Plant Omics Journal POJ

POJ 10(02):53-63 (2017) ISSN:1836-3644
doi: 10.21475/poj.10.02.17.pne370

In silico mining and characterization of simple sequence repeats (SSRs) from *Euphorbia esula* expressed sequence tags (ESTs): A potential crop for biofuel

Surojit Sen[1,*], Budheswar Dehury[2], Jagajjit Sahu[3], Sunayana Rathi[4], Raj Narain Singh Yadav[5]

[1]Centre for Biotechnology and Bioinformatics, Dibrugarh University, Assam, India
[2]Biomedical Informatics Centre, Regional Medical Research centre (ICMR), Chandrasekharpur, Bhubaneswar-751023, Odisha, India
[3]Department of Agricultural Biotechnology, Assam Agricultural University, Jorhat-785013, Assam, India
[4]Department of Biochemistry and Agricultural Chemistry, Assam Agricultural University, Jorhat-785013, Assam, India
[5]Department of Life Sciences, Dibrugarh University, Dibrugarh-786004, Assam, India

***Corresponding author: surojit0368@yahoo.co.in**

Abstract

Euphorbia esula (leafy spurge) is a perennial noxious weed native to Eurasia, which can also be potentially useful as a biofuel, medicine, or a pest control agent. In this study, we analysed publicly available ESTs of 'NCBI dbEST' (http://www.ncbi.nlm.nih.gov/dbEST) using *in silico* tools to have an insight into the genetic makeup of the plant. Mining of simple sequence repeats (SSRs) was performed by MISA, primer designing by Primer3 while functional annotation, gene ontology and enrichment analysis were performed by Blast2GO. SSR mining from 47543 ESTs revealed a total 3248 SSRs, of which di-, tri- and hexa-nucleotide repeats were 352 (10.83%), 822 (25.30%), 17 (0.52%) respectively. A total of 527 primer pairs were designed for the annotated SSR-Contigs. About 77.07% SSR-ESTs could be assigned a significant match to the protein database. 210 unique SSR-FDM (Functional Domain Markers) were assigned for significant functional domains by InterProScan. The gene ontology (GO) analysis provided 1213 number of unique GO terms which were subjected to enrichment analysis to obtain 95 statistically significant GO terms mapped to the SSR containing ESTs. Most frequent enriched GO terms were GO:0005886 for cellular component, GO:0003677 for molecular function and GO:0044255 in case of biological process, indicating the potential of the species as a fuel crop. Many SSR markers were functionally annotated to various biotic and abiotic stress responsive genes. Further studies may help us to understand these traits of extreme adaptive features. This will provide opportunity to genetically manage and modify crops for resistance to these stresses.

Key words: Blast2GO, functional annotation, gene ontology, leafy spurge, primer designing, SSR markers.
Abbreviations: cSSR_Compound simple sequence repeats; EC_Enzyme commission; ESTs_Expressed sequence tags; nr-EST_non redundant-EST; FDM_Functional domain markers; FDR_False discovery rate; GO_Gene ontology; IPR_InterProScan; KEGG_Kyoto encyclopedia of genes and genomes; MISA_MIcro SAtellite identification tool; SSR_Simple sequence repeats; SSR-ESTs_SSR containing ESTs; SSR-Contigs_SSR containing contigs; NCBI_ National Centre for Biotechnology Information.

Introduction

Leafy spurge (*Euphorbia esula*), a member of Euphorbiaceae or spurge family is a perennial noxious weed that grows up to 3 feet tall, invades the crop area and creates trouble in cultivation. Leafy spurge is native to Eurasia competing with desirable forage species as well as cultivated plants. Although this plant is not available in the Indian subcontinent, but recently the species has been reported from Kashmir Himalaya (Singh et al., 2014). The acrid sap of the plant is useful externally for warts, and when used internally it shows anti-helminthic, vasodilator and potentially violent purgative properties. Cattle's generally avoid it because it causes scours and weakness in them (Lym and Kirby, 1987; Hein and Miller, 1992). Lot of efforts have been taken for eradication of the weed every year and studies have been carried out to find its potential use as a substitute energy crop (Maxwell et al., 1985, Zixia et al., 2009, Yang et al., 2013), or as pest control agent (Yang and Tang, 1988). Musci et al. (2001) have isolated diterpenes from leafy spurge that showed antiviral properties against herpes simplex virus. Moreover, in China, a dilution of boiled leafy spurge plant material is used to control maggots, mosquito larvae, rats, and some plant diseases. Leafy spurge is being developed as a model to answer fundamental questions of weed biology and to study seed and adventitious root bud dormancy (Chao et al., 2005).

Molecular markers have been used as a tool for detecting genetic diversity as well as for management of plant genetic resources (Ford-LIoyd et al., 1997; Virk et al., 2000; Song et al., 2003; Gepts, 2006). Microsatellites (also known as short tandem repeats, STR; simple sequence repeats, SSRs) are repetitive DNAs composed of tandemly repeated short motifs of 1-6 base pairs (bp). They are hyper-variable and widely spread in both coding and non-coding regions of plant and animal genomes (Kota et al., 2001). SSRs existing in the

coding region are termed type I SSRs and those in non-coding region as type II SSRs (Borstnik and Pumpemik, 2002; Pan et al., 2004; Tassanakajon et al., 2006). The traditional genomic library-dependent approach for SSR markers development is time consuming and expensive. So, *in silico* approach to the development of SSRs form expressed sequence tags (ESTs) is efficient, cost-effective and more preferable.

ESTs represent short, unedited, randomly selected single pass sequence reads derived from cDNA libraries and serve as a main source for *in silico* identification of microsatellites (Sahu et al., 2014). With the rapid increase of ESTs in the public database in recent years, the development of EST containing simple sequence repeats (SSRs) became an attractive choice for the development of SSR markers. Availability of various bioinformatics tools also aided in development of EST-SSR markers at a large scale and cost-effective way (Kantety et al., 2002; Yan et al., 2008). Compared to other DNA markers like amplified fragment length polymorphisms (AFLPs), random amplified polymorphic DNA (RAPD) and genomic SSR markers, EST-SSR markers or genic SSRs are advantageous because they are derived from expressed regions and are more conserved. They have more potential for cross-genera transferability and are used for identifying conserved genomic regions among species, genera, comparative genomics and evolutionary studies (Cordeiro et al., 2001; Kantety et al., 2002; Thiel et al., 2003; Feng et al., 2009). EST-SSR markers are also used for marker assisted selection (Ashkani et al., 2012) and have significant contribution in construction of genetic linkage maps (Kalia et al., 2011; Sraphet et al., 2011).

EST-SSR markers are more abundant in coding regions than in non-coding regions in eukaryotes (Toth et al., 2000; Karaoglu et al., 2005), hence are more conserved and used as functional markers. EST-SSRs are present in gene-rich regions of the genome; they exhibit less polymorphism than genomic based SSRs. Since they are directly associated with genes affecting a particular trait, they have been proved to be a better resource for their use in breeding (Andersen and Lubberstedt, 2003). EST sequences are likely conserved evolutionarily; hence, cross-species PCR (polymerase chain reaction) amplification of EST-SSRs is expected to be more successful than cross-species amplification of SSRs developed from genomic DNA (Arnold et al., 2002; Saha et al., 2003).

Over 47543 ESTs of leafy spurge sequenced from all plant tissues are available at NCBI dbEST (https://www.ncbi.nlm.nih.gov/dbEST/) including tissues from plants that were cold stressed, drought stressed, or attacked by both flea beetles and gall midges. Genomic analysis of the EST sequences would contribute to a better insight of the genetic architecture and correlate the genes, their function, physiological processes that regulate bud dormancy. The current study was designed to identify and characterize EST-SSR markers from the ESTs available at 'dbEST' of NCBI and to develop PCR primers from these EST-SSR markers which will be useful for functional genome mapping and facilitate breeding programs. Functional annotation of EST-SSRs with the corresponding GO (gene ontology), metabolic pathway mapping and enrichment of GO terms with respect to the presence and absence of the SSR markers was performed. Functional annotation of SSR markers associated with various abiotic and biotic stresses may help in understanding the genetics of extreme adaptive traits of the species.

Studies related to SSR markers have been reported from many plants, but no such work has been done on leafy spurge. In this study an attempt has been made to develop EST-SSR markers, and to functionally annotate them to have a better insight of the genetic makeup of the plant. Leafy spurge is a neglected weed; hence very little information is available about the exact genetic background and the genetic relationship among the members of euphorbiaceae. Therefore, EST-SSR markers obtained in this study will be a valuable resource for the comparative mapping in evolutionary studies and in improvement of members of euphorbiaceae. SSRs located within ESTs have the advantage of providing candidate genes that are known to be expressed and tightly linked to each locus. Microsatellite linked association studies will help to explore the properties involved with the markers and designing primers will help in exploring them across the genera or family as SSR markers are known to have cross genera transferability.

Results

Pre-processing and assembly

Sequence pre-processing involved cleaning, repeat and vector masking and organelle masking of 47543 EST sequences into 42761 unique sequences. Assembly of these pre-processed ESTs by CAP3 resulted into 9268 contigs and 14644 singletons. A total of 65.75% of ESTs formed contigs indicating that most of the sequences had overlapping regions (Table 1).

SSR detection and statistics

MISA aided in the detection of 3248 numbers of SSRs and 184 numbers of compound SSR in 23912 assembled sequences. The SSR density was found to be 1 SSR per 5.5 kb and SSR containing EST was 11.77% of the total number of ESTs. Out of 9268 contig sequences, 1271 contigs contained a total of 1446 numbers of SSRs and 59 numbers of cSSR. Out of 14644 singletons, 1544 singletons contained a total of 1802 SSRs and 125 cSSRs (Table 2).

The frequency distribution (density) of SSRs in the EST sequence was critically analysed. In this study, mono-nucleotide repeats were the most abundant repeat type (2057, 63.33%) followed by di-nucleotide (352, 10.84%), tri-nucleotide (822, 25.31%) and hexa-nucleotide (17, 0.52%). Tetra and penta-nucleotide repeats were completely absent. The most frequent classified repeat type was A/T (95.82%) in mono-nucleotide, AG/CT (64.77%) in di-nucleotide, AAG/CTT (34.06%) in tri-nucleotide and AGATGG/ ATCTCC (29.41%) in hexa- nucleotide (Fig. 1a, 1b, 1c). While analysing distribution of SSR with respect to repeat numbers, it was observed that with the change in the number of repeats (mer) for each type of SSR the frequency of occurrence changed. In our study, we have taken repeat number from 5 mer to 10 mer and a separate class for more than 10 mers. It was found that in case of di-nucleotide SSRs 6 mers were highest (185), for tri-nucleotide SSRs 5 mer were highest (615) and for hexa -nucleotide SSR most abundant were five mers (13). Among other type of repeats from 6 mer to more than 10 mer frequency of di-nucleotide was high compared to others (Fig. 2).

Primer designing and screening

It was not possible to design primers for all SSR containing ESTs with the optimum parameters of Primer3. The output of Primer3 was analysed which provided details of primers successfully created. Out of 3064 numbers of ESTs, primers

were successfully designed for 2486 numbers of ESTs with a success rate of 81.13%. A total of 7458 numbers of primer pairs were designed of which 4359 pairs were designed for mononucleotide repeats, 660 primer pairs for di-nucleotide repeats, 2091 primer pairs for tri-nucleotide repeats and for hexa-nucleotide repeats 36 primer pairs were designed. In case of forward primer, average size was 20.27 and average Tm 59.87 and for reverse primer average size was 20.53 with average Tm of 59.80. Average product size in case of contigs was estimated to be 209.15 bp and in case of singleton 206.726 bp (Table 3).

Primers designed against SSR-Contigs having functional annotation was further optimised by PCR Primer Stats (http://www.bioinformatics.org/sms2/pcr_primer_stats.html) and the number of primers was reduced to 527 pairs. They were further checked by FastPCR version 6.5 for successful *in silico* amplification. The list of primers designed against SSR-Contigs which showed virtual amplification through *in silico* PCR is summarised in supplementary table (Table S1).

Functional annotation

Blastx annotation for 1387 SSR-Contigs showed significant matches with 1069 (77.07%) numbers of SSR containing contigs. The average sequence length was 979.75 and average E-value (1.49646E-06), average sequence similarity was 83.67%. Details of blastx annotation of SSR-Contigs for which primers have been screened, is summarised in supplementary table (Table S2). Out of 1271 SSR-Contigs, only 125 SSR-Contigs were assigned to functional domain markers. The domains were analysed from InterPro member databases such as Gene3D, Hamap, PANTHER, Pfam etc. The IntroProScan result is summarised with the obtained 239 numbers of functional domain markers (FDM) for 125 numbers of unique SSR-Contigs (Table 4).

The functional domains were responsible for 2Fe-2S ferredoxin binding, iron sulphur binding domain, ABC transporter conserved site, 3-oxoacyl-[acyl-carrier-protein] synthase 2, Alfin, Annexin repeat, Aquaporin-like domain, ARID/BRIGHT DNA-binding domain, Aspartic peptidase, Beta galactosidase small chain/domain5, C2domain, Calreticulin family, Concanavalin A-like lectin/glucanase subgroup, Cyclophilin-like peptidyl-prolyl cis-trans isomerase domain, DDRGK domain containing protein, DNA-binding WRKY, DnaJ domain, Endopeptidase, NLPC/P60 domain, FAS1 domain, Fatty acyl-CoA reductase, F-box associated interaction domain, Formate-tetrahydrofolate ligase, Galactose mutarotase-like domain, Gibberellin regulated protein, Glucose/ribitol dehydrogenase, Glycoside hydrolase, catalytic domain, Glycoside hydrolase, superfamily, Glycosyl transferase family, HAD-like domain, Heat shock protein DnaJ, cysteine-rich domain (Contig167-IPR001305), Hyaluronan/mRNA-binding protein, Late embryogenesis abundant protein LEA-14 (Contig73-IPR004864), Leucine-rich repeat, Male sterility, NAD-binding (Contig99-IPR013120), Nodulin-like (Contig452-IPR010658), Palmitoyl protein thioesterase, Pollen Olee 1 allergen/extension (Contig455-IPR006041), Protein of unknown function DUF642, Putative small multi-drug export (Contig775), Serine/threonine-/ dual specificity protein kinase, catalytic domain, Stomatin family, Thioredoxin, Water stress and hypersensitive response domain (Contig73-IPR013990) SM00769 (SMART) and several domain of unknown function.

Data of InterproScan results were analysed and a graphical circular image was created with the help of circos to visualise the relationship of the SSR-Contigs with the IPR IDs. It was found that the maximum numbers of significant matches (21) were shown for IPR 001884 whose function is translocation elongation factor IF5A. It was also observed that many EST SSR-Contigs matched with more than one function (Fig.3).

Pathway mapping using KEGG

The SSR-Contigs were annotated with KEGG pathway database and data obtained were analysed to create a circular image to visualise the significant matches of SSR-Contigs with the enzymes with the help of circos. A total of 149 numbers of contigs was assigned to 143 enzyme commissions (EC) related to various metabolic pathways like lipid metabolism, steroid metabolism, purine metabolism, carbon fixation, biotin metabolism, pyruvate metabolism, methane metabolism, pentose phosphate pathway, glutathione metabolism, zeatin biosynthesis, spingolipid metabolism, terpenoid metabolism, flavonoid metabolism and various other biosynthetic processes. Several SSR-Contigs like Contig1833, Contig8097, matched with enzymes of lipid metabolism pathways. Contig9198 matched with EC:1.11.1.9 (peroxidase) of arachidonic acid metabolism. Contig2141 showed maximum weightage of match with EC:2.4.1.17 (1-naphthol glucuronyltransferase) an enzyme of the drug metabolism pathway. Interestingly, Contig3825 matched with EC:6.4.1.2 (carboxylase) of tetracycline and aflatoxin biosynthetic pathway, while Contig5926 matched with EC:1.1.1.133 (reductase) and EC:5.1.3.13 (3,5-epimerase) of streptomycin biosynthetic pathway. Some of the SSR-Contigs were assigned to more than one enzyme commission (Fig. 4).

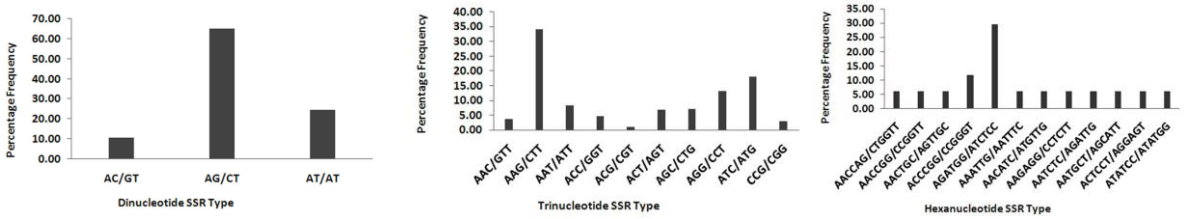
Gene ontology analysis

Gene ontology terms were assigned to SSR-ESTs with significant matches. GO term annotation provides hierarchical common controlled vocabulary to describe gene and gene products across species, i.e., it provides a common terminology for functional description of a transcript as well as the relationship between them. It enables us to communicate unambiguously between different group's annotations of various genomes. It also helps us in discovering patterns across hierarchies. The classified three categories of GO are molecular function (elemental activity), biological process (biological objective or goal), cellular component (location or complex). The GO term annotation revealed a total of 62, 96 and 55 numbers of molecular function, biological process and cellular component respectively (Fig. 5).

Several important molecular functions like binding, catalytic activity, organic cyclic compound binding, heterocyclic compound binding, nucleic acid binding, protein binding, hydrolase activity, transferase activity, kinase, protein kinase and phosphotransferase activity were assigned to SSR containing ESTs. Important biological processes related to oxidation-reduction, metabolic, cellular process, response to stimulus, biosynthesis, macromolecule metabolic process, response to stress, transport, translation were assigned to SSR containing contigs. Among cellular components, cell part, intracellular membrane, intracellular membrane bounded organelle, cytoplasmic, plastid and nuclear components were abundant. Binding activity function was found to be the most

Table 1. Summary of pre-processing and assembly of ESTs.

Total no. of ESTs	ESTs after Pre-processing	No. of ESTs forming contigs (%)	No. of Contigs	No. of Singletons (%)	No of assembled sequences
47543	42761	28117 (65.75%)	9268	14644 (34.24%)	23912



(a) (b) (c)

Fig 1. Frequency of individual repeat types in the SSRs obtained from MISA analysis. (a) Frequency of di-nucleotide repeats; (b) Frequency of tri-nucleotide repeats; (c) Frequency of hexa-nucleotide repeats.

Table 2. Summary of SSR Identified From ESTs.

No. of sequence Examined	No. of SSRs	Frequency (SSR/Kb)	No. of SSR-ESTs	No. of cSSRs
23912	3248	5.50	2815	184

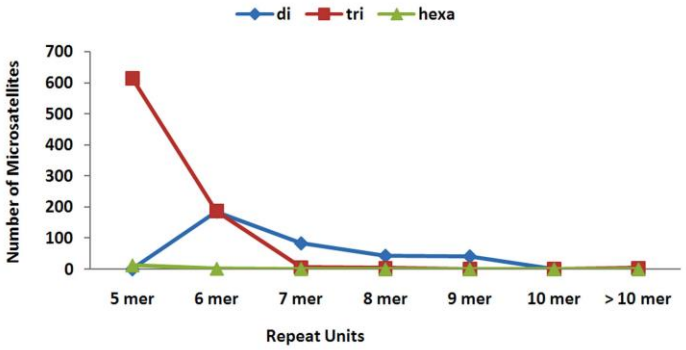


Fig 2. Distribution of SSR with respect to repeat numbers.

Table 3. Summary of Primers designed from the SSR containing ESTs.

No of ESTs containing Primers		No of primer pairs for SSR types							Total no. of primer pairs	Average length of the primers	Average Size Average Size	Average Tm		Average length of the product size
No of Contigs	No of Singlets	Mono-	Di-	Tri-	Tetra-	Penta-	Hexa-	Compound		Forward	Reverse	Forward	Reverse	
1121	1365	4359	660	2091	-	-	36	312	7458	20.28	20.53	59.87	59.80	207.99

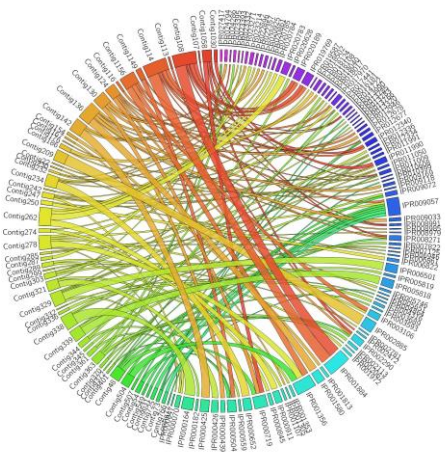


Fig 3. Circos image of InterProScan Data representing significant matches of SSR-Contigs with corresponding function.

Table 4. Summary of functional annotation (IntroProScan results).

No of Contigs	No of Interpro member database										No of FDMs	No of Interpro IDs
	Gene3D	Hamap	PANTHER	Pfam	PIRSF	PRINTS	ProSite	SMART	SUPER-FAMILY	TIGR-FAM		
125	62	3	189	111	4	18	59	29	62	14	239	210

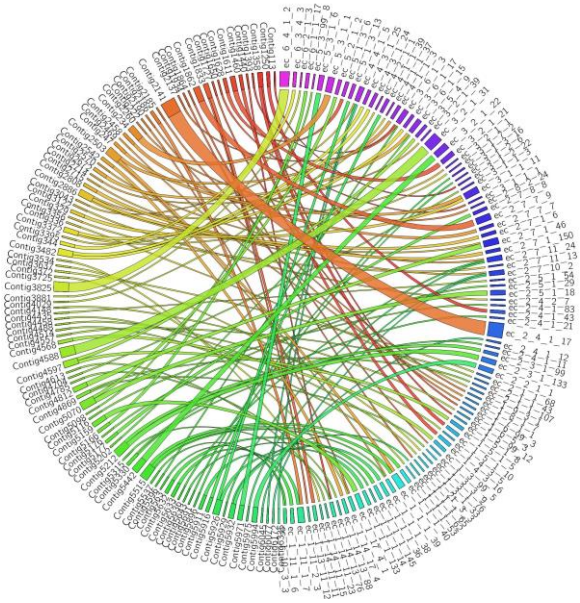


Fig 4. Circos image of KEGG Data representing significant matches of SSR-Contigs with the enzyme Commission of KEGG pathways.

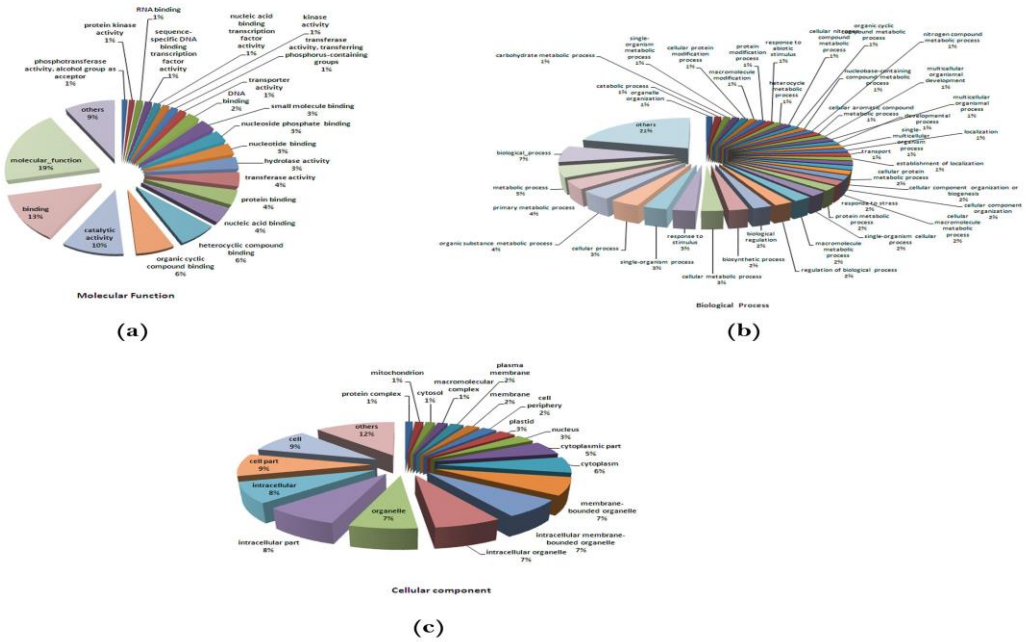


Fig 5. Statistics of GO annotation results of SSR-Contigs. (a) Molecular Function; (b) Biological Process; (c) Cellular Component.

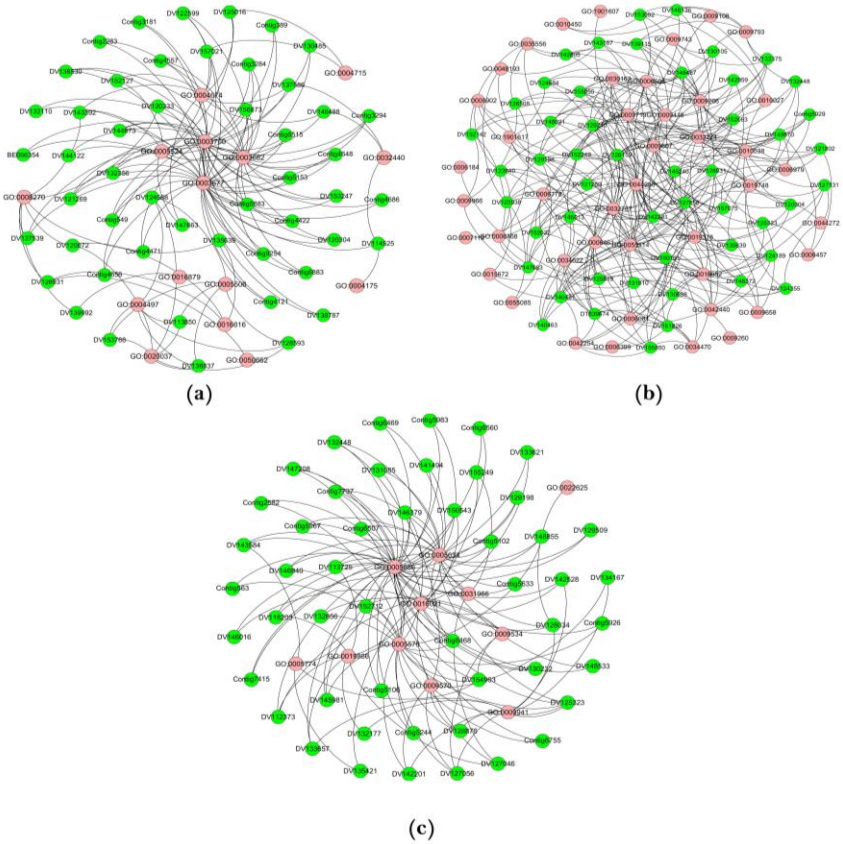


Fig 6. Modular architecture illustrating SSR containing EST and enriched GO terms associated with them. Green and pink nodes represent the EST-IDs, and GO term respectively. (a) Molecular function (Average degree- 2.031 modularity 0.517); (b) Biological process (Average degree: 3.581, modularity: 0.371); (c) Cellular component (Average degree 2.098, modularity 0.354).

abundant molecular function (>13%) followed by binding activity (10%) in addition to this other binding related activity such as cyclic and heterocyclic compound binding, protein binding, ion binding etc. were found. Most abundant biological process was related to metabolic process and most abundant cellular component was intracellular followed by intracellular organelle. Many transcripts did not show any significant similarities with the nr (non-redundant) protein database.

Enrichment analysis

The Gene ontology analysis provided 1213 numbers of GO terms which were subjected to enrichment analysis to obtain statistically significant GO terms mapped to the SSR containing ESTs. Enrichment analysis was done by Fisher exact test on two data sets SSR containing ESTs (Test set) and ESTs without SSR (Reference set). At $p<0.05$ FDR a total of 95 specific enriched GO terms distributed in three subcategories (31 molecular function, 51 biological process, 13 cellular component) were detected. A modular graph was constructed with the help of Gephi which clearly illustrates the networks between SSR containing EST and enriched GO terms associated with them.

There were 15 most frequent enriched GO terms (at $p<0.05$ FDR) found in case of molecular function of which most frequent GO term was GO:0003677 (DNA binding function) and GO:0003700 (sequence-specific DNA binding transcription factor activity) and they were found to be mapped to 33 and 31 unique SSR-ESTs. GO:0003682 associated with 19 specific SSR-ESTs has specific chromatin binding activity. GO:0050662 (coenzyme binding), GO:0004175 (endopeptidase activity), GO:0016879 (ligase activity forming carbon-nitrogen bonds) were least associated terms and were associated with EST singleton DV128931, DV114525 and DV139992 respectively (Fig. 6a). In case of biological process, 43 numbers of GO terms among the 51 enriched GO terms were found to be more frequently associated with SSR containing ESTs. They were involved in various biological processes like, oxidation reduction process, RNA biosynthetic process, secondary metabolic process, cellular lipid metabolic process, transmembrane transport, response to oxidative stress, response to biotic stimulus and response to temperature stimulus. Most frequent enriched GO term was GO:0044255 (cellular lipid metabolic process) and it was found to be associated with 27 numbers of unique SSR ESTs, followed by GO:0032787 (mono carboxylic acid metabolic pathway) found associated with 22 numbers of unique SSR-ESTs. GO:1901607 (alpha-amino acid biosynthetic process), GO:0009108 (coenzyme biosynthetic process), GO:0006399 (tRNA metabolic process), GO:0007112 (male meiosis cytokinesis), GO:0009260 (ribonucleotide biosynthetic process) were found to be associated to least number of unique SSR-ESTs (Fig. 6b). A total of 11 most frequent enriched GO terms was obtained at $p<0.04$ FDR in case of the cellular component category. Significant GO terms were GO:0005886 (plasma membrane), GO:0005634 (nucleolus) and they were found to be mapped to 39 and 21 unique SSR-ESTs respectively. The GO term GO:0004175 (endopeptidase activity) is the least frequent term associated with only one EST singleton DV114525 (Fig. 6c).

Discussion

Euphorbia esula (leafy spurge), apart from its weedy nature has potential to be used as a biofuel, medicine against herpes simplex virus, or a pest control agent. The plant is highly adaptable to various stress condition like, cold, drought, or attack of pathogens. So it can be used as a model for understanding fundamental questions of weed biology. EST sequences are ever increasing in the public database and mining of the EST data may help us to get a better insight of the genomic structure and understand the physiological process involved in bud dormancy as well as resistance to other stresses. Bud dormancy is one of the reasons of extreme tolerance and weediness character of the plant for which it is difficult to control. Simple sequence repeats are an important class of markers to scrutinize the polymorphism in DNA sequence. They are the marker of choice because of multi allelic nature, high abundance, high transferability and extensive genomic coverage to get an insight of genetic constitution of plants (Tautz and Renz, 1984; Gupta et al., 1996; Katti et al., 2001; Kantety et al., 2002; Varshney et al., 2002) and nr-EST (non redundant-EST) sequence provides a more accurate representation of the densities of SSR motifs in the transcribed portion of the genome (Varshney et al., 2005; Poncet et al., 2006). Computational approach of detecting SSRs is a cheaper and faster way of mining SSRs from ESTs than the conventional way.

47543 nr-EST sequences collected from the public domain after preprocessing and assembly resulted into 23912 numbers of EST sequence (9268 contigs and 14644 singletons) which were found to have 3248 numbers of SSRs. The frequency of SSR-EST was found to be 11.77% (i.e. nearly one SSR-EST in every 12 unique EST sequence), which is quite higher than the previous reports in various plants such as 1.5% in maize, 3.2% in wheat, 3.6% in sorghum and 4.7% in rice (Kantety et al., 2002), 2.4% in Arabidopsis, 4.1% in almond and peach, 4.8% in Rosa (Jung et al., 2005), but lower than citrus EST where it was found to be highly abundant 33.3% (Palmieri et al., 2007). The overall density is 1 SSR per 5.5kb which is higher than that in soybean (1/7.4kb), maize (1/8.1kb), tomato (1/11.1kb), and Arabidopsis (1/13kb), poplar (1/14kb) and cotton (1/20kb) (Qiu et al., 2010). However, it is lower than that of rice (1/3.4kb) (Varshney et al., 2002). As found in many other plants, mono-nucleotide repeat motif A/T is the main motif in leafy spurge EST sequence. Apart from mono-nucleotide repeats, tri-nucleotide repeats were more predominant (25.31%) than the di-nucleotide repeats (10.84%) and only 17 numbers of hexa-nucleotide repeats (0.52%) were found. The higher frequencies of TNR (tri-nucleotide repeats) is in accordance with previous studies (Corderio et al., 2001; Kantety et al., 2002; Varshney et al., 2002; Thiel et al., 2003; Nicot et al., 2004; Sahu et al., 2014). Abundance of tri-nucleotide SSR repeats in ESTs can be attributed to absence of frameshift mutations in the coding regions having SSRs of varying length (Metzgar et al., 2000). Among the di-nucleotide repeats most prevalent repeat type was AG/CT (64.77%) which is in accordance with earlier studies in cassava (Lopez et al., 2007; Zou et al., 2011) and in Arabidopsis, rice, soybean, maize, oil palm, coffee, barley, wheat, rubber tree (Morgante et al., 2002; Nicot et al., 2004; Thiel et al., 2003; Poncet et al., 2006; Low et al., 2008; Feng et al., 2009). In fact, AG/CT is the most common repeat type in ESTs of all vascular plants (Victoria et al., 2011; Cardle et al., 2000). Further, it was noticed that di-nucleotide repeat type GC/CG was completely absent which correlates with earlier studies carried out in sequences of other plant species (Morgante and Oliveri 1993, Cardle et al., 2000) and is in fact the least frequent SSR motif in almost all studied organisms except *Escherichia coli* (Trivedi S, 2004). Among tri-nucleotide classified repeats AAG/CTT is most frequent (34.06%) as found in castor bean SSR (Qiu et al., 2010) and many plants like *Carica papaya, Ananas comosus, Catharanthus roseus, Dioscorea alata, Jatropha curcas, Mangifera indica* and others studied by Sahu et al. (2014) and reconfirms the results of a study performed by Morgante et al. (2002), that AAG/CTT is predominant and CCG/CGG is rare in dicotyledonous plants. Tetra-nucleotide and penta-nucleotide repeats were totally absent in the EST sequence and only 17 numbers of hexa-nucleotide repeats were found of which the motif AGATGG/ATCTCC was predominant.

Compound microsatellites (cSSRs) are special variants of microsatellites consisting two or more repeats in close proximity. The cSSRs, although present in very low density in nucleotide sequences are proven to be most important (Bull et al., 1999). Majority of compound microsatellites has originated by duplication of imperfection in microsatellite tract (Kofler et al., 2008). In the present study a total of 184 numbers of cSSRs was detected, i.e., 5.36% of the total SSRs present. The frequency of compound SSR was found better than previous studies done on fully sequenced species (*Maccaca mulata, Mus musculus, Rattus norvegicus, Ornithirhynchus anatinus, Gallus gallus, Danio rerio*, and *Drosophila malanogaster*) where it ranged from 4-25% (Weber, 1990; Kofler et al., 2008). In case of complete *Escherichia coli* genome it was found to be 1.75%-2.85% (Chen et al., 2011). Compound microsatellites are composed of two or more individual microsatellites adjacent to each other. The composition of compound microsatellite is complicated because of the variable number of repeats in cSSR. For instance, the motif 'm1-Xn-m2' is termed '2-microsatellite' and 'm1-Xn-m2-Xt-m3' as '3-microsatellite' and so on. Analysis of compound microsatellites found from MISA results revealed that in both contigs and singletons, '2-microsatellites' has maximum density followed by '3-microsatellite'. In the contigs, the largest compound microsatellites were found to be '3-microsatellites' whereas in singletons, many large sized compound SSRs were observed with a maximum size of '8-microsatellites'. Compound microsatellite without any interrupting sequences are very few in number and are denoted as C* in the MISA results only four numbers were found in contigs (GAA)5(A)10*, (T)13(TG)8*, (TCT)6(TC)6*, (A)17(AC)6* and only one (CTT)6c(GT)6(T)11* was found in singleton.

Primers were developed for large numbers of SSR containing ESTs with a success rate of 81.13%. But it was not possible to design the primers for remaining SSR-ESTs, may be due to inadequate size of the sequence flanking at both ends of the SSRs. Designed primer pairs may be used for gene tagging, genetic mapping and population studies. Screening of developed primers based on properties like percent GC content, self annealing, hairpin formation, reduced the numbers to 527 pairs. They were further tested for selectivity by FastPCR and amplicons of expected size were observed. Although EST-SSRs are generally less polymorphic than genomic SSRs (Eujayl et al., 2001; Gupta et al., 2003), the value of EST-SSRs is enhanced by their transferability across taxa, and their potential as functional markers in defining genes that affect traits of interest. Transferability of polymorphic SSR markers within genera is more (60% success rate in eudicots) than between genera (10% success rate in eudicots) within same family (Barbará et al., 2007). Primers developed from these cross transferable polymorphic markers can be used randomly for comparative mapping as

well as for gene cloning by identifying cross referencing gene.

IntroProScan annotated functional domain to the SSR containing contigs, the translated nucleotides were classified into families, domains and important sites. A total of 239 functional domain markers were associated with 129 SSR containing contigs. The sequences having SSRs matching with FDMs signify that functional domains provide predicted functions to the molecular markers. Important functions like late embryogenesis, water stress and hypersensitive domain, heat shock protein DnaJ, oxidative stress response, biotic and abiotic stress response was identified justifying the extreme tolerance and weediness character. Several domains of unknown function were also identified, such as Contig1055-IPR008217, Contig495-IPR009769, Contig888-IPR019446, Contig22-IPR010608, Contig285-IPR006946, which may serve as novel genes.

Gene ontology annotation identified a total of 213 unique GO terms and majority of the SSR loci were associated with 62 molecular functions, 96 biological processes and 55 cellular components. Among the important molecular function predicted, enzyme activity was found to be more frequent and other functions like, cyclic and heterocyclic compound binding activity, nucleic acid binding, protein binding, small molecule binding, DNA and RNA binding activities were less frequent. Among the least frequent activities were trans-membrane transporter activity, motor activity, oxygen binding activity, receptor activity, antioxidant activity, nuclease activity and many other activities, all grouped under head 'others' totaling to 9% of the total activity. Among the different biological functions predicted by GO analysis the most significant biological function was found to be biological process and metabolic process including primary metabolism and organic substance metabolic process. Among other processes, notably found to be frequent were responses to stimulus, protein metabolic process, response to stress, indicating the susceptible nature of the plant to various biotic and abiotic stresses. Various other processes like aminoacylation, protein modification, nitrogen metabolism, cell signaling, cell death, embryo development (0.27%), secondary metabolic process, were found to be negligible. Among the cellular components, the intracellular component was of highest frequency followed by intracellular organelle, cytoplasm, and nucleolus. The less frequent components included chromosomes, cytoskeleton, peroxisome, microbody, ribosome, endoplasmic reticulum, golgi etc. and all were less than 1% and hence grouped in the head 'others'. SSR-Contigs with no significant similarities with the protein database indicates the potential of discovery of new genes in the plant.

Gene enrichment analysis revealed the most dominant EST IDs which were mapped to the highest number of GO terms. A total of 1213 numbers of unique GO terms obtained from the gene ontology analysis were subjected to enrichment analysis to get statistically significant GO terms mapped to the SSR containing ESTs. Results of enrichment analysis were used to construct network graphs in three different categories. Network graph revealed that in case of molecular function most frequent enriched GO terms were GO:0003677 (DNA binding function) and GO:0003700 (sequence-specific DNA binding transcription factor activity) and they were found to be mapped to 33 and 31 unique SSR-ESTs, indicating high levels of transcription activities.

SSR markers containing ESTs were found to be associated with various biological processes like, oxidation reduction process, RNA biosynthetic process, secondary metabolic process, cellular lipid metabolic process, trans-membrane transport, response to oxidative stress, response to biotic stimulus and response to temperature stimulus. Most frequent enriched GO term was GO:0044255 (cellular lipid metabolic process) and the least frequent was ribonucleotide biosynthetic process and tRNA metabolic process. The high rate of cellular lipid metabolism indicates the potentiality of the species as fuel crop which has been indicated by studies done earlier (Maxwell et al., 1985; Zixia et al., 2009; Yang et al., 2013).

A total of 11 numbers of enriched GO terms in the cellular component category was found while considering 50 most frequent SSR-ESTs. It was found that the maximum number of SSR EST IDs (39 numbers) were associated to GO term GO:0005886 (plasma membrane), followed by GO:0005634 (nucleus) associated to 24 numbers of SSR containing EST IDs. GO:0005576 (extracellular region) was associated with 17 numbers of IDs followed by GO:0016021 (integral membrane), GO:0009941 (chloroplast membrane), GO:0009570 (chloroplast stroma), GO:0019866 (organelle inner membrane), GO:0031966 (mitochondrial membrane) each associated with 5 numbers of EST IDs. GO:0009534 (chloroplast thylakoid), GO:0005774 (vacuolar membrane) were associated with 4 numbers of EST IDs. The least associated GO term was GO:0022625 (cytosolic large vacuolar subunit). Many of these organelles are involved in lipid metabolism.

Materials and Methods

EST sequence retrieval

EST sequences of *Euphorbia esula* were collected from the public domain NCBI dbEST (http://www.ncbi.nlm.nih.gov/dbEST/), a public domain of NCBI. A total of 47543 (as of 1st December, 2015) sequences from different tissues were retrieved and was subsequently subjected to EST pre-processing.

EST pre-processing and assembly (Sequence cleansing)

Pre-processing and assembly were done using the most popular web server EGassembler (Masoudi-Nejad et al., 2006). The sequences were cleaned using cross_match with both min-match and min-score value of 20 to remove vector contamination. Trimming of polyA tail was conducted with Trimest tool of EMBOSS with minimum length value 4 and mismatch value 1. Quality sequences obtained after preprocessing were assembled using CAP3 with default parameters as described and used in the supplementary materials (Huang and Madan, 1999). In this study a standalone processing was opted allotting 6 CPUs for the analysis.

Mining of SSRs

The assembly of sequences resulted in contigs and singletons which were further analysed for identification of SSR using MISA (MIcro SAtellite identification tool) (Thiel et al., 2003). Both "misa.pl" (executive file of MISA tool) and "misa.ini" (search parameters) were downloaded from the website (http://pgrc.ipkgatersleben.de/misa/misa.html). In this study following search detection parameters were employed:

-Maximum difference between two SSR interrupting to form a compound microsatellite was 100.

-Minimum length parameter for repeated units i.e. (unit size/minimum number of repeats) were at least 10 mono-

nucleotides (1/10), at least six di-nucleotide (2/6), five tri-nuclotides (3/5), five tetra-nuclotide (4/5), five penta-nucleotide (5/5), and five hexa-nucleotide (6/5).

Primer design

Primer designing for SSR containing singletons and contigs was done by Primer3 tool (Rozen and Skaletsky, 2000). The results obtained from MISA were used as input for Primer3 tool with the help of Perl scripts (http://pgrc.ipk-gatersleben.de/misa/primer3.html). Primers were designed only for SSR containing ESTs and primer specificity was optimised for primers greater than 10 bp on either side of the SSR with maximum product size 100-280 bp. Following parameters were employed for primer designing:

- Optimal size of primer was set as 18 bp maximum upto 27 bp
- Melting temperature 55^0 C with minimum 50^0 C and maximum 70^0 C
- Maximum GC content of 65%.

The properties of each designed primers of SSR-contigs were further optimised for melting temperature, percent GC content, self-annealing, hairpin formation and PCR suitability by PCR Primer Stats (http://www.bioinformatics.org/sms2/pcr_primer_stats.html). The screened primers were then crosschecked for *in silico* amplification by FastPCR (Kalendar et al., 2014) version 6.5.

Functional domain marker and GO term analysis

Microsatellites containing EST sequences were functionally annotated and GO terms depicting three descriptors, i.e., BP (biological process), CC (cellular component) and MF (molecular function) were assigned to them with the help of B2G (Blast2GO) (Conesa et al., 2005) version 2.8.0 (http://blast2go.com). Putative functions were assigned by remotely blasting to the NCBI database with the help of qblast or remote blast with default parameters like blastx programme, non-redundant database, expect (E) value of 1.0E-3, number of blast hits 20 were selected. Mapping was done to get GO terms associated to each of the obtained hits to evaluate GO annotation of the query sequences. InterPro motifs were queried at the InterProScan web service incorporated in Blast2GO to obtain functional domain markers. InterPro is a collection of protein signature databases and matches of IPR IDs against the queried sequences was obtained.

Pathway mapping was performed using Blast2GO, the annotated sequences were queried against the KEGG database to get enzyme commission (EC) IDs associated to the SSR containing contigs. Data obtained from InterProScan and mapping through KEGG database were exported to Microsoft Excel sheet and a separate datasheet was prepared for the creation of a circular image with the help of online circos (Krzywinski et al., 2009) tool to visualise the relationship of the SSR-Contigs with the corresponding IPR IDs and enzyme codes.

Enrichment analysis of SSR containing ESTs

To test the annotation specificity between two sets of sequences with respect to GO terms, enrichment analysis was performed with the help of BLAST2GO. In this study both contigs and singletons were taken into consideration. Two sequence datasets were prepared, i.e., one with ESTs containing SSR markers (as test) and the other without SSR markers (as reference). Fisher's exact test (two tailed) was performed between the two data sets with term filter value 0.05 FDR (False Discovery Rate) using BLAST2GO. Enriched GO terms for each category (viz., MF, BP and CC) were subjected to Gephi (https://gephi.org) for construction of network graph (Bastian et al., 2009). While constructing network graphs, 50 most frequent SSR-EST IDs associated to maximum number of GO terms were considered in each category (i.e., MF, BP and CC).

Conclusion

SSRs located within ESTs have the advantage of providing candidate genes that are known to be expressed and tightly linked to each locus. SSR-ESTs associated with high rate of cellular lipid metabolism indicate the potentiality of the species as a fuel crop. Many SSR markers were found to be associated with oxidative stress response, biotic and abiotic stress response, peptidase activity. Further studies may help us to understand these traits of extreme adaptability features which helped leafy spurge to become successful colonizer and also to understand how these extreme adaptive features evolved. By understanding the genetic basis, we will be able to track the evolution of these traits, their inheritance as well as dispersal. Understanding stress-specific genes, functional repertoire, pathways and phenotypes associated with various stress signals would help to devise plants that could endure adverse environmental conditions and may eventually help to reduce crop productivity losses. An effective understanding of such key pathways and molecular connections would help to develop plants with traits that confer tolerance to biotic and abiotic stresses. Knowledge developed in this direction will help in development of stress resistant crops as well as in the study of weed management.

References

Andersen JR, Lubberstedt T (2003) Functional markers in plants. Trends Plant Sci. 8:554-560.

Arnold C, Rossetto M, McNally J, Henry RJ (2002) The application of SSRs characterized for grape (*Vitis vinifera*) to conservation studies in Vitaceae. Am J Bot. 89:22-28.

Ashkani S, Rafii MY, Ibrahim R, Meon S, Abdullah SNA, Rahim HA, Latif MA (2012) SSRs for marker-assisted selection for blast resistance in rice (*Oryza sativa* L.). Plant Mol Biol Report. 30(1):79-86.

Barbará T, Palma-Silva C, Paggi GM, Bered F, Fay MF, Lexer C (2007) Cross-species transfer of nuclear microsatellite markers: Potential and limitations. Mol Ecol. 16:3759-67.

Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

Borstnik B, Pumpemik D (2002) Tandem repeats in protein coding regions of primate genes. Genome Res. 12:909-915.

Bull LN, Pabo´n-Pen˜a CR, Freimer NB (1999) Compound microsatellite repeats: Practical and theoretical features. Genome Res. 9:830-838.

Cardle L, Ramsay L, Milbourne 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.

Chao WS, Horvath DP, Anderson JV, Foley MP (2005) Potential model weeds to study genomics, ecology, and physiology in the 21st century. Weed Sci. 53:929-937.

Chen M, Zeng G, Tan Z, Jiang M, Zhang J, Zhang C, Lu L, Lin Y, Peng J (2011) Compound microsatellites in complete *Escherichia coli* genomes. FEBS Lett. 585:1072-1076.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21(18):3674-6.

Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. Plant Sci. 160:1115-1123.

Eujayl I, Sorrells M, Baum M, Wolters P, Powell W (2001) Assessment of genotypic variation among cultivated durum wheat based on EST-SSRs and genomic SSRs. Euphytica. 119:39-43.

Feng SP, Li WG, Huang HS, Wang JY, Wu YT (2009) Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (*Hevea brasiliensis*). Mol Breed. 23:85-97.

Ford-Lloyd BV, Jackson MT, Newbury HJ (1997) Molecular markers and the management of genetic resources in seed gene banks: A case study of rice. In: Biotechnology and Plant Genetic Resources. Conservation and Use. CAB International, Wallingford, UK, pp. 103-118.

Gepts P (2006) Plant genetic resources conservation and utilization. Crop Sci. 46(5):2278-92.

Gupta PK, Balyan HS, Sharma PC, Ramesh B (1996) Microsatellites in plants: a new class of molecular markers. Curr Sci.70:45-54.

Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS (2003) Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. Mol Genet Genomics. 270:315-323.

Hein DG, Miller SD (1992) Influence of leafy spurge on forage utilization by cattle. J Range Manage. 45:405-407.

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res. 9(9):868-77.

Jung S, Abbott A, Jesudurai C, Tomkins J, Main D (2005) Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs. Funct Integr Genomics. 5(3):136-43.

Kalendar R, Lee D, Schulman AH (2014) FastPCR software for PCR, *in silico* PCR, and oligonucleotide assembly and analysis. In: Svein Valla and Rahmi Lale (ed) DNA Cloning and Assembly Methods, Methods in Molecular Biology, Humana Press, New York 1116:271-302.

Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK (2011) Microsatellite markers: an overview of the recent progress in plants. Euphytica. 177(3):309-334.

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.

Karaoglu H, Lee CM, Meyer W (2005) Survey of simple sequence repeats in completed fungal genomes. Mol Biol Evol. 22:639-349.

Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol. 18:1161-1167.

Kofler R, Schlo¨tterer C, Luschu¨tzky E, Lelley T (2008) Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites. BMC Genom. 9:612-626.

Kota R, Varshney RK, Thiel T, Dehmer KJ, Graner A (2001) Generation and comparison of EST-derived SSRs and SNPs in barley (*Hordeum vulgare* L.). Hereditas. 135:145-151.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, HorsmanD, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res. 19(9):1639-1645.

L´opez CE, Quesada-Ocampo LM, Boh´orquez A, Duque MC, Vargas J, Tohme J, Verdier V (2007) Mapping EST-derived SSRs and ESTs involved in resistance to bacterial blight in *Manihot esculenta*. Genome. 50(12):1078-1088.

Low ETL, Alias H, Boon SH, Shariff EM, Tan CYA, Ooi LCL, Cheah SC, Raha AR, Wan KL, Singh R (2008) Oil palm (*Elaeis guineensis* Jacq.) tissue culture ESTs: Identifying genes associated with callogenesis and embryogenesis. BMC Plant Biol. 8:62.

Lym RG, Kirby DR (1987) Cattle foraging behavior in leafy spurge infested rangeland. Weed Technol. 1:314-318.

Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T, Goto S (2006) EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. Nucleic Acids Res. 34 (Web Server issue), W459-62.

Maxwell BD, Wiatr SM, Fay PK (1985) Energy potential of leafy spurge (*Euphorbia esula*). Econ Bot. 39(2):150-156.

Metzgar D, Bytof J, Wills C (2000) Selection against frame shift mutations limits microsatellite expansion in coding DNA. Genome Res. 10:72-80.

Morgante M, Olivieri AM (1993) PCR-amplified microsatellites as markers in plant genetics. Plant J. 3(1):175-182.

Morgante M, Hanafey M, and Powell W. (2002). Microsatellites are preferentially associated with non-repetitive DNA in plant genomes. Nat Genet. 30:194-200.

Mucsi I, Molnar J, Hohmann J, Redel D (2001) Cytotoxicities and anti-herpes simplex virus activities of diterpenes isolated from Euphorbia species. Planta Med. 67:672-674.

Nicot N, Chiquet V, Gandon B, Amilhat L, Legeai F, Leroy P, Bernard M, Sourdille P (2004) Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs). Theor Appl Genet. 109:800-805.

Palmieri DA, Novelli VM, Bastianel M, Cristofani-Yaly M, Astua-Monge G, Carlos EF, de Oliveira AC, Machado MA (2007) Frequency and distribution of microsatellites from ESTs of citrus. Genet Mol Biol. 30(3):1009-1018.

Pan YW, Chou HH, You EM, Yu HT (2004) Isolation and characterization of 23 polymorphic microsatellite markers for diversity and stock analysis in tiger shrimp (*Panaeus monodon*). Mol Ecol Notes. 4:345–347.

Poncet V, Rondeau M, Tranchant C, Cayrel A, Hamon S, deKochko A, Hamon P (2006) SSR mining in coffee tree EST databases: Potential use of EST-SSRs as markers for the Coffea genus. Mol Genet Genomics. 276:436–449.

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. 10:278.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 132:365-386.

Saha S, Karaca M, Jenkins JN, Zipf AE, Reddy O, Umesh K, Kantety RV (2003) Simple sequence repeats as useful resources to study transcribed genes of cotton. Euphytica. 130:355-364.

Sahu J, Sen P, Choudhury MD, Dehury B, Barooah M, Modi MK, Talukdar AD (2014) Rediscovering medicinal plants' potential with OMICS: microsatellite survey in expressed sequence tags of eleven traditional plants with potent antidiabetic properties. OMICS. 18(5):298-309.

Singh G, Khuroo AA, Ganie AH , Tali BA, Malik AH (2014) *Euphorbia esula* L. (Euphorbiaceae): a new plant record for Indian subcontinent from Kashmir Himalaya. Phytodiversity. 1(1&2):1-6.

Song ZP, Xu X, Wang B, Chen JK, Lu BR (2003) Genetic diversity in the northernmost *Oryza rufipogon* populations estimated by SSR markers. Theor Appl Genet. 107:1492-1499.

Sraphet S, Boonchanawiwat A, Thanyasiriwat T, Boonseng O, Tabata S, Sasamoto S, Shirasawa K, Isobe S, Lightfoot DA, Tangphatsornruang S, Triwitayakorn K (2011) SSR and EST-SSR-based genetic linkage map of cassava (*Manihot esculenta* Crantz). Theor Appl Genet. 122(6):1161-1170.

Tassanakajon A, Klinbunga S, Paunglarp N, Rimphanitchayakit V, Udomkit A, Jitrapakdee S, Sritunyalucksana K, Phongdara A, Pongsomboon S, Supungul P, Tang S, Kuphanumart K, Pichyangkura R, Lursinsap C (2006) *Panaeus monodon* gene discovery project: the generation of an EST collection and establishment of a database. Gene. 384:104-112.

Tautz D, Renz M (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res. 12:4127-4138.

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

Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10:967-981.

Trivedi S (2004) Microsatellites (SSRs): puzzles within puzzle. Indian J Biotechnol. 3(3):331–347.

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 VK, Dayal R, Bhandari RS, Jyoti KN, Prasuna AL, Prasad AR, Yadav JS (2005) Behavioral response of the borer beetle *Hoplocerambyx spinicornis* to volatile compounds of the tree *Shorea robusta*. Chem Biodivers. 2:785-791.

Victoria FC, da Maia LC, de Oliveira AC (2011) *In silico* comparative analysis of SSR markers in plants. BMC Plant Biol. 11:15.

Virk PS, Newbury JH, Bryan GJ, Jackson MT, Ford-Lloyd BV (2000) Are mapped or anonymous markers more useful for assessing genetic diversity? Theor Appl Genet. 100:607-613.

Weber JL (1990) Informativeness of human (dC-dA)n.(dGdT)n polymorphisms. Genomics. 7:524-530.

Yan Q, Zhang Y, Li H, Wei C, Niu L, Guan S, Li S, Du L (2008) Identification of microsatellites in cattle unigenes. J Genet Genomics. 35:261-266.

Yang JH, Liu YJ, Li JK, Huang JX, Zhang WY, Li SY (2013) Potential species and character of wild diesel plant in Tianjin. Adv Mat Res. 641:578-582.

Yang RZ, Tang CS (1988) Plants used for pest control in China: a literature review. Econ Bot. 42(3):376-406.

Zixia G, Baocheng W, Lin YW, Yue YH (2009) Analysis of total lipid contents and fatty acids composition of three species of *Euphorbia* in Jiangsu Province. Chemistry and Industry of Forest Products. 29(4):63-66.

Zou M, Xia Z, Ling P, Zhang Y, Chen X, Wei Z, Bo W, Wang W (2011) Mining EST-derived SSR markers to assess genetic diversity in cassava (*Manihot esculenta* Crantz). Plant Mol Biol Report. 29(4):961-971.