC ACT GAA AA	CTC	52 54	52 50	132
CR1559	F:TTA CAA CAC CGA CGC CTA TG R:TGT TGT TCT CAG GGT TGG AA	AAC	54 54	50 45	164
CR1668	F:GGA CCA AAG CTT CTT CTG R:CTC AAG AGC CAC AAT AAC TC	TA	54 56	50 45	137
CR1694	F:CCC CCA ACT AGT CCA AAA CA R:AAC AGC AGCAGC CTT TGA A	TTC	59 55	50 47	147
CR1774	F:CCA TTT TGT TGA CCC TTT CC R:AAC CCA AGA TTT GGC ACA AG	TTC	55 56	45 45	156
CR1968	F:GGC CAA GTC GAA ATA CAT TAC C R:CCA GGG AAG AAT TGT CAA GC	TC	54 56	45 50	172
CR2317	F:CAG TGC AAG CAG TCC TCA AG R:AGC ACC ACCACCACC ACT A	TGG	59 60	55 57	151
CR2332	F:TCT GGA GAT GCA AGT TCG TG R:ATC ATC AGC TTG GAG GAG GA	CTT	59 59	50 50	155
CR2482	F:GCC CAT GAC TTG GTG GTT AC R:GCA GGG ATT CAA AGG AAC AG	AGC	60 59	55 50	154
CR2607	F:TCG GGT CCC TCA CTA TTG TT R:GCC GCT ATT CTC TTC AGC TT	AG	56 55	50 50	144
CR2724	F:GCT GCT AGC TGA GCA AAA AGA R:GGA ACG ATT GGG AAT TTC AG	CA	56 52	47 45	217
CR2725	F:GGT CAT GAT GAG GAGGAG GA R:TGC TCA GCT AGC AGC TTT TG	TCT	56 56	55 50	181
CR2773	F:GCT GCA TGG CAT CTG AAT AA R:CAT AAT ATT TGG CCT CCA CCT C	TA	53 53	45 45	178
CR2784	F:AAA CTC CAA CCG TTC GGA TAR:CCC TTC GTC AGT CAC CAT TT	AG	54 55	45 50	153

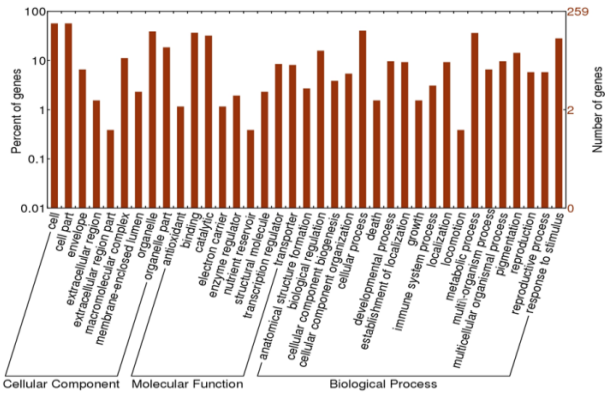

Fig. 5: Gene ontology (GO) classification of *C. reseus*

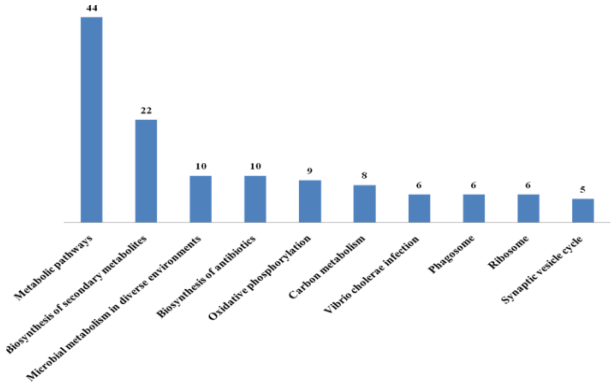

Fig. 6: Pathway assignment based on Kyoto Encyclopedia of Gene and Genome

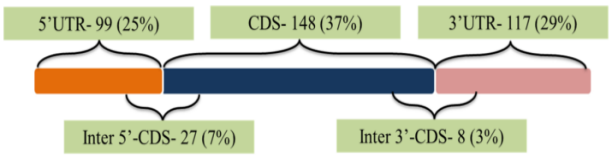

Fig. 7: Localization of SSRs in *C. roseus*

characterization proved that the found markers are targeting higher proportion of expressed part of genome in *C. roseus.*

Pathway Identification and SSRs Localization

Sequences containing SSR markers motifs were used to identify pathways using KEGG database. A total of 141 pathways were found with highest number of metabolic pathways (44) followed by biosynthesis of secondary metabolites (22), biosynthesis of antibiotics (10), oxidative phosphorylation (9) and carbon metabolism (8). Top 10 KEGG pathways are shown in Fig. 6. A total of 53 enzyme commission numbers were found to belong to 6 categories of enzymes. To predict the location of SSR markers, ORF regions of sequences was predicted based on which 5 different locations (5'UTR, CDS, 3'UTR, inter5' CDS and inter 3'CDS) were identified as shown in Fig. 7.

Table 4: Diversity statistics for 5 SSR loci studied in *C. roseus*

S.No.	Locus	na	Ne	I	Ho	He	PIC
1	CR651	2	1.987	0.69	0.92	0.507	0.373
2	CR1668	2	1.967	0.685	0.522	0.502	0.371
3	CR1774	3	1.615	0.659	0.24	0.389	0.333
4	CR1968	2	1.268	0.367	0	0.216	0.189
5	CR2607	2	1.24	0.344	0.217	0.198	0.175
Mean		2.2	1.615	0.549	0.38	0.362	0.288

Where, bp= Base pair, n_a= Observed number of alleles; n_e = Effective number of alleles; I= Shannon's information index; H_e= Expected heterozygosity; H_o= Observed heterozygosity; PIC = Polymorphism information content

Fig. 8: Genomic relationship among 25 accessions of C. roseus as revealed 5 SSR pri-mers using Jaccard's similarity co-deficient method

Genetic Relationship and Cluster Analysis

In the present study, total 260 primer pairs were designed and out of them, 28 class I type primers were synthesised (Table 3), used for validation and assessment of the polymorphism among the 25 *C. roseus* accessions out of which 17 showed amplification and 5 were found to be polymorphic. A total of 5 SSR markers were found to be polymorphic in 11 alleles of *C. roseus* with average number of 2.20 alleles per locus. The values of PIC were found to be ranged from 0.175 in primer (CR2607) to 0.373 in (CR651), with an average value of 0.288. The mean for the effective number of alleles (Na) is 2.2 and 1.615 for observed number of alleles (Ne). The highest (0.507) and the lowest (0.198) expected heterozygosity (H_e) values were obtained with (CR651) and (CR2607) primers, respectively, having average of 0.362. The range for the observed heterozygosity is 0 to 0.92 with a mean value of 0.38. Shannon's Information index (*I*) was recorded as 0.69 (maximum) and 0.344 (minimum) with an average of 0.549 (Fig. 8, Table 4). Neighbors joining (N-J tree) was constructed using Jaccard's coefficient for measuring genetic diversity and relatedness among accessions. The 25 accessions grouped into 3 major groups with high (>50%) boot strap value. Cluster I comprises of 11 accessions included Cr00PFRE, Cr00LPFLPE, Cr00BPF, Cr00DPF, Cr00WFYE, Cr00DRYE, Cr00CLF, Cr00FRFF, Cr00SFP, Cr00SBRF and Cr00TDRF. Cluster II comprises four accessions Cr00SAF, Cr00WFRE2, Cr00PFWE and Cr00CHEF. Cluster III comprises ten accessions Cr00LPNF, Cr00LP, Cr00PFYE, Cr00PLMF, Cr00WFSRE, Cr00SNFF, Cr00WFYE, Cr00DP, Cr00CAF and Cr00WFRE. The clustering outcomes represent the genetic similarity and relatedness among the studied accessions. The bootstrap value showed the relationship stability. High bootstrap values indicated a strong statistical support of outcomes based on genetic analysis. Dendrogram clustering grouped different *Catharanthus* accessions based on ESR-SSR markers response that till had divided only on their morphology.

Discussion

Publically available EST data has been already proven to be a potential and most valuable resource for the identification and developing of large scale SSRs molecular markers to examine the evolutionary studies. As EST markers are derived from the genomic regions which can be transcribed, these are also conserved across the various genomes in a wider range then the other markers (Pashley *et al.*, 2006). In this study, 22867EST sequences have been searched for *C. roseus* to find EST-SSR markers and finally, 427 SSR were found. We found that di-repeat nucleotides were more abundant (29.5%) among all found markers also earlier reported in other plant species (Feng *et al.*, 2009). In the present investigation, di-nucleotide motif was also found to be very common SSRs, followed by tetra-, tri-, penta- and hexa-nucleotide motifs. The most common and dominant di-nucleotide repeat motif type was AG/CT in the dataset (27.7%). Similar results have been reported in Rice, Soybean, *Arabidopsis*, Oil palm, Barley, Coffee, Wheat, Maize and Rubber (Kantety *et al.*, 2002). The high level of occurrence of AG/CT repeat motif results the high translated amino acid product. These findings are accordingly the agreement with the important Chinese traditional medicinal plant (*Epimedium sagittatum)* (Zeng *et al.*, 2010). Further analysis revealed in the tri-nucleotides the repeat motif AAG/CTT was most frequently occurring (29.5%) in *C. roseus* followed by AGC/CTG (11.4%) and AAT/ATT (10.6%). The previous studies of Soybean (*Glycine max*) (Gao *et al.*, 2003) and model plant like *Arabidopsis* (Cardle *et al.*, 2000) also suggested that AAG tri-nucleotide motif may

be very common in dicots. Similarly, AAT/ATT was predominant in genomic SSRs region of *Dendrocalamus latiflorus* (Bhandawat *et al.,* 2017). $(AAG)_n$ is the richest recurrent motif in *Gossypium barbadense* (Lu *et al.,* 2010). $(AAG)_n$ to be very commonly found repeat motif in Turmeric (Siju *et al.,* 2010). The most frequently occurring tetrameric repeat motif was AAAG/CTTT (18%), followed by AATT/AATT (14%).

For the functional annotation and characterization of the assembled contigs, all assembled contigs were considered to identify SSRs and contigs with SSR regions were executed for functional analysis by comparative analysis with publicly available databases like NCBI-nr, Swiss-Prot and TAIR and resulted annotation for 293 contigs (81.84%) contigs. It was also observed that the top BLAST similarity of contigs was with the members of Rubaceae family. In the present study, the result also indicated that most of the annotated genes were not annotated earlier; otherwise it would have resulted in higher similarity. *C. roseus* showed highest sequence homology with members of Solanceae and Vitaceae.

Gene ontology analysis was performed, which provided significant information in the form of 3 major categories- Biological, cellular and molecular functions. In case of biological process, the cellular process was the highest, followed by metabolic process in *C. roseus.* Observations for cellular process were reported in *Physcomitrella patens* (Xiao *et al.,* 2011) and metabolic process in *Tortula ruralis* (Triwitayakorn *et al.,* 2002). Molecular function for binding was observed maximum followed by catalytic, transcription regulator, transporter and structural molecule, also reported earlier (Triwitayakorn *et al.,* 2011), where most categories fall into binding and catalytic activity among molecular function. In cellular component, most of unigene involved in cell, cell parts and organelles, which was consistent with the earlier report (Xia *et al.,* 2011). KEGG pathway annotation using enzyme commission number assigned to annotated sequences is a very important approach for functional genomics, which identify the different biochemical and metabolic pathways. It was observed that the pathways involved in metabolism of various core components of sugar, synthesis of secondary metabolites, microbial metabolism in diverse environment, biosynthesis of antibiotics and oxidative phosphorylation were highly rich in *C. roseus.* The result of annotation of SSR-EST sequences helps in providing comprehensive and fundamental classification with the identification of important pathways (metabolic) from which future studies on the *C. roseus* can be beneficial for alkaloid production.

Of the 28 genic SSRs microsatellite marker validated, 17 (60.71%) produced reliable amplification and 5 (17.85%) of them were found polymorphic. Data analysis shows CR651primerhad highest PIC value of 0.373 with highest expected heterozygosity (H_e) of 0.507 and maximum value of Shannon's Information index of 0.690. Similar results were also reported earlier in *C. roseus* (Mishra *et al.,* 2011; El-Domyati *et al.,* 2012) by studying molecular markers for flower characteristics, producing anti-cancer compounds, using 9ISSRand20 RAPD primers andrevealed the highest PIC of 0.37 for ISSR and 0.35 for RAPD. Highest expected heterozygosity (H_e) value for ISSR and RAPD of 0.49 and 0.46, respectively. Suggesting that SSRs markers used in the present study were polymorphic and suitable for characterizing *C. roseus* accessions in MAS for crop improvement programmes, to analyse the genetic diversity studies through various types of comparative mappings.

Conclusion

In this study molecular markers were developed to highlight the significant insights relevant to this area. A computational based approach was used to develop and identify SSR markers from publicly available EST database, which was further validated by wet lab means. Marker development from coding regions of DNA has great advantage as the gene function previously known may help in utilizing marker for specific trait. The finding from this study may be helpful in fascinating various agronomic interests directly related to the genetic analysis means.

Reference

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde and R.F. Moreno, 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252: 1651–1656

Anderson, J.A., G.A. Churchill, J.E. Autrique, S.D. Tanksley and M.E. Sorrells, 1993. Optimizing parental selection for genetic linkage maps. *Genome*, 36: 181–186

Bhandawat, A., G. Singh, R. Seth, P. Singh and R.K. Sharma, 2017. Genome-wide transcriptional profiling to elucidate key candidates involved in bud burst and rattling growth in a subtropical bamboo (*Dendrocalamus hamiltonii*). *Front. Plant Sci*., 7: 1–16

Cardle, L., L. Ramsay, D. Milbourne, M. Macaulay, D. Marshall and R. Waugh, 2000. Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics*, 156: 847–854

Chen, C., P. Zhou, Y.A. Choi, S. Huang and F.G. Gmitter, 2006. Mining and characterizing microsatellites from citrus ESTs. *Theor. Appl. Genet.*, 112: 1248–1257

Cruz, C.D., 1998. Programa GENES: Aplicativo computacional em estatística aplicada à genética (GENES-Software for Experimental Statistics in Genetics). *Genet. Mol. Biol.*, 21: 1

Doyle, J.J., 1990. Isolation of plant DNA from fresh tissue. *Focus*, 12: 13–15

Du, Z., X. Zhou, Y. Ling, Z. Zhang and Z. Su, 2010. Agri GO: a GO analysis toolkit for the agricultural community. *Nucl. Acids Res.*, 38: 64–70

Durand, J., C. Bodenes, E. Chancerel, J.M. Frigerio, G. Vendramin, F. Sebastiani, A. Buonamici, O. Gailing, H.P. Koelewijn, F. Villani, C. Mattioni, M. Cherubini, P.G. Goicoechea, A. Herran, Z. Ikaran, C. Cabane, S. Ueno, F. Alberto, P.Y. Dumoulin, E. Guichoux, A.D. Daruvar, A. Kremer and C. Plomion, 2010. A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. *BMC Genomics*, 11: 1–13

El-Domyati, F.M., A.M. Ramadan, N.O. Gadalla, S. Edris, A.M. Shokry, S.M. Hassan, S.E. Hassanien, M.N. Baeshen, N.H. Hajrah, M.A. Al-Kordy, O.A. Abuzinadah, A.S.M. Al-Hajr, C.C. Akoh and A. Bahieldin, 2012. Identification of molecular markers for flower characteristics in *Catharanthus roseus* producing anticancer compounds. *Life Sci. J.*, 9: 5949–5960

Ewing, R.M., A.B. Kahla, O. Poirot, F. Lopez, S. Audic and J.M. Claverie, 1999. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.*, 9: 950–959

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

Gao, L., J. Tang, H. Lia and J. Jia, 2003. Analysis of microsatellites in major crops assessed by computational and experimental approaches. *Mol. Breed*., 12: 245–261

Gupta, P.K., S. Rustgi, S. Sharma, R. Singh, N. Kumar and H.S. Balyan, 2003. Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Mol. Genet. Genomics*, 270: 315–323

Heijden, R.V.D., D. I. Jacobs, W. Snoeijer, D. Hallard and R. Verpoorte, 2004. The *Catharanthus* alkaloids: pharmacognosy and biotechnology. *Curr. Med. Chem.*, 11: 607–628

Joshi, R.K., B. Kar and S. Nayak, 2011. Exploiting EST databases for the mining and characterization of short sequence repeat (SSR) markers in *Catharanthus roseus* L. *Bioinformation*, 5: 378–381

Kantety, R.V., M.L. Rota, D.E. Matthews and M.E. Sorrells, 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol. Biol.*, 48: 501–510

Lu, Y., C. Cai, L. Wang, S. Lin, L. Zhao, L. Tian, J. Lu, T. Zhang and W. Guo, 2010. Mining, characterization, and exploitation of EST-derived microsatellites in *Gossypium barbadense*. *Chin. Sci. Bull*., 55: 1889–1893

Mishra, R.K., B.H. Gangadhar, J.W. Yu, D.H. Kim and S.W. Park, 2011. Development and characterization of EST based SSR markers in Madagascar Periwinkle (*Catharanthus roseus*) and their transferability in other medicinal plants. *Plant Omics J*., 4: 1–9

Pashley, C.H., J.R. Ellis, D.E. McCauley and J.M. Burke, 2006. EST databases as a source for molecular markers: lessons from *Helianthus*. *J. Hered*., 97: 381–388

Pinto, L.R., K.M. Oliveira, E.C. Ulian, A.A.F. Garcia and A.P.D. Souza, 2004. Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. *Genome*, 47: 795–804

Ronning, C.M., S.S. Stegalkina, R.A. Ascenzi, O. Bougri, A.L. Hart, T.R. Utterbach, S.E. Vanaken, S.B. Riedmuller, J.A. White, J. Cho, G.M. Pertea, Y. Lee, S. Karamycheva, R. Sultana, J. Tsai, J. Quackenbush, H.M. Griffiths, S. Restrepo, C.D. Smart, W.E. Fry, R.V.D. Hoeven, S. Tanksley, P. Zhang, H. Jin, M.L. Yamamoto, B.J. Baker and C.R. Buell, 2003. Comparative analyses of potato expressed sequence tag libraries. *Plant. Physiol.*, 131: 419–429

Scott, K.D., P. Eggler, G. Seaton, M. Rossetto, E.M. Ablett, L.S. Lee and R.J. Henry, 2000. Analysis of SSRs derived from grape ESTs. *Theor. Appl. Genet*., 100: 723–726

Shaw, R.K., L. Acharya and A.K. Mukherjee, 2009. Assessment of genetic diversity in a highly valuable medicinal plant *Catharanthus roseus* using molecular markers. *Crop Breed. Appl. Biotechnol*., 9: 52–59

Siju, S., K. Dhanya, S. Syamkumar, B. Sasikumar, T.E. Sheeja, A.I. Bhat and V.A. Parthasarathy, 2010. Development, characterization and cross species amplification of polymorphic microsatellite markers from expressed sequence tags of turmeric (*Curcuma longa* L.). *Mol. Biotechnol*., 44: 140–147

Sokal, R.R., 1958. A statistical method for evaluating systematic relationship. *Univ. Kansas Sci. Bull.*, 28: 1409–1438

Triwitayakorn, K., A.J. Wood and F.C. Botha, 2002. Characterization of two desiccation-stress related cDNAs TrDr1 and TrDr2 in the resurrection moss *Tortula Ruralis1*. *S. Afr. J. Bot.*, 68: 545–548

Triwitayakorn, K., P. Chatkulkawin, S. Kanjanawattanawong, S. Sraphet, T. Yoocha, D. Sangsrakru, J. Chanprasert, C. Ngamphiw, N. Jomchai, K. Therawattanasuk and S. Tangphatsornruang, 2011. Transcriptome sequencing of *Hevea brasiliensis* for development of microsatellite markers and construction of a genetic linkage map. *DNA Res.*, 18: 471–482

Vaidya, E., R. Kaur and S.V. Bhardwaj, 2012. Data mining of ESTs to develop dbEST-SSRs for use in a polymorphism study of cauliflower (*Brassica oleracea* var. botrytis). *J. Hortic. Sci. Biotechnol*., 87: 57–63

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

Verma, M. and L. Arya, 2008. Development of EST-SSRs in watermelon (*Citrullus lanatus* var. lanatus) and their transferability to *Cucumis* spp. *J. Hortic. Sci. Biotechnol*., 83: 732–736

Wang, Z., G. Taramino, D. Yang, G. Liu, S.V. Tingey, G.H. Miao and G.L. Wang, 2001. Rice EST with diseases-resistance gene or defense-response gene like sequences mapped to region containing major resistance gene or QTL. *Mol. Genet. Genom*., 265: 301–310

Xia, Z., H. Xu, J. Zhai, D. Li, H. Luo, C. He and X. Huang, 2011. RNA-Seq analysis and *de novo* transcriptome assembly of *Hevea brasiliensis*. *Plant Mol. Biol.*, 77: 1–10

Xiao, L., H. Wang, P. Wan, T. Kuang and Y. He, 2011. Genome-wide transcriptome analysis of gametophyte development in *Physcomitrella patens*. *BMC Plant Biol*., 11: 1–16

Yeh, F.C., 1999. POPGENE (version 1.3. 1). Microsoft Window-Bases Freeware for Population Genetic Analysis. http://www. ualberta. ca/~ fyeh/

Zeng, S., G. Xiao, J. Guo, Z. Fei, Y. Xu, B.A. Roe and Y. Wang, 2010. 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: 94

Zhu, Y., Y. Hao, K. Wang, C. Wu, W. Wang, J. Qi and J. Zhou, 2009. Analysis of SSRs information and development of SSR markers from walnut *ESTs*. *J. Fruit Sci.*, 26: 394–398

(Received 22 November 2018; Accept 31 December 2018)