Utilization of in silico EST–SSR markers for diversity studies in castor (*Ricinus communis* L.)

Ramesh Thatikunta¹ · A. Siva Sankar¹ · J. Sreelakshmi¹ · Gouthami Palle¹ ·
C. Leela¹ · Ch. V. Durga Rani² · V. Gouri Shankar³ · B. Lavanya¹ ·
P. Narayana Reddy⁴ · M. Y. Dudhe⁵

Received: 11 January 2016/Revised: 13 July 2016/Accepted: 25 July 2016/Published online: 8 August 2016
© Prof. H.S. Srivastava Foundation for Science and Society 2016

Abstract Castor (*Ricinus communis* L.) a chief non-edible oilseed crop has numerous industrial applications. Systematic genetic diversity analysis utilizing DNA based markers has been quick and reliable method that ensures selection of diverse parents for exploitation of higher levels of heterosis in breeding programs. From NCBI database, 63,852 EST sequences of castor were mined. One thousand one hundred and five (1105) EST–SSRs and 1652 repeat motifs sequences were identified from 20,495 non-redundant unigene sequences. Repeat motifs consisted of 29.7 % mono nucleotide repeats, 24.8 % di nucleotide repeats, 27.27 % tri nucleotide repeats and 3.94 % tetra nucleotide repeats. Twenty eight primer pairs were chosen from SSR-containing ESTs to determine genetic diversity among 27 castor accessions. Twelve EST–SSRs showed polymorphism. Number of alleles detected were 2–3 with an average of 2.33 per locus. 150–400 bp was the size of an allele. Dendrogram analysis grouped the 27 accessions into two separate clusters. Genetic similarity coefficient of

dendrogram ranged from 0.24 to 0.83. The polymorphic information content value of 0.28–0.49 revealed medium level of diversity in castor. Results of present study indicated that EST–SSRs to be efficient markers for genetic diversity studies. Knowledge on level of diversity existing in castor genotypes would be useful for breeders to plan efficient hybrid breeding programme.

Keywords Castor · EST–SSRs · Polymorphism · Genetic diversity

Abbreviations

EST	Expressed sequence tag
SSR	Simple sequence repeat
PIC	Polymorphic information content
MNR	Mono nucleotide repeats
DNR	Di nucleotide repeats
TNR	Tri nucleotide repeats
TTNR	Tetra nucleotide repeats
NCBI	National Center for Biotechnology Information
UPGMA	Unweighted pair group method with arithmetic mean
UTR	Untranslated region

Electronic supplementary material The online version of this article (doi:10.1007/s12298-016-0367-x) contains supplementary material, which is available to authorized users.

✉ Ramesh Thatikunta
thatikuntaramesh@gmail.com

- ¹ Department of Crop Physiology, College of Agriculture, Rajendranagar, Hyderabad 500030, India
- ² Institute of Biotechnology, PJTSAU, Rajendranagar, Hyderabad 500030, India
- ³ SRTC, PJTSAU, Rajendranagar, Hyderabad 500030, India
- ⁴ Department of Plant Pathology, College of Agriculture, Rajendranagar, Hyderabad 500030, India
- ⁵ ICAR-Indian Institute of Oilseeds Research, Rajendranagar, Hyderabad 500030, India

Introduction

Castor (*Ricinus communis* L., $2n = 20$) belongs to *Euphorbiaceae* family. The crop has use as non-edible oil across the world. Major castor growing countries include India, Brazil, China, Russia and Thailand (Nagesh Kumar et al. 2015). India accounts for a total production of 17.33 lakh tonnes from an area of 11.05 lakh ha and a productivity of 1568 kg ha^{-1} during 2014–2015. Castor seeds contain highest oil content of 40–55 % when compared

with other oilseed crops. Kernels contain oil that ranges between 64–71 %. Among all the vegetable oils, castor oil has the distinct advantage due to its high level (>85 %) of fatty acid known as ricinoleic acid which makes it world's most useful and economically important natural oil. It has the best raw industrial oil owing to its highly stable nature and as it shows minimal variation in fatty acid. Castor oil and its derivatives mainly used in industry sector (Ogunniyi 2006) and castor cake is very useful as organic manure. The non-edible nature of castor oil is due to the presence of ricin / *Ricinus communis* agglutinin (RCA).

Castor a highly cross pollinated crop has recorded 15–20 % of heterosis in the hybrids. India has been credited with release of world's first castor hybrid GCH-3. Several high yielding hybrids coupled with resistance / tolerance to various biotic and abiotic stresses were released and are under commercial cultivation in varied agro-climatic conditions. Already released hybrids occupy most of the castor growing area because hybrids are high yielders than natural varieties (Gouri Shankar et al. 2013). Selection of diverse potential parents is necessarily used to exploit higher levels of heterosis in hybrid breeding programme. Genetic diversity primarily has been assessed with both conventional and molecular markers. In the recent past large number of DNA based markers have been developed, such as restriction length polymorphism (RFLP), random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), simple sequence repeat polymorphism (SSR), intron length polymorphism (ILP) and single nucleotide polymorphism (SNP) (Dudhe et al. 2014). Conventional diversity analysis procedures consume time, labour intensive and affected by environments. Genetic diversity assessment utilizing DNA based markers has been rapid and reliable. To study the extent of genetic diversity in castor, DNA based markers have been extensively used viz., RAPD, AFLP, RFLP and SSR (Anjani 2012). It has been well known that SSR markers are co-dominant, abundantly found throughout the genome and highly polymorphic, easily reproducible, hence widely favored by researchers. The expressed sequence tags (ESTs) which is expressed portion of gene is a valuable genomic resource for the mining of gene targeted SSR markers. Due to availability of large number of bioinformatic tools and huge sequence level information, mining of EST–SSRs and genomic SSRs has gained importance in recent past (Dudhe and Sarada 2012). The interspecific transferability of EST–SSRs and genomic SSRs has been well proved (Wen et al. 2010). EST–SSR markers have been widely applied in gene tagging, linkage map construction and QTL mapping (Qiu et al. 2010). Hence, for various crops, EST–SSRs have been developed and widely used in crop improvement (Qiu et al. 2010). In Castor the availability of EST–SSR markers are limited when compared with other crops viz., wheat, rice,

maize, chickpea, pigeon pea, groundnut, sunflower and rapeseed mustard. Hence, the present study was taken up to mine EST–SSR markers by using available ESTs in public domain, to study the occurrence and distribution pattern of motifs and their loci and to prove their working nature by utilizing them in the genetic diversity study in certain castor genotypes.

Materials and methods

Twenty seven castor accessions including checks were grown in a randomized block design with two replications with four plants each. Spacing of 90 × 60 cm was followed. The plants were grown during *Rabi*, 2012. Pistillate and male lines used in present study are the parental lines of the existing hybrids. The other elite material used is either at par or superior to existing parental lines or existing checks for oil and yield parameters. The elite lines combine the genes from different existing donors for particular trait. These lines were systematically evaluated at different AICRP castor research stations viz., IIOR, Hyderabad; RARS, Palem (PJTSAU, Telangana); Junagadh (JAU, Gujrat) and Sardar Krishi Nagar (SDAU, Gujarat) for particular trait. The information on characters like stem colour, capsule nature, bloom, pistillateness, disease and pest tolerance of the accessions used in the study have been presented (Table 1).

DNA extraction, amplification

Leaf material was collected at 30 DAS. DNA was extracted by utilizing CTAB (Doyle and Doyle 1990). Purified total DNA was checked for quality utilizing spectrophotometer (Glaser 1995). DNA in samples was quantified by loading 1 µl on to 0.8 % agarose gels. To assess the quality and the quantity uncut λ DNA was loaded as control. DNA was amplified in 10 µl containing 25 ng µl⁻¹ DNA template 3 µl, 1× PCR reaction buffer 15 mM Tris HCl 1 µl, 1U Taq DNA polymerase (Genei, Bangalore), forward primer 10 µM, reverse primer 0.5 µM, 0.2 mM dNTPs 1 µl and sterile distilled water 3.9 µl. PCR was carried out in a Veriti™, 96-well Thermal cyclor of Applied Biosystems, CA, USA. The PCR set up used was initial denaturation at 94 °C for 5 min, followed by 35 cycles of 45 s at 94 °C denaturation, 1 min at specific annealing temperatures and 1 min at 72 °C extension followed by a final extension at 72 °C for 10 min. PCR products were separated by electrophoresis in 3 % Lonza agarose gel in 0.5× TAE buffer at 120 V with 0.05 µg µl⁻¹ ethidium bromide used for screening of genotypes. Samples were loaded along with 100 bp ladder as reference (Genei, Bangalore). Gels were documented after separation of bands using the Molecular Imager Gel Doc (Bio-RAD).

Table 1 Information on morphological diversity in castor accessions

Genotypes	Source	Information about genotypes	Morphological traits		
			Stem colour	Capsule	Bloom
1. DPC-9	IIOR, Rajendranagar	Pistillate line	Green	Spiny	Zero
2. RG-48	IIOR, Rajendranagar	Elite line (germplasm accession)	Red	Non spiny	Double
3. RG-43	IIOR, Rajendranagar	Early flowering, resistant to reniform nematode and leaf hopper	Red	Spiny	Triple
4. RG-1354	IIOR, Rajendranagar	Fusarium wilt resistant	Green	Spiny	Double
5. RG-47	IIOR, Rajendranagar	Fusarium wilt resistant and root rot	Green	Spiny	Triple
6. RG-67	IIOR, Rajendranagar	Elite line (germplasm accession)	Green	Spiny	Double
7. RG-1686	IIOR, Rajendranagar	Male line (germplasm accession)	Green	Spiny	Double
8. DCS-78	IIOR, Rajendranagar	Male line	Green	Spiny	Double
9. RG-20	IIOR, Rajendranagar	Elite line (germplasm accession)	Green	Non spiny	Double
10. PCS-293	RARS, Palem, Telangana	Elite line	Green	Spiny	Double
11. PCS-312	RARS, Palem, Telangana	Elite line	Red	Spiny	Double
12. PCS-106	RARS, Palem, Telangana	Elite line	Green	Spiny	Double
13. PCS-236	RARS, Palem, Telangana	Elite line	Red	Spiny	Double
14. PCS-252	RARS, Palem, Telangana	Elite line	Red	Spiny	Double
15. PCS-230	RARS, Palem, Telangana	Elite line	Green	Spiny	Double
16. PCS-171	RARS, Palem, Telangana	Elite line	Green	Spiny	Triple
17. PCS-265	RARS, Palem, Telangana	Elite line	Red	Spiny	Double
18. PCS-228	RARS, Palem, Telangana	Elite line	Red	Spiny	Double
19. PCS-224	RARS, Palem, Telangana	Elite line	Green	Spiny	Double
20. PCS-302	RARS, Palem, Telangana	Elite line	Green	Spiny	Double
21. JP-65	JAU, Junagadh,	Pistillate line	Green	Spiny	Zero
22. SKI-215	S.K. Nagar, Gujarat	Male line, Fusarium wilt resistant	Red	Non spiny	Double
23. M-574	IIOR, Rajendranagar	Pistillate line	Green	Spiny	Triple
24. RG-1(Aruna)	IIOR, Rajendranagar	Check variety	Green	Spiny	Double
25. Kiran	RARS, Palem, Telangana	Check variety	Red	Non spiny	Double
26. Kranthi	RARS, Palem, Telangana	Check variety	Red	Spiny	Double
27. Haritha	RARS, Palem, Telangana	Check variety	Green	Spiny	Double

Detection of EST SSR containing sequences

In the present study, a total of 63,852 ESTs (54.4 Mb) were downloaded in FASTA format from National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/). Total length of EST sequence was 45,046,231 bp with an average of 706 space length base pair. EST sequences were assembled by CAP3 software to seek unigene sequences (<http://genome.cs.mtu.edu/sas.html>). These sequences were used to identify microsatellites utilizing SSR Locator (da Maia et al. 2008). SSRs (motifs ranging from 1 to 10 nucleotides) were searched to evaluate the pattern of EST–SSRs distribution. The present study was taken up to select the repeat number parameters ≥ 20 for mono, ≥ 10 di, ≥ 7 for tri, ≥ 5 for tetra, ≥ 4 for penta and hexa and ≥ 3 for hepta, octa, nano and deca nucleotides respectively. Since,

longer SSRs had higher probability to be polymorphic, longest sequence used for SSR identification was 1361 bp.

Primer designing and in silico PCR of EST–SSRs

Non-redundant 20,495 unigene sequences containing EST sequences were used for primer designing utilizing standard parameters of primer3 software inbuilt in SSR locator. The parameter set up used for primer designing were PCR primer length of 15–25 and 20 bp as optimum with product size of 150–400 bp, 55–61 °C annealing temperature with 60 °C as optimum and 40–60 % minimum GC content with 50 % being the optimum. In silico PCR was run which has inbuilt SSR Locator software to identify the suitable primers that yielded desired product size.

Data analysis

List of 28 randomly selected EST–SSR markers with expected amplicon size has been represented (Table 2). Accessions were genotyped based on the presence (1) or absence (0) of peaks represented as bands for each EST–SSRs. Faint or unclear bands were not considered for scoring to calculate the polymorphic information content (PIC) value (Anderson et al. 1993). The qualitative data generated was used to generate a matrix determining similarity among the samples. Similarity matrix was used to draw a dendrogram with Unweighted Pair Group Method with Arithmetic Mean (UPGMA) clustering method via SHAN module of NTSYS-pc version 2.2j software of Applied Biostatistics Inc, Setauket, NY, USA for a phylogenetic representation of genetic relationships calculated by using the similarity coefficient (Sneath and Sokal 1973).

Results

Detection and distribution of EST–SSR motifs

From 63,852 ESTs, 20,495 (32.09 %) non-redundant unigene sequences were identified (Table 3). These 20,495 unigene sequences were used for SSR mining as explained in materials and method. The longest sequence containing microsatellite was of 1361 bp. Total 1105 microsatellites have been identified from 63,852 ESTs which contributes to 1.73 % (Supplementary Table 1). Twenty eight primer pairs were synthesized for further study which accounts for 2.53 % of the total number of EST–SSRs identified. The number and percentages of the identified repeat motifs revealed that mono, di and trinucleotides are present in large numbers and in more or less in equal proportion (Table 4).

Occurrence patterns of motifs and their loci

Number and percentages of repeat motifs identified have been presented (Table 4). For 4-mer to 6-mer only frequent motifs have been studied and presented. For other motifs all possible arrangements were studied and presented (Supplementary Table 2). A/T monomer repeats were found in 475 loci with breakup of 260 (55 %) and 215 (45 %) loci formed by A and T nucleotides respectively. C/G motifs were found in 7 loci with breakup of five (71 %) and two (29 %) respectively. SSRs with predominant repeats were A/T i.e. 100 % monomer loci. In the present study monomers occurred at 29.17 % of 1672 identified loci. AG/CT motifs were the most abundant

adding upto 60 and 92 loci respectively. For the tri repeats AAG/CTT are the most frequent and comprised of 25 and 33 loci respectively. Among 4-mers, many different arrangements were found. AAGA/TCTT motif has been presented at 10 loci, 5-mers AGAAA/TTTCT at 7 loci, 6-mer AGAAGG/CCTTCT at 1 locus. From all the remaining repeats occurrences have been widely distributed with low percentage values for each arrangement. For 7-mer, 8-mer, 9-mer and 10-mer repeats, total occurrences were found to be 35, 4, 21 and 4 respectively.

Distribution of amino acids, number of loci and total repeats

Another area of this investigation has been to study the abundance of amino acids predicted from EST–SSRs. The distribution of amino acids has been presented (Supplementary Table 3; Fig. 1). Among these leu was most abundant (92 % loci and 663 repeats), followed by ser (91 % of loci and 589 repeats), Phe, Ile, Gln, Lys, Pro, Glu, Arg, Ala, His (50–36 % loci and 372–212 repeats) followed by Thr, Term, Asn, Gly, Cys and Asp (34–21 % of loci and with 196–119 repeats). Met, Trn, Val and Tyr have least percentage of loci (12–6) and repeats (92–30).

Genetic diversity

Castor cultivation has been for seed oil in tropical and subtropical regions. In the present study, 12 polymorphic EST–SSRs detected twelve alleles. A wide range of fragments from 150 to 400 bp were observed. Castor genotypes were tested with 28 primer sets. Among them, 12 primer sets showed polymorphism with PIC values that ranged between 0.28 and 0.49 with an average of 0.339. Remaining primer set did not show any polymorphism. Number of alleles detected by individual EST–SSR averaged 2.3. Cluster analysis was performed to generate a dendrogram using 27 castor genotypes (Fig. 2). Cluster I included 26 accessions (96.30 %) while cluster II accommodated single accession (3.70 %) with 0.83 % similarity coefficient. M-574 a single genotype formed separate OUT in cluster showed less similarity coefficient of 0.24 %. M-574 is a pistillate (female) line generated through mutation breeding of VP-1. VP-1 genotype is widely used female parent in castor hybridization programmes and is female parent for several commercially released hybrids. However, it is susceptible to wilt. VP-1 is subjected to mutations using gamma rays to develop a line for wilt tolerance or resistance. M-574 is one of the derivatives of the mutation breeding with wilt resistance.

Table 2 List of 28 EST–SSR markers with expected amplicon size

S. no.	Locus name	Forward and reverse primers sequences	Amplicon size
1	RcDES593	F:GTTCAAGCTCTAGTTCGGTGAG R: TACACTTCTGTCTGTCCATGC	371
2	RcDES 598	F:GAAAGCAACAAGAAAAGGTCTG R: AAGCTGGTGGTTTACTTGCTAC	346
3	RcDES 599	F:CTTCATTAAGCCATCAAAGAGC R:AATGTAGTGTCTGTCTGTTGCG	256
4	RcDES 596	F:TCCACATTGCTAACAAGCATAG R:AAATGCACACAGCTAAGACAAG	381
5	RcDES 597	F:CAGCCAAACATAAGATTCATGC R:ATCATCTGGGTGTCTCAAAATC	379
6	RcDES 595	F:TCTCATCACTAACAACAGGCC R:GGTTCTGAAAGAAGTGAAATGG	128
7	RcDES 592	F:GGCTTCTCCTCTGTTTGTATG R: ACCACATGCAACAATCACAAAG	139
8	RcDES 594	F:GGCTTCTAATCTTTACTTCACC R: TATTTTCATCACCGACCCTAAC	35
9	RcDES 600	F:CCGCTCACTACAAGTGGTACTC R:GAGTACTGGACAGCGATGAGAC	387
10	RcDES 601	F:TAACCTGTCTGTTTGGGTGAC R:GATTGGAGTCATGGAGAGAGAG	258
11	RcANGRAU 1 (3)	F:ACGAGGCTCTGTCTCTGTGT R:GGCAAAATTCAACACCATTTC	125
12	Rc ANGRAU 18 (34)	F:TTAGGGTTTGGTCTTTTGA R:TCAAGTGCCCATTTTAGAGC	186
13	Rc ANGRAU 20 (39)	F:AGCAGCAACAACTACGCTT R:GTTGCATCGATAGAGCCTGT	279
14	Rc ANGRAU 22 (41)	F:AAAAGGTGGGATAGAGCCTGT R:CCATCAAGAATACCCTCCCT	400
15	Rc ANGRAU 25 (46)	F:CTGCTTCTCTCTGCATTGTG R:CATGGGATCTTTGGTTCTTG	276
16	RcANGRAU 33	F:ATTTATTGCATGTTGCAGCC R:GCAAAGTTCACCGATGAGAA	378
17	Rc ANGRAU 34(56)	F:AAAAGACAAAGCAGCACCAC R:GCAGCTCCGAGTTGTACG	225
18	Rc ANGRAU 43 (65)	F:AACGAAGCATTTTAATCCCC R:GTGAGCGAGATTATCGGAGA	367
19	Rc ANGRAU 46 (68)	F:TCCCATCAGTTTTTGCTGTT R:AGAATAGGATCCATCCTGCC	259
20	Rc ANGRAU 6 (13)	F:ATACTCTCAGTGCTGCTGCC R:AAAACGTGTAAACGGGATCA	252
21	GB RC 019	F:GTATGGTACGATCTCTTTGGAC R:TACGCAGAGAAGCACTAATAGA	277
22	GBRC 021	F:AAGCTCAACTTAAAGCCTAGA R:ATTTTGGTGGTCTCTAGTTCAGTC	211
23	GBRC 080	F:TTGAGGAAACAGAAGATCAAAT R:AGCACGCCAATACCTCTTGTA	213
24	GB RC 105	F:TAGATTTTTATGGATAGGTGCC R:ATGTAGACACTTGACTCACGAA	206

Table 2 continued

S. no.	Locus name	Forward and reverse primers sequences	Amplicon size
25	Rc DES 28	F:CCCCAAAGACTAACAACAACC R:ATGCATCTGTTGGAGTTGC	239
26	Rc DES 45	F:CACAAACACACATATCATGTCC R:CTCAAGTGCATCTGAAACG	218
27	Rc DES 46	F:ACGAGGAGGGAGACTAAATGC R:CACTGATATACACACCACAGTGAC	229
28	Rc DES 100	F:CTGGCATTGCAGATCGTATGA R:GCGCCACCACCTTGATCTT	233

Table 3 Summary statistics of ESTs assembled and total number of EST–SSRs identified

Feature	Statistics
Total number of ESTs used	63,852 (100 %)
Total length of EST sequence	45,046,231 bp (100 %)
Average EST length	720 bp (0.001 %)
Unigene sequences identified	20,495 (32.09 %)
Longest sequence containing EST–SSR	1361 bp (0.003 %)
Total number of EST–SSRs identified	1105 (1.73 %)
Total number of primers synthesized	28 (2.53 %)

Discussion

Frequency and distribution of EST–SSRs

Due to availability of large scale genomic information for majority of oilseed crop, the in silico studies have been an important area of interest for the identification of novel useful genes (Dudhe et al. 2012). Genomic resources available for castor have been successfully utilized for the development of SSR markers but still the number of markers available for castor improvement has been limited. It is well known fact that SSRs can be mined from genic and non genic regions. Earlier in castor Sharma and Chauhan (2011) mined 73 % of SSRs from genic regions and 27 % in non genic regions. SSRs were dense in the putative genic regions of 5'UTRs (26 %), introns (25 %), 3'UTRs (16 %) and in exons (6 %). SSRs were present at very high frequency in castor (one SSR per every 1.7 kb or 2.45 EST or 67 % of sequenced clones) as given in table owing to its small genome size (680 bp to 1.2 kb) when compared to either *Jatropha* or *Hevea* or *Cassava* (2.2–7 kb) to which castor belongs or even compared to

Fig. 1 Total number of amino acid distribution, total repeats and number of loci identified

Rice (6 kb) or *Arabidopsis* (6 kb) or even plant genome (6.4 kb) (Table 5). High density of SSRs (excluding mono repeats) present in castor genic sequences, compared to smaller SSRs makes them liable to polymorphism (Qiu et al. 2010). They identified 7732 SSR sites from 5122 ESTs (unigene) in total of 18,928 ESTs (17.03 Mb). In our study we have mined 1672 SSR sites from 20,495 ESTs (unigene) in total of 63,852 ESTs (54.4 Mb). We have downloaded nearly 3 times more EST sequences and identified 4 times unigene compared with Qiu et al. (2010) which is the greatest advancement of this study hence, in the present study the chances of getting novel EST–SSRs are more.

Sharma and Chauhan (2011) stated that castor genome has a lot of repeat/motif variation. SSRs in UTR regions showed 80.2 % MNRs, 70.3 % DNRs, 83.2 % TTNRs and 77.1 % PNRs as compared to exons which contained 76.1 % TNRs and 73.2 % HNRs. Among the MNRs, A/T accounted to 98.54 % in our study, which indicates the larger presence of these repeats (Supplementary Table 2)

Table 4 Number and percentages of repeat motifs identified

Mono	Di-	Tri-	Tetra-	Penta-	Hexa-	Hepta-	Octa-	Nova-	Deca-	Total
483	411	456	66	65	127	35	4	21	4	1672
28.89 %	24.58 %	27.27 %	3.94 %	3.89 %	7.6 %	2.09 %	0.24 %	1.26 %	0.24 %	100 %

Fig. 2 UPGMA dendrogram of genetic relationship among 27 castor accessions analysis using 28 EST–SSR markers

whereas Qiu et al. (2010) reported 96.9 %. Hence, from both the studies it can be concluded that in castor genome MNR, A/T are the richest repeat.

DNRs varied between 24.8 and 80.4 % in the castor genome (Table 6). AG/GA and AG/CT were the dominant motifs. DNRs were present in non genic regions, introns, 5'UTRs and 3'UTRs. Earlier Seo et al. (2011) observed di-

nucleotide repeats to range from 4 to 52 (mean 12.4). Generally AG/CT and AC/GT abundance has been common in plant genome. Higher frequency of DNR motif is a characteristic of plant genomes (Wang et al. 1994). It is evident from earlier reports that DNRs in SSRs from 5'UTR regions with repeat unit length of 16–30 bp showed higher polymorphism (Sharma and Chauhan 2011). GA/CT

Table 5 Frequency of SSRs in the genome of various crop plants

Crop	Frequency of SSRs in non redundant sequences	Reference
Castor	1.7 kb or one SSR every 2.45 EST	Qiu et al. 2010
Castor	66.7 % of sequenced clones	Seo et al. 2011
Castor	1.2 kb	Pranavi et al. 2011
Castor	680 bp	Sharma and Chauhan 2011
Jatropha	5.95 kb	Yadav et al. 2011
Hevea	2.2 kb	Feng et al. 2009
Cassava	7.0 kb	Raji et al. 2009
Rice	6 kb	Varshney et al. 2002
Arabidopsis	6 kb	Cradle et al. 2000
Plant genome	6.4 kb	Gupta and Varshney 2000

motifs were understood to be high due to high level of occurrence of the translated amino acid products of the motifs (Kantety et al. 2002). Slipped strand mispairing during replication and mutations have been the cause for the variation in frequencies and abundance of different repeat motifs in SSR length (Levinson and Gutman 1987). Even in the present study, AG/CT (36.9 %) was followed by GA/TC (30.9 %). Least present were CA/TG and AC/GT (Supplementary Table 2). The result confirms the finding that AG/CT repeat motif occurs predominantly with a frequency of about twice that of AC/GT (Seo et al. 2011). Abundance of AG/CT regions can be explained by the fact that they were the preferred regions for amplification. On the other hand, it was recorded that DNA fragment that was directed to primers with AT microsatellite repeats did not get amplified (Vasconcelos et al. 2012).

Overall TNRs varied between 17 and 47.8 % in the castor genome. TNR number ranged from 4 to 56 with a mean of 7.35. Tri-nucleotide repeats were dominant (5.2–25.9 %) followed by di-nucleotide repeat motifs. Among TNRs, AAG/CTT was the most dominant, followed by AAG/AGA and AAG/CTT. In cereals also, TNRs were recorded and the most frequent followed by DNRs and TTNRs when MNRs were excluded (Varshney et al. 2005). CCG/CGG was relatively rare in dicotyledonous plants (Morgante et al. 2002). Tri motifs were harbored in exons. The di and tetra motifs / SSRs present in UTR regions were identified to give higher polymorphism when compared with tri motifs which are abundant in exonic regions. In earlier reports, di, tri and tetra motif loci polymorphism proportion amounted to 54.8, 28.8 and 47 % whereas TTNRs and PNRs were randomly distributed (Sharma and Chauhan 2011). Higher order repeat motifs (more than three bases) as well as the compound repeat were found to have less polymorphism as compared to lower order repeats (Feng et al. 2009). Such amplification and polymorphism represented primers which contained more than 15 repeat units (Sharma and Chauhan 2011). The

present investigation showed MNRs, DNRs and TNRs to be in equal proportions (29.7, 24.8 and 27.6 %) and followed by other repeats (Table 4).

TNRs have been understood to be runs of particular amino acids. Most frequent amino acid runs in castor SSRs as reported were those belonging to serine (TCT)_n, glutamate (GAA)_n, arginine (CGC)_n and phenylalanine (TTC)_n (Sharma and Chauhan 2011). In plant ESTs, TNRs abundance in SSRs may be attributed to the absence of frame shift mutations within the transcribed genes which have resulted in the suppression of non-tri nucleotide SSRs in coding regions (Metzgar et al. 2000). Among the TNRs, codon repeats correspond to small hydrophilic amino acids which have been easily tolerated than hydrophobic and basic amino acids (Katti et al. 2001). Four thousand nine hundred and seventy eight (4978) codons were identified in castor EST–SSRs sequence data. Maximum codons were of leucine 663 repeats (13.31 %, CTT) followed by serine 589 repeats (11.83 %, TCT) and tyrosine with least 6 repeats (0.12 %, TAT) (Supplementary Table 3). Most frequent amino acid runs reported in castor included 16.5 % serine (TCT)_n followed by 13.6 % glutamate (GAA)_n, 12.3 % arginine (CGC)_n and 9.7 % phenylalanine (TTC)_n (Sharma and Chauhan 2011). Few other essential amino acids more than 5 % present were phenylalanine (7.47 %, TCT), isoleucine (7.03 %, ATT) and lysine (5.68 %, AAA). The number of loci identified was maximum for leucine (92 loci) and minimum for tyrosine (6 loci) (Fig. 1).

The predicted genome size of castor was estimated to be 323 Mb (Arumuganathan and Earle 1991). In the present study we have mined 54.4 Mb ESTs data to identify 1105 EST–SSRs. The high frequency and distribution of EST–SSRs in this study may be attributed to small genome size and due to criteria used for SSR search by using SSR Locator. Variation in the frequency in SSRs in the EST of a particular species have been attributed to the criteria adapted in SSR search in the database mining approaches and to the quantity of data used to identify SSRs (Varshney et al. 2002 and Yan et al. 2008).

Table 6 Distribution and abundance of repeat motif in Castor genome expressed in percent

DNR			TNR			TTNR			Dominant repeat	Reference
DNR	% in genome	DNR %	TNR	% in genome	TNR %	TTNR	% in genome	TTNR %		
AT	51	43	TCT	29	–	–	12	–	DNR	Sharma and Chauhan (2011)
AG/GA	62.4	68.9	AAG/AGA	33.5	25.9	–	2.2	–	DNR	Pranavi et al. (2011)
AG/CT	80.4	51.7	CTT/AAG	17	5.2	AAAT/AGAT	2.6	0.7	TNR	Seo et al. (2011)
AG/CT	25.1	69.6	AAG/CTT	47.8	23.7	AAAG/CTTT	2.84	40.9	TNR	Qiu et al. (2010)
AG/CT	24.8	36.9	AAG/CTT	27.60	12.7	AAGA/TCTT	2.9	20.8	TNR	Present study

Polymorphism of EST–SSRs

Twelve randomly chosen EST–SSRs were identified as polymorphic with products of expected size which contributes to 42.85 % of the total primer pairs synthesised. The high rate of polymorphism may be due to compound nature of SSRs. Twenty eight alleles were detected with an average of 2.3 alleles for each locus containing TNRs as a dominant motif. In previous studies, SSRs with DNRs showed higher allele numbers of 2.7 per locus followed by 2.3 TNRs per locus (Sharma and Chauhan 2011). From 118 loci which contained 350 alleles i.e. 223 from di loci and 107 from tri loci, an average of 2.97 alleles per locus was recorded. Di, tri and tetra motif loci proportion was 54.8, 28.8 and 47 % respectively. The proportion of polymorphic primers was 41.1 % when null allele primers and those that harbor introns were excluded resulting in 2.97 alleles per marker. This revealed that the polymorphic ratio of EST–SSR primers to be at medium level in castor (Qiu et al. 2010).

PIC values for the tested primers varied from a low of 0.27 (Rc DES 596) to a high of 0.49 (Rc DES 592) with an average value of 0.332. Allan et al. (2008) reported 9 genomic SSR markers, 0.403 gene diversity (PIC) with an average of 3.01 alleles per locus. Bajay et al. (2009) developed 12 genomic SSR markers with an average of 0.416 gene diversity (He) and 3.3 alleles per locus. Qiu et al. (2010) reported 118 polymorphic SSR loci, 0.36 PIC and 2.97 alleles per locus which suggested that EST–SSR markers developed had moderate level of polymorphism. Seo et al. (2011) developed 28 polymorphic SSR loci with PIC value of 0.26 and an average of 3.18 alleles per locus. Gajera et al. (2010) reported 5 polymorphic ISSRs with an average PIC value of 0.88. Pecina-Quintero et al. (2013) used the proven SSRs of Bajay et al. (2009) and obtained PIC values that ranged from 0.5 to 0.812 and 5.5 alleles per marker. Data indicated that genetic variation was mostly due to variation within populations, 10.5 % to among populations and 0.5 % to among regions. PIC values as reported can range from 0 to 1. PIC of 0 markers has only one allele and at PIC of 1 marker would have an infinite

number of alleles. More than 0.7 is considered to be highly informative (Hildebrand et al. 1992). All the above results which also include the present study on castor show PIC values between 0.27 and 0.49. PIC values are important criteria to access diversity or level of polymorphism. The PIC value will be almost zero if there is no allelic variation and it can reach a maximum of 1.0 if a genotype has only new allele which is a rare phenomenon. Hence, in our study PIC values indicate the moderate level of polymorphism or gene diversity in castor genome, which may help future researcher in selection of the informative markers.

UPGMA analysis showed 27 accessions to be grouped into two main clusters. Similarity coefficient between accessions ranged between 0.24 and 0.83 (Fig. 2). Cluster I included 26 accessions (96.30 %) while cluster II accommodated single accession (3.70 %) with 0.83 % similarity coefficient. M-574 a single genotype formed separate OUT in cluster showed less similarity coefficient of 0.24 %. M-574 is a wilt resistant pistillate line with condensed internodes and cup shaped leaves, developed from VP-1 through mutation breeding. The castor genotypes used in the study showed variation in their morphological characters (bloom, stem colour and capsule spininess). Despite the observed morphological variation, variability recorded by use of DNA markers could not be observed at molecular level and this was reflected in the low PIC (gene diversity) and clustering of genotypes into few groups. Studies also confirm that *Ricinus* species does not have a specific genetic pattern or markers that do not allow differentiation based on geographical origin and DNA fragments do not appear to be associated with differentiation of population (Pecina-Quintero et al. 2013). The first report which came out with low polymorphism indicated the narrow genetic base (Allan et al. 2008). Milani et al. (2009) studied diversity structure based on phenotypic traits and RAPD markers. They reported accessions from different countries that showed lack of geographically structured clusters. Earlier reports on diversity study conducted with 200 RAPD and 21 ISSR primers revealed low variability (Gajera et al. 2010). Studies with 232 high-quality single nucleotide polymorphisms showed large molecular

variance within populations as compared to those among populations which confirm a very weak geographic stature among castor populations. This result further confirms the previous findings obtained with dominant RAPD markers (Foster et al. 2010).

Finally we conclude that by using SSR Locator tool, castor EST databases can be systematically searched for EST–SSR markers. The newly identified and synthesized EST–SSR markers based on PIC values indicate the moderate level of polymorphism or gene diversity in castor genome. The results of the study inferred that EST–SSR markers have potential to study genetic diversity and can be used for trait marker relationship studies. Plant breeders with the knowledge of genetic diversity of castor genotypes can design well organized crop improvement programme.

Acknowledgments Authors sincerely acknowledge the financial support provided under the University Grants Commission (UGC), New Delhi Major Research Project Grant No. 40-40/2011 (SR) to carry out the present work to the corresponding author and PI of the project.

References

- Allan G, Williams A, Rabinowicz PD, Chan AP, Ravel J, Keim P (2008) Worldwide genotyping of castor bean germplasm (*Ricinus communis* L.) using AFLPs and SSRs. *Genet Resour Crop Evol* 55:365–378
- Anderson JA, Churchill GA, Autrique JE, Tanksley SD, Sorrells ME (1993) Optimizing parental selection for genetic linkage maps. *Genome* 36:181–186
- Anjani K (2012) Castor genetic resources: a primary gene pool for exploitation. *Ind Crops Prod* 35:1–14
- Arumuganathan K, Earle E (1991) Nuclear DNA content of some important plant species. *Plant Mol Biol Rep* 9:208–218
- Bajay MM, Pinheiro JB, Batista CEA, Nobrega MBM, Zucchi MI (2009) Development and characterization of microsatellite markers for castor (*Ricinus communis* L.), an important oleaginous species for biodiesel production. *Conserv Genet* 1:237–239
- Cradle L, Ramsay L, Milroure D, Macaulay M, Marshall D, Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156(2):847–854
- da Maia LC, Palmieri DA, de Souza VQ, Kopp MM, de Carvalho FIF, de Oliveira AC (2008) SSR locator: tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *Int J Plant Genom*. doi:10.1155/2008/412696
- Doyle JJ, Doyle JL (1990) Isolation of plant DNA fresh tissue. *Focus* 12:13–15
- Dudhe MY, Sarada C (2012) Comparative assessment of microsatellite identification tools available in public domain. *DOR News Lett* 18(2 & 3):8–9
- Dudhe MY, Meena HP, Ranganatha ARG, Mukta N, Lavanya C (2012) In silico-identification of conserved domains from EST database in safflower. *J Oilseeds Res* 29(Special issue):178–181
- Dudhe MY, Sudhakarbabu O, Sudhakar C (2014) Development of intron-flanking EST-specific primers from drought responsive expressed sequence tags (ESTs) in safflower. *SABRAO J Breed Genet* 46(1):56–66
- Feng SP, Li WG, Huang HS, Wang JY, Wu TY (2009) Development, characterization and cross-species/genera transferability of EST–SSR markers for rubber tree (*Hevea brasiliensis*). *Mol Breed* 23:83–97. doi:10.1007/s11032-008-9216-0
- Foster JT, Allan GJ, Chan AP, Rabinowicz PD, Ravel J, Jackson P, Keim P (2010) Single nucleotide polymorphisms for assessing genetic diversity in castor bean (*Ricinus communis*). *BMC Plant Biol*. doi:10.1186/1471-2229-10-13
- Gajera BB, Kumar N, Singh AS, Punvar BS, Ravikiran R, Subhash N, Jadeja GC (2010) Assessment of genetic diversity in castor (*Ricinus communis* L.) using RAPD and ISSR markers. *Ind Crops Prod* 32:491–498
- Glase JA (1995) Validity of nucleic acid purities monitored by 260 nm/280 nm absorbance ratios. *Biotechniques* 118:62–63
- Gouri Shankar V, Venkata Ramana Rao P, Bindu Priya P, Nagesh Kumar MV, Ramanjaneyulu AV, Vishnuvardhan Reddy A (2013) Genetic purity assessment of castor hybrids using EST–SSR markers. *SABRAO J Breed Genet* 45(3):504–509
- Gupta PK, Varshney RK (2000) The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* 113:163–185. doi:10.1023/A:1003910819967
- Hildebrand CE, Torney DC, Wagner RP (1992) Informativeness of polymorphic DNA markers, vol 20. Los Alamos Science Inc, New Mexico, pp 100–102
- Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol Biol* 48:501–510
- Katti MV et al (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol Biol Evol* 18:1161–1167
- Levinson G, Gutman GA (1987) Slipped-strand mispairing—a major mechanism for DNA-sequence evolution. *Mol Biol Evol* 4:203–221
- Metzgar D, Bytof J, Wills C (2000) Selection against frame shift mutations limits microsatellite expansion in coding DNA. *Genome Res* 10:72–80
- Milani M, Dantas FV, Martins WFS (2009) Genetic divergence among castor bean genotypes by morphologic and molecular characters. *Revista Brasileira de Oleaginosas e Fibras* 3:61–71
- Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet* 30:194–200
- Nagesh Kumar MV, Gouri Shankar V, Ramya V, Bindu Priya P, Ramanjaneyulu AV, Seshu G, Vishnu Vardhan Reddy D (2015) Enhancing castor (*Ricinus communis* L.) productivity through genetic improvement for *Fusarium* wilt resistance—a review. *Ind Crops Prod* 67:330–335
- Ogunniyi DS (2006) Castor oil: a vital industrial raw material. *Bioresour Tech* 97:1086–1091
- Pecina-Quintero V, Anaya-López JL, Núñez-Colín CA, Zamarripa-Colmenero A, Montes-García N, Solís-Bonilla JL, Aguilar-Rangel MR (2013) Assessing the genetic diversity of castor bean from Chiapas, México using SSR and AFLP markers. *Ind Crops Prod* 41:134–143
- Pranavi B, Sitaran G, Yamini KN, Dinesh Kumar V (2011) Development of EST–SSR markers in castor bean (*Ricinus communis* L.) and their utilization for genetic purity testing of hybrids. *Genome* 54:684–691
- Qiu L, Yang C, Tian B, Yang JB, Liu A (2010) Exploiting EST databases for the development and characterization of EST–SSR markers in castor bean (*Ricinus communis* L.). *BMC Plant Biol* 10:1–10
- Raji AA, Anderson JV, Kolade OA, Ugwu CD, Dixon AG, Lngelbrecht IL (2009) Gene-based microsatellites for cassava

- (*Manihot esculenta* Crantz): prevalence, polymorphisms, and cross-taxa utility. BMC Plant Biol 9:118
- Seo K-I, Lee G-A, Ma K-H, Hyun D-Y, Park Y-J, Jung J-W, Lee S-Y, Gwag J-G, Kim C-K, Lee M-C (2011) Isolation and characterization of 28 polymorphic SSR loci from castor bean (*Ricinus communis* L.). J Crop Sci Biotechnol 14:97–103. doi:[10.1007/s12892-010-0107-7](https://doi.org/10.1007/s12892-010-0107-7)
- Sharma A, Chauhan RS (2011) Repertoire of SSRs in the castor bean genome and their utilization in genetic diversity analysis in *Jatropha curcas*. Comp Funct Genom. 2011:1–9, Article ID 286089. doi:[10.1155/2011/286089](https://doi.org/10.1155/2011/286089)
- Sneath PHA, Sokal RR (1973) Numerical Taxonomy. Freeman Press, San Francisco
- Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett 7:537–546
- Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. Trends Biotechnol 23:48–55
- Vasconcelos S, Onofre AV, Milani M (2012) Molecular markers to access genetic diversity of castor bean. In: Abdurakhmonov I (ed) Current status and prospects for breeding purposes. InTech, Rijeka, pp 201–217
- Wang Z, Weber JL, Zhong G, Tanksley SD (1994) Survey of plant short tandem DNA repeats. Theor Appl Genet 88:1–6
- Wen M, Wang H, Xia Z, Zou M, Lu C, Wang W (2010) Development of EST-SSR and genomic-SSR markers to assess genetic diversity in *Jatropha curcas* L. BMC Res Notes 3:42. doi:[10.1186/1756-0500-3-42](https://doi.org/10.1186/1756-0500-3-42)
- Yadav HK, Ranjan A, Asif MH, Mantri S, Sawant SV, Tuli R (2011) EST-derived SSR markers in *Jatropha curcas* L. development, characterization, polymorphism, and transferability across the species/genera. Tree Genet Genom 7:207–219
- Yan QL, Zhang YH, Li HB, Wei CH, Niu LL, Guan S, Li SG, Du LX (2008) Identification of microsatellites in cattle unigenes. J Genet Genom 35:261–266