American Journal of Plant Sciences, 2014, 5, 1521-1541
Published Online May 2014 in SciRes. http://www.scirp.org/journal/ajps
http://dx.doi.org/10.4236/ajps.2014.511167

# *In Silico* Mining of EST-SSRs in *Jatropha curcas* L. towards Assessing Genetic Polymorphism and Marker Development for Selection of High Oil Yielding Clones

**Neeraj Jain[1*], Ganesh B. Patil[2], Poonam Bhargava[3], Rajani S. Nadgauda[1]**

[1]Plant Cell and Molecular Biology, Indian Institute of Advanced Research, Gandhinagar, India
[2]Molecular Breeding and Applied Genomics, MAHYCO Seeds Ltd., Dawalwadi, India
[3]Gujarat State Biotechnology Mission, Gandhinagar, India
Email: [*]neerajjain@iiar.res.in

Received 14 March 2014; revised 12 April 2014; accepted 22 April 2014

## Abstract

**In recent years, *Jatropha curcas* L. has gained popularity as a potential biodiesel plant. The varying oil content, reported between accessions belonging to different agroclimatic zones, has necessitated the assessment of the existing genetic variability to generate reliable molecular markers for selection of high oil yielding variety. EST derived SSR markers are more useful than genomic markers as they represent the transcriptome, thus, directly linked to functional genes. The present report describes the *in silico* mining of the microsatellites (SSRs) using *J. curcas* ESTs from various tissues viz. embryo, root, leaf and seed available in the public domain of NCBI. A total of 13,513 ESTs were downloaded. From these ESTs, 7552 unigenes were obtained and 395 SSRs were generated from 377 SSR-ESTs. These EST-SSRs can be used as potential microsatellite markers for diversity analysis, MAS etc. Since the *Jatropha* genes carrying SSRs have been identified in this study, thus, EST-SSRs directly linked to genes will be useful for developing trait linked markers.**

## Keywords

**Bio-Diesel, *Jatropha curcas* L., EST, SSR, MAS, Diversity, Marker, Mining, Polymorphism, *In Silico***

## 1. Introduction

In recent years, *Jatropha curcas* L. has gained popularity as a potential biodiesel plant. It is commonly known as

[*]Corresponding author.

**How to cite this paper:** Jain, N., *et al.* (2014) *In Silico* Mining of EST-SSRs in *Jatropha curcas* L. towards Assessing Genetic Polymorphism and Marker Development for Selection of High Oil Yielding Clones. *American Journal of Plant Sciences*, **5**, 1521-1541. http://dx.doi.org/10.4236/ajps.2014.511167

purging nut/Barbados nut. This plant belongs to the family Euphorbiaceae and is a native of Mexico and Central America and was later on introduced in many parts of tropics and subtropics. *J. curcas* is commonly known to be a poisonous plant. It is a semi-evergreen shrub or small tree reaching a height of 6 mt (20 ft). It can survive arid conditions; therefore, can be grown on drylands and wastelands. The seeds of this plant are highly toxic but produce oil that can be used as biodiesel after transesterification, besides that, in soap and candle making. Being traditionally considered as a weed, its oil has recently started gaining importance as "fuel of the future" or "green fuel" and has been in news, with transport companies eager to run trains, cars and aeroplanes using biodiesel to cut down both on cost and pollution.

The oil content in *Jatropha curcas* is reported to be varying between accessions belonging to different agroclimatic zones (40% to 58% in kernels) of India [1]-[3]. In recent years, emphasis has been laid on producing high oil yielding *Jatropha* plant which can be achieved through genetic selection and crop improvement methods. As a means to this end, it is necessary to assess the existing genetic variability and generate reliable molecular markers for selection.

DNA markers are not typically influenced by environmental conditions, therefore, can be used to describe patterns of genetic variation among plant populations and to identify duplicated accessions within germplasm collections [4]. To assess the genetic diversity, several types of popular PCR based markers like, RAPD (Random Amplified Polymorphic DNA) [5], ISSR (Inter Simple Sequence Repeat) [6] [7] and AFLP (Amplified Fragment Length Polymorphism) [8] [9] are routinely used due to the advantage of no requirement of prior sequence information [3].

The existing information regarding the extent and pattern of genetic variation in *J. curcas* population is limited [10]. Common molecular markers like AFLP [3] and, RAPD and ISSR [10] [11] have been used to assess the genetic diversity of *J. curcas*. The assessment of genetic diversity using molecular markers disclosed low interaccessional variability in local *J. curcas* germplasm [12]. Basha and Sujatha [11] used RAPD, ISSR and SSR markers to study the diversity between *J. curcas* accessions from different countries, which revealed low genetic variability between accessions from same country and maximum divergence between Indian accessions and a non-toxic Mexican accession. They also developed SCAR markers to differentiate Indian accessions from non-toxic Mexican accession.

There are less popular but extremely useful markers like SSRs (Simple Sequence Repeats) and SNPs (Single Nucleotide Polymorphisms) [13] which can be used for genetic diversity profiling. Of these markers, SSRs [14], also known as Microsatellites or Tandem repeats are short repeating nucleotide sequences in DNA that provide greater confidence for the assessment of genetic diversity and relationship [15]. These are the markers of choice for plant genetics and breeding applications [16] [17] as the data generated by these markers can be used for selections during backcross breeding programs [15], and also because of their reproducibility, multiallelic nature, codominant inheritance, relative abundance and good genome coverage [17]. Marker Assisted Selection (MAS) has proved to be the best resource for improvement of many crops [18]. SSRs have been used for MAS in crops like rice [19] and common bean [20].

The traditional methods of developing SSR markers are usually time consuming and labor-intensive [21] [22]. In contrast to this approach, *in silico* mining of SSRs from available ESTs in public databases, with an increasing data accumulating at a fast rate, is an expeditious and cost effective alternative [21]. The search of SSRs in ESTs (representing genes or coding region) becomes more attractive in wake of report of abundance of SSRs in single or low-copy rather than in repetitive or non-coding sequences as assumed earlier [23]. Therefore, molecular SSRs can be searched in EST databases and employed for designing locus-specific primers [24]. Such markers are termed as EST-SSRs. By convention, the EST sequences containing SSRs are generally referred to as SSR-ESTs, whereas the markers developed from SSR-ESTs are called EST-SSRs [17] [25], the same has been followed throughout this paper.

Expressed Sequence Tags (ESTs) are generated by end sequencing of large number of randomly picked clones from cDNA library constructed using mRNA isolated from specific tissue or specific developmental stage of an organism. EST-derived SSR markers are generally less polymorphic than genomic SSRs [26] due to an associated lower polymorphism of coding regions in contrast to non-coding ones [27]. There are also reports of moderate [28] to very high polymorphism associated with EST-SSRs [29] [30]. In spite of contrasting reports about the level of polymorphism related to EST-SSRs, there are several advantages of using expressed sequences compared with genomic sequences as genetic markers. As the EST derived markers represent the functional component of the genome and are transferable across species [31], they can serve as efficient tool for gene discovery and genetic mapping of genes [32] [33]. Therefore, EST-SSRs enhance the role of genetic markers by assaying variations in transcribed and known function of genes [21] [26] [34]. In spite of several studies, till date no genetic map of *Jatropha* has been reported [22] and there is a very recent report of SNP-based linkage map by Wang *et al.* [35]. There is also a need to develop molecular markers for MAS for high oil yielding variety and assessing the genetic diversity.

The present report describes the *in silico* mining of the microsatellites (SSRs) using the *J. curcas* ESTs from various tissues viz., embryo, root, leaf and seed available in the public domain of NCBI. At the time of mining, a total of 13513 ESTs were available and downloaded. From these ESTs, 7552 unigenes were obtained, and 395 EST-SSRs were generated from 377 SSR-ESTs. The EST-SSRs obtained through computational method in this study can be used as potential microsatellite markers for various studies like diversity analysis, MAS etc. Since, the *Jatropha* genes carrying SSRs have been identified in this study, thus, EST-SSRs directly linked to genes will be useful for developing trait linked markers.

## 2. Materials & Method

### Search for EST-SSRs and Primer Designing

EST sequences of *J.curcas* were downloaded from NCBI's dbEST database (http://ncbi.nlm.nih.gov/) [36] which contains sequences generated from different tissue specific cDNA libraries of embryo, root, leaf and seed. These sequences were arranged in a single FASTA file, which was used for the sequence analysis using different softwares and Analysis Tools.

To find the singletons and to assemble the contigs from the total ESTs, an online tool "EGassembler" (http://egassembler.hgc.jp/) [37] was used. The main parameter provided was 'Overlap Identity cutoff (N > 65): 85'. From the unigenes (singletons+contigs), EST-microsatellites [EST-SSRs] were searched using "SSRlocator version 1" (http://www.ufpel.tche.br/faem/fitotecnia/fitomelhoramento/faleconosco.html) [38].

The SSR search was carried out for repeat motifs (ranging from mono- to hexa-nucleotides). For each repeat motif the parameters were: Mononucleotide repeat-20, Dinucleotide repeat-10, Trinucleotide repeat-07, Tetranucleotide repeat-05, Pentanucleotide repeat-04, Hexanucleotide repeat-04 (the numbers indicating repeat unit *i.e.* minimum number of times the motif was repeated at a stretch); Space between SSRs-100, Space between imperfect SSRs [<=]-05. After obtaining the motifs, the sequence complementarity was taken into consideration and accordingly the complementary motifs like AG and CT or AC and GT or AAC and GTT motifs were grouped into a single class under mono-, di-, tri-, tetra-, penta- or hexa-nucleotides, respectively. After getting SSRs, the primers were designed from the flanking regions using the same software as for SSR search. The parameters provided in the software for primer designing are given in **Table 1**.

EST Sequences, which have credit in the primer designing, were searched for their gene annotations using BLASTX at The Arabidopsis Information Resource (TAIR) (http://www.arabidopsis.org/index.jsp) [39]. This data was used to get the Gene Ontology (GO) Annotations and functional categorization of ESTs using locus identifiers at Bulk Data Retrieval System of TAIR (http://www.arabidopsis.org/tools/bulk/go/index.jsp) [40].

## 3. Results and Discussions

### 3.1. Assembling of ESTs as Unigenes

The size of the available EST data used in this study has been calculated in accordance with the size of the

**Table 1.** Parameters for primer designing.

| Sr. No. | Criteria | Minimum | Maximum | Optimum |
|---|---|---|---|---|
| 1 | Amplicon size | 150 | 1000 | - |
| 2 | GC Clamps | 0 | - | - |
| 3 | Primer Size | 18 | 22 | 20 |
| 4 | Tm | 55 | 61 | 59 |
| 5 | Content G/C | 45 | 50 | - |
| 6 | Region scanned | Auto | Auto | - |
| 7 | End Stability | 250 | - | - |

genome of *J. curcas* (C = 416 Mb) reported by Carvalho and coworkers [41]. The ESTs of *J. curcas* generated from tissue specific cDNA library of various tissues (viz. embryo, root, leaf and seed) available in the NCBI's public database dbEST, were downloaded and pooled. These downloaded ESTs were inclusive of the seed specific ESTs generated in our laboratory. The pooled set consisted of 13513 ESTs (~6.2 MB) in all, which comprised of 9844 ESTs of embryo, 1000 of leaf, 1304 of root and 1375 of seed library. Using the EGassembler, all the sequences were categorized into singletons and contigs. The EGassembler segregated 13513 ESTs into 6098 singletons and 7415 redundant sequences. Then it assembled the redundant sequences into 1454 contigs. Therefore, through the software, the total ESTs were categorized into contigs and singletons, which were together grouped as 7552 (~3.8 MB) Unigenes. These data showed that the 45% of the total ESTs, downloaded from the database, were singletons and the rest 55% were assembled into contigs (**Figure 1**). The assembling of the redundant ESTs into contigs was beneficial in reducing the errors in sequence analysis in addition to removing the redundancy so that only the unigenes were used for SSR mining and for annotation. As reported by Raji and coworkers [18], these unigenes, when used for the mining of SSRs result in a realistic estimate of the microsatellite repeat frequency and ensures that non redundant EST-SSR markers that correspond to unique loci in the genome are obtained. Therefore, in this study the unigenes were used for SSR search. The mining of the EST-SSRs starting with downloading of all the *Jatropha* ESTs is outlined in **Figure 1**.

## 3.2. Occurrence and Frequency of Microsatellites

For searching the SSRs, the repeat motifs in the software, were selected from mono- to hexa-nucleotide as

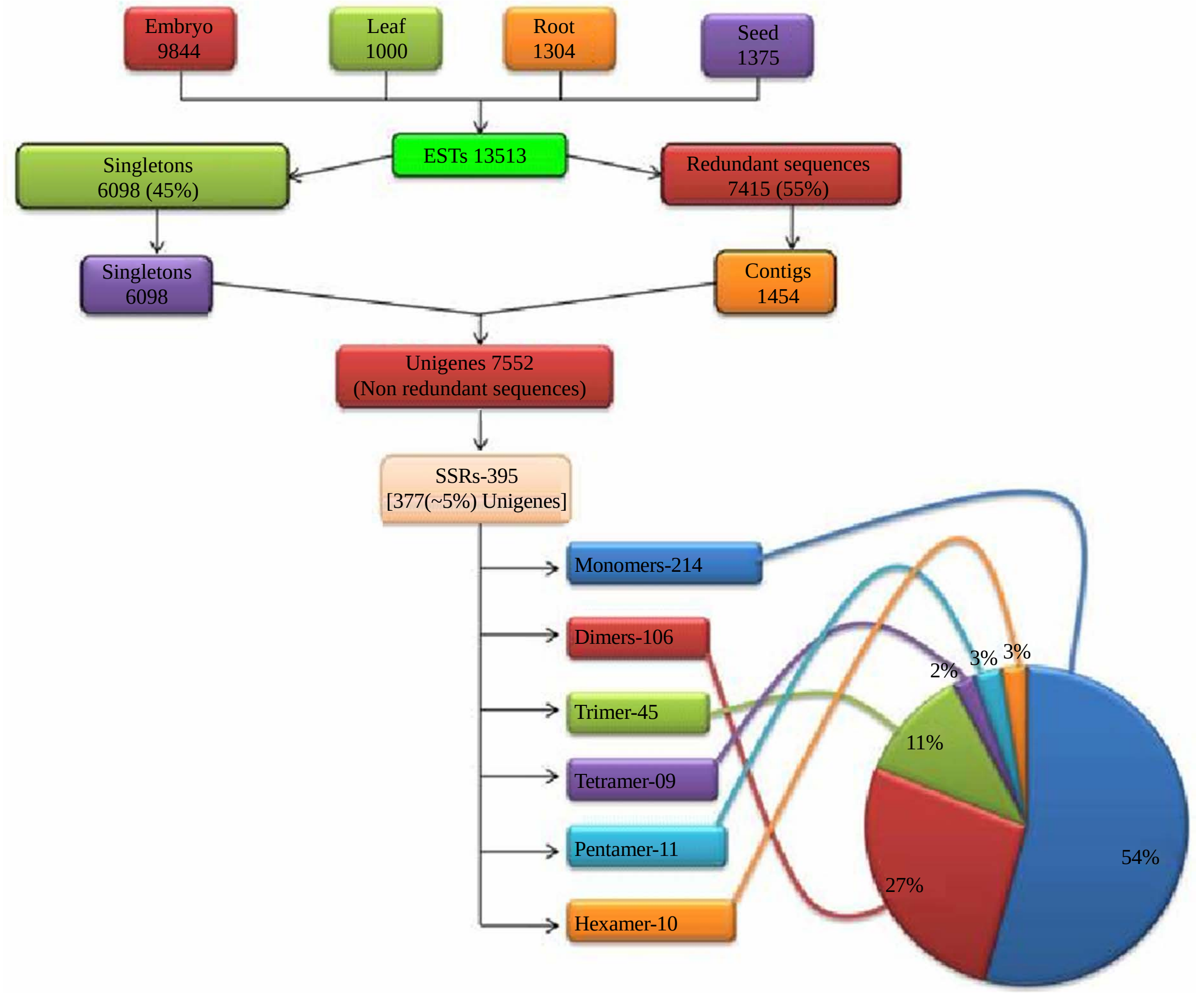

**Figure 1.** Overview of the study indicating the major steps and the statistics leading to generation of the EST-SSRs.

going above this motif range, the frequency of occurrence of SSRs is drastically reduced. Thus, the SSRs were obtained in the form of repeat motifs ranging from mono- to hexa-nucleotides. Out of the 7552 unigenes searched for SSRs, 395 SSRs (**Table 2**) were generated from 377 unigenes. These 395 SSRs can be termed as EST-SSRs and 377 unigenes possessing SSRs can be termed as SSR-ESTs according to the convention. The 377 SSR-ESTs amounted to approximately 5% (inclusive of the mononucleotide repeat motif) of total unigenes and 2.78% of total downloaded EST data set. The various studies show a representation ranging from 2.65% - 16.82% [25] to 26.84% [42] in dicot species and 7% - 10% [43] in cereals or monocots. The workers [17] [21] [25] who have carried out similar studies are of the view that the variation in the percentage may be due to variation in sample size, search criteria, size of database, and the tools used for EST-SSR development. The percentage of SSR-ESTs in the present study could be owing to more stringent preset parameters for EST mining compared to other similar studies [21] [42] that reported a higher percentage of SSR-ESTs.

The 395 SSRs were present in 377 SSR-ESTs as 17 (4%) SSR-ESTs contained more than one SSR e.g. FM889616.1 with 3 SSRs, having motifs $(GA)_{32}$, $(AG)_{14}$, $(AG)_{14}$ (data not shown). The SSRs in mononucleotide class were found to be the most abundant with a frequency of 1/17.83 kb followed by dinucleotide 1/36.00 kb, trinucleotide 1/84.82 kb, tetranucleotide 1/424.11 kb, pentanucleotide 1/347.00 kb and hexanucleotide 1/381.70 kb.

## 3.3. Distribution of Microsatellite Classes and Motifs

The overall analysis of the distribution of the microsatellites into various classes of the repeat types (mono-, di-, tri-, tetra- penta- and hexa-nucleotides) showed that the number of the microsatellites decreased with increasing motif size (**Figure 2**, **Table 3**). It was observed that mononucleotide repeats were the most abundant (representing 54% of the total microsatellites), followed by dinucleotide (27%) and trinucleotide (11%). The least frequent were tetra-, penta- and hexa-nucleotides (2% - 3%). The abundance of mononucleotides is in accordance with several previous reports [23] [25] [44] and also that these contributed to nearly half of all the SSRs, is similar to those in certain species of dicots analysed previously [25]. The dinucleotides were the second most abundant class as reported across most of the dicots investigated by Kumpatla and Mukhopadhyay [25], suggesting an over-representation of UTRs (un-translated regions) compared with ORFs (Open Reading Frames).

The non-dominance of trinucleotides compared to other classes, by virtue of which the decreasing trend of various classes with increasing motif size, is in contrast to several earlier studies but in concurrence to that reported for several dicots [25]. These observations about the abundance and therefore, the dominance of one SSR motif category over other categories, holds significance in the chances of fixation of mutations against selection pressure [45]. The trinucleotides have more chances of getting fixed against mutation pressure due to selection against frameshift events [45] The prevalence of di- over tri-nucleotide in this study could be attributed to 1. increased stringency of preset parameters in this study compared to previous studies [21] [22], so as not to compromise on polymorphism level and thus their utility as markers. The results were also computed with relaxed preset parameter of repeat length which gave a higher percentage of total SSRs especially trinucleotides (data not shown). But, the results reported here are those obtained with more stringent parameter of minimum repeat length 2. a bias in representation of 5' and 3'UTRs in the EST dataset used for mining. A lowered representation of tetranucleotides, as also observed in this study, is also suggestive of under representation of 3'UTRs [25].

In terms of SSR coverage of available Unigenes data (~3.8 MB), it was observed that a total of 11.7 kb (0.31%) region was covered by SSR motifs. Out of this, mono-represented 6.5 kb (0.17%) region, di—3.3 kb (0.08%), tri—1.1 kb (0.03%), tetra—0.18 kb (0.004%), penta—0.26 kb (0.006%) and hexa-nucleotides 0.24 kb (0.006%).

The various classes of repeat motifs, when analyzed further, showed that some motifs in each category were more abundant than others (**Table 4**), e.g. among the dinucleotide repeats, the AG/CT motif was the most common (33%) followed by the motifs GA/TC (31%) and, the least common was AC/GT (0.94%). The abundance of AG/CT/GA/TC motifs are in concurrence with previous studies [25] [43] [44] where ESTs were used for mining SSRs, in contrast to abundance of AT motif when genomic data was used for mining SSRs [44] [46]. Thus, abundance of the motifs is attributed to systematic bias resulting from the use of ESTs (coding sequences) instead of genomic sequences (non-coding) as a source for SSR mining [43]. The CG motif was found to be totally absent, which is in concurrence to earlier studies, where it has been observed to be either the least [43] or absent [44]. Among the trinucleotide repeats, the most common motif is AGA/TCT subclass amounting to

**Table 2.** Categorization of SSRs by repeat units and repeat motif.

| Repeat Type | Repeat Motif | Number of Repeat Units | | | | | | | | | | | | | Total | Analysis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | >20 | | |
| Mononucleotide | A/T | - | - | - | - | - | - | - | - | - | - | - | - | 211 | **211** | **214 (54%)** |
| | G/C | - | - | - | - | - | - | - | - | - | - | - | - | 3 | **3** | |
| | | | | | | | | | | | | | | **214** | | |
| Dinucleotide | AG/CT | - | - | - | - | - | - | 6 | 4 | 4 | 1 | 4 | 2 | 14 | **35** | **106 (27%)** |
| | AT/AT | - | - | - | - | - | - | - | 1 | 2 | 2 | 1 | 1 | 13 | **20** | |
| | AC/GT | - | - | - | - | - | - | - | - | - | - | - | 1 | - | **1** | |
| | TA/TA | - | - | - | - | - | - | 3 | 3 | 1 | - | 1 | 2 | 7 | **17** | |
| | GA/TC | - | - | - | - | - | - | 4 | 7 | 5 | 4 | - | 3 | 10 | **33** | |
| | | | | | | | | **13** | **15** | **12** | **7** | **6** | **9** | **44** | | |
| Trinucleotide | AAC/GTT | - | - | - | 1 | - | - | - | - | 1 | - | - | - | - | **2** | **45 (11.5%)** |
| | AAT/ATT | - | - | - | 1 | 1 | - | - | - | 1 | - | - | - | - | **3** | |
| | ACC/GGT | - | - | - | 1 | - | - | - | - | - | - | - | - | - | **1** | |
| | AGA/TCT | - | - | - | 4 | - | 3 | 1 | 3 | - | 1 | - | - | - | **12** | |
| | AGC/GCT | - | - | - | 2 | - | - | - | - | - | - | - | - | - | **2** | |
| | ATA/TAT | - | - | - | 1 | 2 | - | 1 | 1 | - | - | - | - | - | **5** | |
| | ATG/CAT | - | - | - | - | - | 1 | - | - | - | - | - | - | - | **1** | |
| | CAC/GTG | - | - | - | 1 | - | - | - | - | - | - | - | - | - | **1** | |
| | CAG/CTG | - | - | - | 1 | - | 1 | - | 1 | - | - | - | - | - | **3** | |
| | CTT/AAG | - | - | - | 1 | - | - | - | - | - | - | - | - | - | **1** | |
| | GAA/TTC | - | - | - | 2 | 1 | 1 | - | - | - | - | - | - | - | **4** | |
| | GCA/TGC | - | - | - | 1 | 1 | - | - | - | - | - | - | - | - | **2** | |
| | GGA/TCC | - | - | - | 1 | - | - | - | - | - | - | - | - | - | **1** | |
| | TAA/TTA | - | - | - | 1 | - | 2 | 1 | 2 | - | - | - | - | - | **6** | |
| | TTG/CAA | - | - | - | 1 | - | - | - | - | - | - | - | - | - | **1** | |
| | | | | | **19** | **5** | **8** | **3** | **7** | **2** | **1** | | | | | |
| Tetranucleotide | AAGA/TCTT | - | 2 | - | - | - | - | - | - | - | - | - | - | - | **2** | **9 (2%)** |
| | AATT/AATT | - | 1 | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | CATA/TATG | - | 1 | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | TATT/AATA | - | 1 | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | TTAA/TTAA | - | 1 | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | TTAT/ATAA | - | - | 1 | - | - | - | - | - | - | - | - | - | - | **1** | |
| | TTCT/AGAA | - | 1 | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | TTTA/TAAA | - | 1 | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | | | **8** | **1** | | | | | | | | | | | | |
| Pentanucleotide | AAGAA/TTCTT | - | 1 | - | 1 | - | - | - | - | - | - | - | - | - | **2** | **11 (3%)** |
| | AGGAA/TTCCT | 1 | - | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | ATTTT/AAAAT | | 1 | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | CTTCT/AGAAG | 1 | - | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | TAAAA/TTTTA | 1 | - | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | TATTT/AAATA | 1 | - | 1 | - | - | - | - | - | - | - | - | - | - | **2** | |
| | TCTTT/AAAGA | - | 1 | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | TTATA/TATAA | 1 | - | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | TTTCT/AGAAA | - | 1 | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | | **5** | **4** | **1** | **1** | | | | | | | | | | | |
| Hexanucleotide | AAAAAG/CTTTTT | 1 | - | - | - | - | - | - | - | - | - | - | - | - | **1** | **10 (2.5%)** |
| | CAGCTC/GAGCTG | 1 | - | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | GCTGGT/ACCAGC | 1 | - | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | GGATCA/TGATCC | 1 | - | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | GTTTCA/TGAAAC | 1 | - | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | TTCCAT/ATGGAA | 1 | - | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | TTTATT/AATAAA | 1 | - | - | - | - | - | - | - | - | - | - | - | - | **1** | |
| | TTTCTC/GAGAAA | 3 | - | - | - | - | - | - | - | - | - | - | - | - | **3** | |
| | | **10** | | | | | | | | | | | | | | |
| | | | **Total** | | | | | | | | | | | | | **395** |

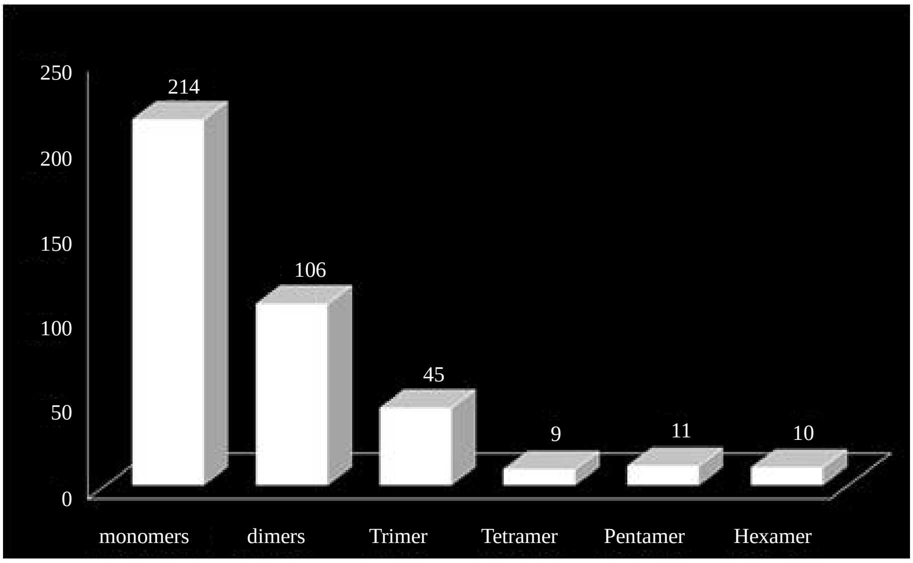

**Figure 2.** Distribution of SSRs into various classes.

**Table 3.** Abundance of SSRs of various types.

| Repeat Types | No. of SSRs | Abundance (%) |
|---|---|---|
| Mononucleotide | 214 | 54 |
| Dinucleotide | 106 | 27 |
| Trinucleotide | 45 | 11 |
| Tetranucleotide | 9 | 2 |
| Pentanucleotide | 11 | 3 |
| Hexanucleotide | 10 | 3 |

**Table 4.** Most abundant motifs and their relative abundance in each of the SSR types.

| Repeat Types | Most Abundant Motifs | Relative Abundance (%) |
|---|---|---|
| Mononucleotide | A/T | 98 |
| Dinucleotide | AG/CT | 33 |
| Trinucleotide | AGA/TCT | 27 |
| Tetranucleotide | AAGA/TCTT | 22 |
| Pentanucleotide | AAGAA/TTCTT<br>TATTT/AAATA | 18 each |
| Hexanucleotide | TTTCTC/GAGAAA | 30 |

26.6% and rest of them ranging from 2% - 13% of the total microsatellites in this class. The CCG/CGG motif is reported to be the rarest motif in dicots [23] [25] and was observed to be absent in this study. In the tetranucleotide repeats, most of the motifs were AT rich. The most common motif was AAGA/TCTT (22%) and the rest of them were each ~11% of the total microsatellites in this class. In the pentanucleotide class of motifs the most common one was AAGAA/TTCTT and TATTT/AAATA (18% each) and others were each 9%. The hexanucleotide class TTTCTC/GAGAAA (30%) formed the most abundant subclass and the rest of them were 10% each. In general, the motifs were observed to be AT rich and less of GC rich motifs, similar to that observed for dicots

[25].

The analysis of repeat units under each motif class revealed a varying range of repeat units in each of the classes of repeat motifs. It was observed that, in dinucleotide motif, repeat units ranged from 10 - 45; in trinucleotide motif, from 7 - 13; in tetranucleotide, from 5 - 6 units; in pentanucleotide, from 4 - 6; and hexanucleotide motif was represented by a single class of 6 repeat units only. Further analysis of the number of repeat units in every class of the SSRs, especially tri-, tetra- penta- and hexa-nucleotides, showed that the number of the microsatellites decreased with increasing repeat unit length with little variation, e.g. for trinucleotide motif, SSRs with 7 repeats were represented by 42.2% while 2.2% by 13 repeat units. Amongst the pentanucleotide SSRs, the category with 4 repeat units shared as much as 45.5% of the total class in comparison to 9% for repeat unit of seven (**Figure 3**). Therefore, it can be said that as the class of the SSR motif size increases, like tetra-, penta- and hexa-nucleotide, higher number, rather 100% of microsatellites were found in the category of <10 repeat units (**Table 2**) which is similar to that observed by Varshney and co-workers [43]. These results clearly indicate the effect of increased stringency of parameters which were maintained during this study to retain the polymorphism level and utility of the SSRs as markers because the probability of polymorphism increases with increasing length of SSRs [47]-[49] and, a higher number of repeat followed by shorter stretches would be beneficial for marker development [48]. The polymorphism reported in *Jatropha* in earlier studies was very low, therefore, the parameters for mining the SSRs were kept more stringent in this study, which lead to lower frequency of SSRs but with a longer repeat length; as in the case of trinucleotide repeats, keeping the minimum repeat length of 7 resulted in it not being the most abundant class, as reported in other similar studies.

## 3.4. Designing of Primers towards Marker Development

For the use of SSRs as markers, it is necessary to design the primers. The SSRs commonly used for marker development are those belonging to di-, tri- and tetra-nucleotides [25]. The mononucleotides are useful for population genetic analyses of chloroplast genomes [50] and can also be useful in filling gaps in linkage maps created by di-, tri-, and tetra-nucleotide repeats [25] but, at the same time they cause difficulties in accurate sizing of polymorphisms [18]. Therefore, to design the primers for potential SSR markers, the mononucleotide repeats were not included. Thus, out of 395 EST-SSRs generated from 377 SSR-ESTs, the primers were designed for only 181 SSRs.

For each of the SSRs, a pair of reverse and forward primer was designed from the flanking regions of their respective SSR-ESTs by the software. 181 SSRs generated from 172 SSR-ESTs were used for primer designing and yielded 79 SSR mediated primer pairs (data not shown). These 79 primer pairs were designed from 76 SSR-ESTs as some of these contained more than one SSR e.g. JES 56 and 57 (**Supplementary Table A**). Thus, 76 SSR-ESTs having credit in primer designing have been termed as ESTs-PD and were further annotated. The primers could not be designed for some of the EST-SSRs from their respective SSR-ESTs. As reported by Varshney and coworkers [42], this could be due to any or all of the following reasons, (a) SSR-ESTs are too short, (b) EST-SSRs are too close to the cloning site of the SSR-ESTs, or (c) the flanking sequences are not unique, as was also observed for some of the SSRs in this study.

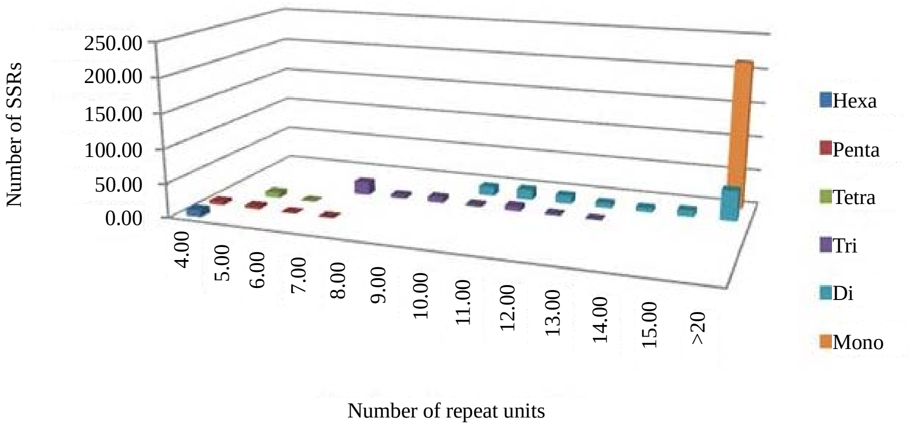

**Figure 3.** Distribution of SSRs as per repeat unit size in different types.

### 3.5. Functional Annotation of the ESTs-PD

The GC level of the genome of *J. curcas* is typical of core dicots, therefore, it should be easy to annotate by sequence comparison with *Arabidopsis* [41], hence, ESTs-PD were searched for their gene annotations using BLASTX at TAIR. The Gene Ontology (GO) Annotations and functional categorization of ESTs-PD obtained using locus identifiers are given in **Supplementary Table A**.

The data showed that most of the ESTs-PD are expressing functional proteins and still there are some for which the protein is not yet predicted. On the basis of the functions related to the predicted protein, the ESTs-PD were classified into three major classes viz. Cellular Component, Biological Process and Molecular Function (**Figure 4**). In the limits of the available data in the public database for the ESTs of *J. curcas*, it was found that one of the ESTs-PD (Contig1345) containing SSR (JES35) expresses gene of oil biosynthesis pathway (AT1G48750).

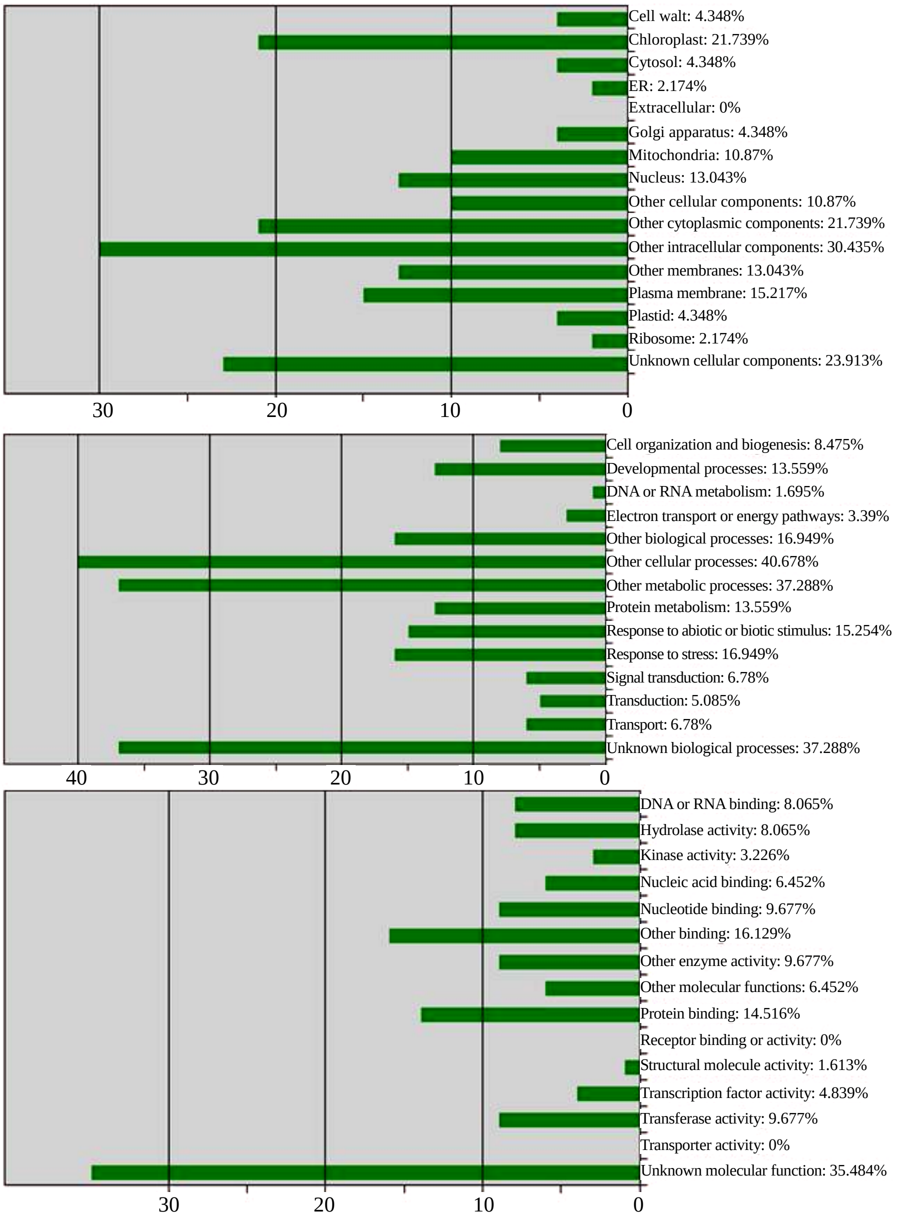

**Figure 4.** Functional categorization of ESTs-PD by loci A: Cellular component, B: Biological process, C: Molecular function.

## 4. Conclusion

The *in silico* mining of EST-SSRs of *Jatropha* was carried out in this study taking advantage of the availability of enormous EST data in the public database, the importance of ESTs in SSR mining and, the potential of modern bioinformatics tools combined with their speed and ease. The stringency of the preset parameters was kept high so as not to compromise on the level of polymorphism in potential EST-SSRs, thus, their utility as markers, more so in this study, as low levels of polymorphisms have been reported in *Jatropha*. The functional annotation of the SSR-ESTs showed that most of them are associated with expressed proteins and therefore, trait linked genes. Thus, in this study, the genes of *Jatropha* carrying SSRs were identified. The EST-SSRs generated would be useful for developing trait linked markers. As the expressed sequences are highly conserved, the SSRs developed from the ESTs are characterized by transferability across species. Owing to this characteristic, these SSRs could also be useful as markers across closely related species like *Ricinus*, thus, saving time and resources in reiteration of SSR mining or; for related species with limited or no sequence information. EST-SSRs like JES35 generated from EST expressing gene of fatty acid biosynthesis pathway (AT1G48750) would be of utmost importance towards marker development in *Jatropha*. With more data being submitted at a rapid pace to the public database, more such SSRs can be looked for in comparative genomic studies and, the knowledge generated in this study is a step towards development of markers in this plant and also related species.

## Acknowledgements

The authors are thankful to the Department of Biotechnology, Government of India, for financial support for research and fellowships of two authors Dr. Poonam Bhargava and Mr. Ganesh B. Patil and, Puri foundation for Education in India, for providing infrastructure facilities.

## References

[1] Ginwal, H.S., Rawat, P.S. and Srivastava, R.L. (2004) Seed Source Variation in Growth Performance and Oil Yield of *Jatropha curcas* Linn. in Central India. *Silvae Genetics*, **53**, 186-192.

[2] Ikbal, K., Boora, S. and Dhillon, R.S. (2010) Evaluation of Genetic Diversity in *Jatropha curcas* L. Using RAPD Markers. *Indian Journal of Biotechnology*, **9**, 50-57.

[3] Tatikonda, L., Wani, S.P., Kannan, S., Beerelli, N., Sreedevi, T.K., Hoisington, D.A., Devi P. and Varshney, R.K. (2009) AFLP-Based Molecular Characterization of an Elite Germplasm Collection of *Jatropha curcas* L., a Biofuel Plant. *Plant Science*, **176**, 505-513. http://dx.doi.org/10.1016/j.plantsci.2009.01.006

[4] Jubera, M.A., Janagoudar, B.S., Biradar, D.P., Ravikumar, R.L., Koti, R.V. and Patil, S.J. (2009) Genetic Diversity Analysis of Elite *Jatropha curcas* (L.) Genotypes Using Randomly Amplified Polymorphic DNA Markers. Karnataka. *Journal of Agricultural Sciences*, **22**, 293-295.

[5] Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A. and Tingey, S.V. (1990) DNA Polymorphisms Amplified by Arbitrary Primers Are Useful as Genetic Markers. *Nucleic Acids Research*, **18**, 6531-6535. http://dx.doi.org/10.1093/nar/18.22.6531

[6] Alhani, M.C. and Wilkinson, M.J. (1998) Inter Simple Sequence Repeat Polymerase Chain Reaction for the Detection of Somaclonal Variation. *Plant Breeding*, **117**, 573-575. http://dx.doi.org/10.1111/j.1439-0523.1998.tb02210.x

[7] Blair, M.W., Panaud, O. and McCouch, S.R. (1999) Inter-Simple Sequence Repeat (ISSR) Amplification for Analysis of Microsatellite Motif Frequency and Fingerprinting in Rice (*Oryza sativa* L.). *Theoretical and Applied Genetics*, **98**, 780-792. http://dx.doi.org/10.1007/s001220051135

[8] Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Zabeau, M. and Kuiper, M. (1995) AFLP: A New Technique for DNA Fingerprinting. *Nucleic Acids Research*, **23**, 4407-4414. http://dx.doi.org/10.1093/nar/23.21.4407

[9] Zhu, J., Gale, M.D., Quarrie, S., Jackson, M.T. and Bryan, G.J. (1998) AFLP Markers for the Study of Rice Biodiversity. *Theoretical and Applied Genetics*, **96**, 602-611. http://dx.doi.org/10.1007/s001220050778

[10] Basha, S.D., Francis, G., Makkar, H.P.S., Becker, K. and Sujatha, M. (2009) A Comparative Study of Biochemical Traits and Molecular Markers for Assessment of Genetic Relationships between *Jatropha curcas* L. Germplasm from Different Countries. *Plant Science*, **176**, 812-823. http://dx.doi.org/10.1016/j.plantsci.2009.03.008

[11] Basha, S.D. and Sujatha, M. (2007) Inter and Intra-Population Variability of *Jatropha curcas* L. Characterized by RAPD and ISSR Markers and Development of Population-Specific SCAR Markers. *Euphytica*, **156**, 375-386. http://dx.doi.org/10.1007/s10681-007-9387-5

[12] Sujatha, M., Reddy, T.P. and Mahasi, M.J. (2008) Role of Biotechnological Interventions in the Improvement of Castor (*Ricinus communis* L.) and *Jatropha curcas* L. *Biotechnology Advances*, **26**, 424-435. http://dx.doi.org/10.1016/j.biotechadv.2008.05.004

[13] Vieux, E.F., Kwok, P.Y. and Miller, R.D. (2002) Primer Design for PCR and Sequencing in High-Throughput Analysis of SNPs. *BioTechniques*, **32**, S28-S32.

[14] Levinson, G. and Gutman, G.A. (1987) Slipped-Strand Mispairing: A Major Mechanism for DNA Sequence Evolution. *Molecular Biology and Evolution*, **4**, 203-221.

[15] Chakravarthi, B.K. and Naravaneni, R. (2006) SSR Marker Based DNA Fingerprinting and Diversity Study in Rice (*Oryza sativa*. L.). *African Journal of Biotechnology*, **5**, 684-688.

[16] Tripathi, K.P., Roy, S., Khan, F., Shasany, A.K., Sharma, A. and Khanuja, S.P.S. (2008) Identification of SSR-ESTs Corresponding to Alkaloid, Phenylpropanoid and Terpenoid Biosynthesis in MAPs. *Online Journal of Bioinformatics*, **9**, 78-91.

[17] Varshney, R.K., Graner, A. and Sorrells, M.E. (2005) Genic Microsatellite Markers in Plants: Features and Applications. *Trends in Biotechnology*, **23**, 48-55. http://dx.doi.org/10.1016/j.tibtech.2004.11.005

[18] Raji, A.A.J., Anderson, J.V., Kolade, O.A., Ugwu, C.D., Dixon, A.G.O. and Ingelbrecht, I.L. (2009) Gene-Based Microsatellites for Cassava (*Manihot esculenta* Crantz): Prevalence, Polymorphisms, and Cross-Taxa Utility. *BMC Plant Biology*, **9**, 1471-1429. http://www.biomedcentral.com/1471-2229/9/118

[19] Wang, Y.H., Liu, S.J., Ji, S.L., Zhang, W.W., Wang, C.M., Jiang, L. and Wan, J.M. (2005) Fine Mapping and Marker-Assisted Selection (MAS) of a Low Glutelin Content Gene in Rice. *Cell Research*, **15**, 622-630. http://dx.doi.org/10.1038/sj.cr.7290332

[20] Carneiro, F.F., Santos, J.B. and Leite, M.E. (2010) Marker-Assisted Backcrossing Using Microsatellites and Validation of SCAR *Phs* Marker for Resistance to White Mold in Common Bean. *Electronic Journal of Biotechnology*, **13**, Online Version. http://ejb.ucv.cl/content/vol13/issue6/full/13/index.html

[21] Gupta, S. and Prasad, M. (2009) Development and Characterization of Genic SSR Markers in *Medicago truncatula* and Their Transferability in Leguminous and Non-Leguminous Species. *Genome*, **52**, 761-771. http://dx.doi.org/10.1139/G09-051

[22] Wen, M., Wang, H., Xia, Z., Zou, M., Lu, C. and Wang, W. (2010) Development of EST-SSR and Genomic-SSR Markers to Assess Genetic Diversity in *Jatropha curcas* L. *BMC Resarch Notes*, **3**, 42. http://www.biomedcentral.com/1756-0500/3/42

[23] Morgante, M., Hanafey, M. and Powell, W. (2002) Microsatellites Are Preferentially Associated with Nonrepetitive DNA in Plant Genomes. *Nature Genetics*, **30**, 194-200. http://dx.doi.org/10.1038/ng822

[24] Riju, A, Chandrasekar, A. and Arunachalam, V. (2007) Mining for Single Nucleotide Polymorphisms and Insertions / Deletions in Expressed Sequence Tag Libraries of Oil Palm. *Bioinformation*, **2**, 128-131. http://dx.doi.org/10.6026/97320630002128

[25] Kumpatla, S.P. and Mukhopadhyay, S. (2005) Mining and Survey of Simple Sequence Repeats in Expressed Sequence Tags of Dicotyledonous Species. *Genome*, **48**, 985-998. http://dx.doi.org/10.1139/g05-060

[26] Garcia, R.A.V., Rangel, P.N., Brondani, C., Martins, W.S., Melo, L.C., Carneiro, M.S., Borba, T.C. and Brondani, R.P. (2011) The Characterization of a New Set of EST-Derived Simple Sequence Repeat (SSR) Markers as a Resource for the Genetic Analysis of *Phaseolus vulgaris*. *BMC Genetic*, **12**, 41-54. http://dx.doi.org/10.1186/1471-2156-12-41 http://www.biomedcentral.com/1471-2156/12/41

[27] Ceresini, P.C., Silva, C.L.S.P., Missio, R.F., Souza, E.C., Fischer, C.N., Guillherme, I.R., Gregorio, I., Da Silva, E.H.T., Cicarelli, R.M.B., Silva, M.T.A., Garcia, J.F., Avelar, G.A., Porto Neto, L.R., Marcon, A.R., Bacci Jr., M. and Marini, D.C. (2005) Satellyptus: Analysis and Database of Microsatellites from ESTs of *Eucalyptus*. *Genetics and Molecular Biology*, **28**, 589-600.

[28] Pinto, L.R., Oliveira, K.M., Ulian, E.C., Garcia, A.A.F. and De Souza, A.P. (2004) Survey in the Sugarcane Expressed Sequence Tag Database (SUCEST) for Simple Sequence Repeats. *Genome*, **47**, 795-804. http://dx.doi.org/10.1139/g04-055

[29] Eujayl, I., Sledge, M.K., Wang, L., May, G.D., Chekhovskiy, K., Zwonitzer, J.C. and Mian, M.A.R. (2004) *Medicago truncatula* EST-SSRs Reveal Cross-Species Genetic Markers for *Medicago* spp. *Theoretical and Applied Genetics*, **108**, 414-422. http://dx.doi.org/10.1007/s00122-003-1450-6

[30] Fraser, L.G., Harvey, C.F., Crowhurst, R.N. and Silva, H.N. (2003) EST-Derived Microsatellites from *Actinidia* Species and Their Potential for Mapping. *Theoretical and Applied Genetics*, **108**, 1010-1016. http://dx.doi.org/10.1007/s00122-003-1517-4

[31] Ellis, J.R. and Burke, J.M. (2007) EST-SSRs as a Resource for Population Genetic Analyses. *Heredity*, **99**, 125-132. http://dx.doi.org/10.1038/sj.hdy.6801001

[32] Sato, K., Nankaku, N. and Takeda, K. (2009) A high-Density Transcript Linkage Map of Barley Derived from a Single Population. *Heredity*, **103**, 110-117. http://dx.doi.org/10.1038/hdy.2009.57

[33] Shirasawa, K., Oyama, M., Hirakawa, H., Sato, S., Tabata, S., Fujioka, T., *et al*. (2011) An EST-SSR Linkage Map of *Raphanus sativus* and Comparative Genomics of the Brassicaceae. *DNA Research*, **18**, 221-232. http://dx.doi.org/10.1093/dnares/dsr013

[34] Varshney, R.K., Mahendar, T., Aggarwal, R.K. and Börner, A. (2007) Genetic Molecular Markers in Plants: Development and Applications. In: Varshney, R.K. and Tuberosa, R., Eds., *Genomics-Assisted Crop Improvement*: 1: *Genomics Approaches and Platforms*, Springer, Berlin, 13-29.

[35] Wang, C.M., Liu, P., Yi, C., Gu, K., Sun, F., Li, L., Lo, L.C., Liu, X., Feng, F., Lin, G., Cao, S., Hong, Y., Yi, Z. and Yue, G.H. (2011) A First Generation Microsatellite and SNP-based Linkage Map of *Jatropha*. *PLoS ONE*, **6**, e23632. http://dx.doi.org/10.1371/journal.pone.0023632

[36] NCBI's dbEST Database. http://www.ncbi.nlm.nih.gov/

[37] EGassembler. http://egassembler.hgc.jp/

[38] SSRlocator Version 1. [Standalone software] http://www.ufpel.tche.br/faem/fitotecnia/fitomelhoramento/faleconosco.html

[39] TAIR. [online] http://www.arabidopsis.org/index.jsp

[40] Locus Identifiers at Bulk Data Retrieval System of TAIR. http://www.arabidopsis.org/tools/bulk/go/index.jsp

[41] Carvalho, C.R., Clarindo, W.R., Praca, M.M., Araújo, F.S. and Carels, N. (2008) Genome Size, Base Composition and Karyotype of *Jatropha curcas* L., an Important Biofuel Plant. *Plant Science*, **174**, 613-617.

[42] Raju, N.L., Gnanesh, B.N., Lekha, P., Jayashree, B., Pande, S., Hiremath, P.J., Byregowda, M., Singh, N.K. and Varshney, R.K. (2010) The First Set of EST Resource for Gene Discovery and Marker Development in Pigeon Pea (*Cajanus cajan* L.). *BMC Plant Biology*, **10**, 45. http://www.biomedcentral.com/1471-2229/10/45

[43] Varshney, R.K., Thiel, T., Stein, N., Langridge, P. and Graner, A. (2002) *In Silico* Analysis on Frequency and Distribution of Microsatellites in ESTs of Some Cereal Species. *Cell and Molecular Biology Letters*, **7**, 537-546.

[44] Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D. and Waug, R. (2000) Computational and Experimental Characterization of Physically Clustered Simple Sequence Repeats in Plants. *Genetics*, **156**, 847-854.

[45] Metzgar, D., Bytof, J. and Wills, C. (2000) Selection against Frameshift Mutations Limits Microsatellite Expansion in Coding DNA. *Genome Research*, **10**, 72-80.

[46] Morgante, M. and Olivieri, A.M. (1993) PCR-Amplified Microsatellites as Markers in Plant Genetics. *The Plant Journal*, **3**, 175-182.

[47] Cho, Y.G., Ishii, T., Temnykh, S., Chen, X., Lipovich, L., McCouch, S.R., Park, W.D., Ayres, N. and Cartinhour, S. (2000) Diversity of Microsatellites Derived from Genomic Libraries and GenBank Sequences in Rice (*Oryza sativa* L.). *Theoretical Applied Genetics*, 100, 713-722. http://dx.doi.org/10.1007/s001220051343

[48] La Rota, M., Kantety, R.V., Yu, J.K. and Sorrells, M.E. (2005) Nonrandom Distribution and Frequencies of Genomic and EST-Derived Microsatellite Markers in Rice, Wheat, and Barley. *BMC Genomics*, **6**, 23. http://www.biomedcentral.com/1471-2164/6/23

[49] Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S. and McCouch, S. (2001) Computational and Experimental Analysis of Microsatellites in Rice (*Oryza sativa* L.): Frequency, Length Variation, Transposon Associations, and Genetic Marker Potential. *Genome Research*, **11**, 1441-1452. http://dx.doi.org/10.1101/gr.184001

[50] Powell, W., Morgante, M., McDevitt, R., Vendramin, G.G. and Rafalski, J.A. (1995) Polymorphic Simple Sequence Repeat Regions in Chloroplast Genomes: Applications to the Population Genetics of Pines. *Proceedings of National Academy of Sciences of the United States of America*, **92**, 7759-7763. http://dx.doi.org/10.1073/pnas.92.17.7759

## Supplementary Table

**Table A.** GO annotations of ESTs-PD[a].

| Sr. No. | Gene | SSR ID | Protein Annotation | GO ID | GO Term | Category |
|---|---|---|---|---|---|---|
| 1 | Contig39 | JES 2 | 1-phosphatidylinosit ol-4-phosphate 5- kinases; zinc ion binding | AT4G33240 | zinc ion binding | func[b] |
| | | | | | pollen development | proc[c] |
| | | | | | endosome | comp[d] |
| | | | | | vacuole organization | proc[c] |
| | | | | | phosphatidylinositol phosphorylation | proc[c] |
| | | | | | 1-phosphatidylinositol-3-phosphate 5-kinase activity | func[b] |
| | | | | | endomembrane system organization | proc[c] |
| 2 | Contig93 | JES 5 | Protein of unknown function | AT1G50630 | --- | proc[c] |
| 3 | Contig332 | JES 12 | heat shock protein 2 | AT4G27670 | response to high light intensity | proc[c] |
| | | | | | response to hydrogen peroxide | proc[c] |
| | | | | | response to heat | proc[c] |
| | | | | | chloroplast | comp[d] |
| 4 | Contig570 | JES 16 | Stress induced protein | AT3G51810 | embryo development ending in seed dormancy | proc[c] |
| | | | | | response to abscisic acid stimulus | proc[c] |
| 5 | Contig635 | JES 19 | unknown protein | AT1G44608 | --- | proc[c] |
| 6 | Contig872 | JES 26 | Papain family cysteine protease | AT4G16190 | proteolysis | proc[c] |
| | | | | | cysteine-type peptidase activity | func[b] (hydrolas e) |
| | | | | | vacuole | comp[d] |
| 7 | Contig896 | JES 27 | Tetratricopeptide repeat (TPR)-like superfamily protein | ATCG00360 | chloroplast | comp[d] |
| | | | | | photosystem I assembly | proc[c] |
| | | | | | unfolded protein binding | func[b] |
| 8 | Contig911 | JES 28 | Thiazole biosynthetic enzyme | AT5G54770 | oxazole or thiazole biosynthetic process | proc[c] |
| | | | | | chloroplast | comp[d] |
| | | | | | thiamine biosynthetic process | proc[c] |
| | | | | | thylakoid | comp[d] |
| | | | | | response to cold | proc[c] |
| | | | | | stromule | comp[d] |
| | | | | | mitochondrion | comp[d] |
| | | | | | protein homodimerization activity | func[b] |
| | | | | | response to DNA damage stimulus | proc[c] |
| | | | | | zinc ion binding | func[b] |

**Continued**

| | | | | | | |
|---|---|---|---|---|---|---|
| 9 | Contig1179 | JES 31 | Cystatin/monellin family protein | AT5G05110 | negative regulation of endopeptidase activity | func[b] |
| | | | | | cysteine-type endopeptidase inhibitor activity | func[b] |
| | | | | | endomembrane system | comp[d] |
| 10 | Contig1259 | JES 32 | RAB GTPase homolog 1C | AT4G17530 | membrane | comp[d] |
| | | | | | GTP binding | func[b] |
| | | | | | plasma membrane | comp[d] |
| | | | | | small GTPase mediated signal transduction | proc[c] |
| | | | | | GTP binding | func[b] |
| | | | | | vacuole | proc[c] |
| | | | | | cytosol | proc[c] |
| | | | | | protein transport | proc[c] |
| 11 | Contig1345 | JES35 | Bifunctional inhibitor/lipid- transfer protein/seed storage 2S albumin superfamily protein | AT1G48750 | endomembrane system | comp[d] |
| | | | | | lipid transport | proc[c] |
| | | | | | lipid binding | func[b] |
| 12 | Contig1364 | JES36 | Thiazole biosynthetic enzyme, chloroplast (ARA6) (THI1) (THI4) | AT5G54770 | oxazole or thiazole biosynthetic process | proc[c] |
| | | | | | chloroplast | comp[d] |
| | | | | | thiamine biosynthetic process | proc[c] |
| | | | | | thylakoid | comp[d] |
| | | | | | response to cold | proc[c] |
| | | | | | stromule | comp[d] |
| | | | | | mitochondrion | comp[d] |
| | | | | | protein homodimerization activity | func[b] |
| | | | | | response to DNA damage stimulus | proc[c] |
| | | | | | zinc ion binding | func[b] |
| 13 | FM895253.1 | JES37 | RNA polymerases | AT4G16265 | DNA-directed RNA polymerase IV complex | comp[d] |
| | | | | | transcription, DNA-dependent | proc[c] |
| | | | | | regulation of transcription, DNA-dependent | proc[c] |
| | | | | | DNA binding | func[b] |
| | | | | | nucleic acid binding | func[b] |
| | | | | | zinc ion binding | func[b] |
| 14 | FM891378.1 | JES40 | unknown protein | AT1G48330 | --- | --- |
| 15 | FM890964.1 | JES43 | FRIGIDA interacting protein | AT2G06005 | biological_process_unkno wn | proc[c] |
| | | | | | protein binding | func[b] |

**Continued**

| | | | | | | |
|---|---|---|---|---|---|---|
| 16 | FM896040.1 | JES45 | dicer-like 2 | AT3G03300 | production of ta-siRNAs involved in RNA interference | proc[c] |
| | | | | | ribonuclease III activity | func[b] |
| | | | | | ATP-dependent helicase activity | func[b] |
| | | | | | maintenance of DNA methylation | proc[c] |
| | | | | | ATP catabolic process | proc[c] |
| | | | | | double-stranded RNA binding | func[b] |
| | | | | | ATP catabolic process | proc[c] |
| | | | | | protein binding | func[b] |
| | | | | | nucleic acid binding | func[b] |
| | | | | | metabolic process | proc[c] |
| | | | | | intracellular | comp[d] |
| | | | | | ATP binding | func[b] |
| | | | | | RNA binding | func[b] |
| | | | | | defense response to virus | proc[c] |
| 17 | FM895395.1 | JES52 | unknown protein | AT1G33860 | --- | --- |
| 18 | FM895019.1 | JES55 | RNI-like superfamily protein | AT2G39940 | protein binding | func[b] |
| | | | | | ubiquitin-protein ligase activity | func[b] |
| | | | | | stomatal movement | proc[c] |
| | | | | | response to far red light | proc[c] |
| | | | | | jasmonic acid mediated signaling pathway | proc[c] |
| | | | | | defense response | proc[c] |
| | | | | | ubiquitin-dependent protein catabolic process | proc[c] |
| | | | | | regulation of flower development | proc[c] |
| | | | | | shade avoidance | proc[c] |
| 19 | FM894820.1 | JES56 | unknown protein | AT3G57450 | --- | --- |
| 20 | FM894820.1 | JES57 | unknown protein | AT3G57450 | mitochondrial respiratory chain complex I | comp[d] |
| 21 | FM890627.1 | JES59 | unknown protein | AT3G08610 | --- | --- |
| 22 | FM893757.1 | JES62 | HSP20-like chaperones superfamily protein | AT1G52560 | response to hydrogen peroxide | proc[c] |
| | | | | | response to high light intensity | proc[c] |
| | | | | | response to heat | proc[c] |
| 23 | FM893708.1 | JES64 | RNA-binding family protein | AT4G27000 | RNA binding | func[b] |
| | | | | | cytosol | comp[d] |
| 24 | FM893961.1 | JES66 | mitochondrion-localized small heat shock protein | AT4G25200 | molecular_function_unknown | func[b] |
| | | | | | response to cadmium ion | proc[c] |
| | | | | | response to heat | proc[c] |

**Continued**

| | | | | | | |
|---|---|---|---|---|---|---|
| 25 | FM896491.1 | JES67 | Eukaryotic aspartyl protease family protein | AT3G20015 | aspartic-type endopeptidase activity | func[b] |
| | | | | | proteolysis | proc[c] |
| | | | | | endomembrane system | comp[d] |
| 26 | FM896553.1 | JES69 | hydroxyproline-rich glycoprotein family | AT4G05220 | --- | --- |
| 27 | FM896533.1 | JES71 | unknown protein | AT2G36470 | --- | --- |
| 28 | FM894501.1 | JES72 | unknown protein | AT2G15860 | cytosol | comp[d] |
| 29 | FM892805.1 | JES74 | Homeodomain-like superfamily protein | AT5G47660 | regulation of transcription, DNA-dependent | proc[c] |
| | | | | | sequence-specific DNA binding transcription factor activity | func[b] |
| 30 | FM894390.1 | JES75 | No hits found | --- | --- | --- |
| 31 | FM894371.1 | JES76 | indole-3-acetic acid inducible 9 | AT5G65670 | regulation of transcription, DNA-dependent | proc[c] |
| | | | | | auxin mediated signaling pathway | proc[c] |
| | | | | | response to auxin stimulus | proc[c] |
| | | | | | regulation of transcription, DNA-dependent | proc[c] |
| | | | | | response to cyclopentenone | proc[c] |
| | | | | | nucleus | comp[d] |
| | | | | | sequence-specific DNA binding transcription factor activity | func[b] |
| 32 | FM894336.1 | JES78 | No hits found | --- | --- | --- |
| 33 | FM894239.1 | JES81 | Uncharacterised protein family | AT4G19390 | chloroplast | comp[d] |
| 34 | FM894182.1 | JES82 | unknown protein | AT5G65250 | chloroplast | comp[d] |
| 35 | FM891673.1 | JES86 | No hits found | --- | --- | --- |
| 36 | FM889541.1 | JES91 | Actin-binding FH2 family protein | AT5G48360 | actin binding | func[b] |
| | | | | | cellular component organization | proc[c] |
| | | | | | actin cytoskeleton organization | proc[c] |
| 37 | FM889451.1 | JES93 | No hits found | --- | --- | --- |
| 38 | FM889323.1 | JES95 | unknown protein | AT3G03570 | cytosol | comp[d] |
| 39 | FM890156.1 | JES98 | No hits found | --- | --- | --- |
| 40 | FM890148.1 | JES99 | elicitor peptide 6 precursor | AT2G22000 | --- | --- |
| 41 | FM889890.1 | JES102 | Protein kinase superfamily protein | AT5G02290 | protein phosphorylation | proc[c] |
| | | | | | N-terminal protein myristoylation | proc[c] |
| | | | | | ATP binding | func[b] |
| | | | | | phosphorylation | proc[c] |
| | | | | | plasma membrane | comp[d] |
| | | | | | protein kinase activity | func[b] |

**Continued**

| | | | | | | |
|---|---|---|---|---|---|---|
| 42 | FM889794.1 | JES104 | Phosphoenolpyruv ate carboxylase family protein | AT4G10750 | acetate fermentation | proc[c] |
| | | | | | Chloroplast | comp[d] |
| | | | | | carbon-carbon lyase activity | func[b] |
| | | | | | cellular aromatic compound metabolic process | proc[c] |
| | | | | | reductive pentose-phosphate cycle | proc[c] |
| | | | | | gluconeogenesis | proc[c] |
| | | | | | mitochondrion | comp[d] |
| | | | | | catalytic activity | func[b] |
| | | | | | glucose catabolic process to butanediol | proc[c] |
| | | | | | glycolysis | proc[c] |
| | | | | | formaldehyde assimilation via xylulose monophosphate cycle | proc[c] |
| 43 | FM889738.1 | JES106 | No hits found | --- | --- | --- |
| 44 | FM888036.1 | JES110 | ARM repeat superfamily protein | AT4G30990 | binding | func[b] |
| | | | | | Golgi apparatus | comp[d] |
| 45 | FM887831.1 | JES111 | unknown protein | AT1G71900 | --- | --- |
| 46 | FM887718.1 | JES113 | Methylthioalkylma late synthase 1 | AT5G23010 | glucosinolate biosynthetic process | proc[c] |
| | | | | | 2-(2'-methylthio)ethylmalate synthase activity | func[b] |
| | | | | | chloroplast | comp[d] |
| 47 | FM887648.1 | JES114 | oxysterol binding protein | AT5G59420 | cytosol | comp[d] |
| | | | | | oxysterol binding | func[b] |
| | | | | | steroid metabolic process | proc[c] |
| 48 | FM887648.1 | JES115 | oxysterol binding protein | AT5G59420 | cytosol | comp[d] |
| | | | | | oxysterol binding | func[b] |
| | | | | | steroid metabolic process | proc[c] |
| 49 | FM889075.1 | JES119 | zinc finger family protein | AT1G51200 | DNA binding | func[b] |
| | | | | | zinc ion binding | func[b] |
| 50 | FM888913.1 | JES122 | ARM repeat superfamily protein | AT3G60740 | cytosol | comp[d] |
| | | | | | protein folding | proc[c] |
| | | | | | tubulin complex assembly | proc[c] |
| | | | | | embryo development ending in seed dormancy | proc[c] |
| | | | | | tubulin binding | func[b] |
| | | | | | cytokinesis | proc[c] |
| | | | | | microtubule-based process | proc[c] |
| 51 | FM888879.1 | JES125 | Galactose oxidase/kelch repeat superfamily protein | AT2G21680 | --- | proc[c] |

**Continued**

| | | | | | | |
|---|---|---|---|---|---|---|
| 52 | FM888681.1 | JES128 | Ribosomal L28e protein family | AT2G19730 | translation | proc[c] |
| | | | | | cytosolic ribosome | comp[d] |
| | | | | | chloroplast | comp[d] |
| | | | | | translation | proc[c] |
| | | | | | plasma membrane | comp[d] |
| | | | | | cytosolic ribosome | comp[d] |
| | | | | | cell wall | comp[d] |
| | | | | | chloroplast | comp[d] |
| | | | | | structural constituent of ribosome | func[b] |
| | | | | | ribosome biogenesis | proc[c] |
| 53 | FM888604.1 | JES129 | RING/U-box superfamily protein | AT3G06330 | ubiquitin-protein ligase activity | func[b] |
| | | | | | zinc ion binding | func[b] |
| | | | | | protein ubiquitination | proc[c] |
| 54 | FM888513.1 | JES130 | Terpenoid cyclases family protein | AT1G78950 | metabolic process | proc[c] |
| | | | | | beta-amyrin synthase activity | func[b] |
| 55 | FM888472.1 | JES131 | CYCLIN | AT4G34160 | regulation of phosphorylation | func[b] |
| | | | | | nucleus | comp[d] |
| | | | | | response to sucrose stimulus | proc[c] |
| | | | | | response to cytokinin stimulus | proc[c] |
| | | | | | response to cyclopentenone | proc[c] |
| | | | | | response to brassinosteroid stimulus | proc[c] |
| | | | | | regulation of cell cycle | proc[c] |
| | | | | | response to cytokinin stimulus | proc[c] |
| | | | | | regulation of catalytic activity | proc[c] |
| | | | | | protein binding | func[b] |
| | | | | | regulation of cell proliferation | proc[c] |
| 56 | FM887318.1 | JES136 | Mannose-binding lectin superfamily protein | AT1G19715 | --- | --- |
| 57 | FM887287.1 | JES137 | polypyrimidine tract-binding protein-1 | AT3G01150 | regulation of translation | proc[c] |
| | | | | | cytoplasmic mRNA processing body | comp[d] |
| | | | | | nucleic acid binding | func[b] |
| | | | | | regulation of RNA splicing | proc[c] |
| | | | | | pollen germination | proc[c] |
| | | | | | nucleus | comp[d] |
| | | | | | cytoplasm | comp[d] |
| 58 | FM892621.1 | JES139 | unknown protein | AT3G19680 | plasma membrane | comp[d] |

**Continued**

| | | | | | | |
|---|---|---|---|---|---|---|
| 59 | FM887106.1 | JES142 | NAC domain containing protein | AT4G29230 | regulation of transcription, DNA-dependent | proc[c] |
| | | | | | sequence-specific DNA binding transcription factor activity | func[b] |
| | | | | | multicellular organismal development | proc[c] |
| 60 | FM889664.1 | JES144 | unknown protein | AT2G25800 | --- | --- |
| 61 | GR716987.1 | JES145 | Transducin like superfamily protein | AT5G13480 | embryo development ending in seed dormancy | proc[c] |
| | | | | | CUL4 RING ubiquitin ligase complex | comp[d] |
| | | | | | mRNA processing | proc[c] |
| | | | | | regulation of flower development | proc[c] |
| | | | | | protein binding | func[b] |
| 62 | GO247057.1 | JES146 | cellulose synthase | AT5G64740 | membrane | comp[d] |
| | | | | | transferase activity, transferring glycosyl groups | func[b] |
| | | | | | metabolic process | proc[c] |
| | | | | | Golgi apparatus | comp[d] |
| | | | | | cellulose synthase activity | func[b] |
| | | | | | plasma membrane | comp[d] |
| | | | | | cellulose biosynthetic process | proc[c] |
| | | | | | plant-type cell wall biogenesis | proc[c] |
| | | | | | plasma membrane | comp[d] |
| | | | | | cell growth | proc[c] |
| | | | | | primary cell wall biogenesis | proc[c] |
| | | | | | cortical microtubule organization | proc[c] |
| | | | | | response to cyclopentenone | proc[c] |
| 63 | GO247026.1 | JES147 | No hits found | --- | --- | --- |
| 64 | GO246782.1 | JES149 | early nodulin-related | AT5G25940 | mitochondrion | comp[d] |
| 65 | GO246705.1 | JES150 | Pyrophosphorylas e | AT3G53620 | response to cadmium ion | proc[c] |
| | | | | | inorganic diphosphatase activity | func[b] |
| | | | | | membrane | comp[d] |
| | | | | | cytosol | comp[d] |
| 66 | GO246573.1 | JES152 | low-molecular-weight cysteine-rich | AT2G02100 | peptidase inhibitor activity | func[b] |
| | | | | | defense response | proc[c] |
| | | | | | negative regulation of peptidase activity | proc[c] |
| | | | | | plant-type cell wall | comp[d] |
| | | | | | plasma membrane | comp[d] |
| 67 | GO247549.1 | JES154 | glutamine dumper | AT4G25760 | regulation of amino acid export | proc[c] |

**Continued**

| | | | | | | |
|---|---|---|---|---|---|---|
| 68 | GT228727.1 | JES159 | Transducin like superfamily protein | AT5G56130 | CUL4 RING ubiquitin ligase complex | comp[d] |
| | | | | | production of ta-siRNAs involved in RNA interference | proc[c] |
| | | | | | gene silencing by RNA | proc[c] |
| 69 | GT228640.1 | JES160 | Chaperone protein | AT4G24190 | nucleotide binding | func[b] |
| | | | | | unfolded protein binding | func[b] |
| | | | | | response to salt stress | proc[c] |
| | | | | | response to water deprivation | proc[c] |
| | | | | | vacuole | comp[d] |
| | | | | | response to cadmium ion | proc[c] |
| | | | | | chloroplast | comp[d] |
| | | | | | response to cold | proc[c] |
| | | | | | vacuolar membrane | comp[d] |
| | | | | | nucleus | comp[d] |
| | | | | | protein folding | proc[c] |
| | | | | | unfolded protein binding | func[b] |
| | | | | | endoplasmic reticulum | comp[d] |
| | | | | | ATP binding | comp[d] |
| | | | | | vacuolar membrane | comp[d] |
| | | | | | plasma membrane | comp[d] |
| | | | | | protein secretion | proc[c] |
| | | | | | chloroplast | comp[d] |
| | | | | | mitochondrion | comp[d] |
| | | | | | regulation of meristem structural organization | proc[c] |
| 70 | GT228466.1 | JES161 | RNI-like superfamily protein | AT1G47920 | --- | --- |
| 71 | GT228457.1 | JES162 | Methyltransferase-related protein | AT5G58375 | --- | --- |
| 72 | GT228457.1 | JES163 | Methyltransferase-related protein | AT5G58375 | --- | --- |
| 73 | GT229336.1 | JES164 | No hits found | --- | --- | --- |
| 74 | GT229302.1 | JES165 | double-stranded RNA binding protein | AT4G20910 | mRNA cleavage involved in gene silencing by miRNA | proc[c] |
| | | | | | RNA methyltransferase activity | func[b] |
| | | | | | regulation of flower development | proc[c] |
| | | | | | RNA methylation | proc[c] |
| | | | | | nucleus | comp[d] |
| | | | | | specification of floral organ identity | proc[c] |
| | | | | | virus induced gene silencing | proc[c] |
| | | | | | production of miRNAs involved in gene silencing by miRNA | proc[c] |
| | | | | | RNA methyltransferase activity | func[b] |
| | | | | | cytoplasm | comp[d] |
| | | | | | production of siRNA involved in RNA interference | proc[c] |

**Continued**

| | | | | | | |
|---|---|---|---|---|---|---|
| 75 | GT229079.1 | JES167 | indole-3-acetic acid inducible 9 | AT5G65670 | regulation of transcription, DNA-dependent | proc[c] |
| | | | | | auxin mediated signaling pathway | proc[c] |
| | | | | | response to auxin stimulus | proc[c] |
| | | | | | response to cyclopentenone | proc[c] |
| | | | | | nucleus | comp[d] |
| | | | | | sequence-specific DNA binding transcription factor activity | func[b] |
| 76 | GT229050.1 | JES168 | NADH-dependent glutamate synthase 1 | AT5G53460 | nitrate assimilation | proc[c] |
| | | | | | glutamate biosynthetic process | proc[c] |
| | | | | | glutamate synthase (NADH) activity | func[b] |
| | | | | | response to cadmium ion | proc[c] |
| | | | | | plastid | comp[d] |
| | | | | | chloroplast | comp[d] |
| | | | | | ammonia assimilation cycle | proc[c] |
| | | | | | oxidation-reduction process | proc[c] |
| | | | | | developmental growth | proc[c] |
| 77 | GT228943.1 | JES169 | RING-H2 finger A2A | AT1G15100 | protein ubiquitination | proc[c] |
| | | | | | zinc ion binding | func[b] |
| | | | | | protein binding | func[b] |
| | | | | | response to salt stress | proc[c] |
| | | | | | ubiquitin-protein ligase activity | func[b] |
| | | | | | positive regulation of abscisic acid mediated signaling pathway | proc[c] |
| | | | | | regulation of response to osmotic stress | proc[c] |
| 78 | JCST109 | JES171 | IBR domain-containing protein | AT2G31510 | endomembrane system | comp[d] |
| | | | | | zinc ion binding | func[b] |
| | | | | | nucleic acid binding | func[b] |
| 79 | JCST328 | JES172 | Expressed protein | AT3G01345 | membrane | comp[d] |
| | | | | | carbohydrate metabolic process | proc[c] |
| | | | | | hydrolase activity, hydrolyzing O-glycosyl compounds | func[b] |

[a] ESTs of *Jatropha curcas* in the present study, which contained SSRs and also yielded primers through the online software; [b] ESTs-PD categorized under "Molecular Function" when analyzed through TAIR; [c] ESTs-PD categorized under "Biological Process" when analyzed through TAIR; [d] ESTs-PD categorized under "Cellular Component" when analyzed through TAIR.