Indian Journal of Agricultural Sciences **85** (3): 429–33, March 2015/Article

Identification and characterization of microsatellites in ESTs of *Rosa* species: Insight in development of SSR markers

SAPNA PANWAR[1], KANWAR PAL SINGH[2], NAMITA[3], HUMIRA SONAH[4], RUPESH K DESHMUKH[5] and T R SHARMA[6]

Indian Agricultural Research Institute, New Delhi 110 012

Received: 11 February 2014; Revised accepted: 12 September 2014

ABSTRACT

The present investigations were undertaken during the year 2012. In this study, analysis of microsatellites identified in expressed sequence tags (ESTs) of *Rosa* species was performed. Entire set of ESTs for genus *Rosa* including 1 792 of *Rosa chinensis*, 5 495 of *R.*× *hybrida* and 1 849 of *R. lucieae* was retrieved from public database and assembled individually. Furthermore, 668 microsatellites (SSR) were identified in the 5 272 assembled ESTs in which di-nucleotide repeats were most frequent. The highest 55.2 percentage of di-nucleotide was identified in *R. chinensis* followed by *R.*× *hybrida* (50.1%) and *R. lucieae* (44.3%). Repeats with motif sequences AG/CT and AAG/CTT were observed as most frequent motifs. Moreover, SSRs were also identified in non-EST sequences available for the genus *Rosa*. The entire sequences were assembled together to develop non-redundant data that was subsequently used to design 293 SSR markers. Analysis of microsatellite and resource of SSR markers developed in present study will help for undertaking the molecular research in *Rosa* species.

Key words: Expressed Sequence Tags (ESTs), Markers, Microsatellites, *Rosa* species, SSR

Rose is one of the most economically important genus belonging to family Rosaceae. The genus possesses very high level of diversity among species and these hybridize very easily (Gridin 2010) which resulted in development of wide range of hybrids. In this regards, assessment of genetic diversity is preferred to improve selection of parental genotypes and for development of elite hybrids (Namita *et al.* 2011)

Study of *Rosa* evolution is not only ecologically important but provides a basis for the effective utilization of available genetic resources. Earlier, morphological characters have been explored for systematic and evolutionary studies (Foote 1997, Panwar *et al*. 2010). Later on with the advancements in DNA sequencing technology, DNA sequences are being used for this purpose (Buckler and Thornsberry 2002) and moreover, the genome sequence of a species has its unique attributes like nucleotide frequency, transposable elements and repetitive sequences. Initially, repetitive sequences were considered as evolutionary neutral but recent studies demonstrated the relevance of repetitive sequences in genome evolution (Zhu *et al*. 2000). Repetitive sequence are generally categorised as microsatellites and minisatellites on the basis of motif length (McCouch *et al*. 1997). The microsatellites (SSRs) are having 1–6 bp motif length which tandemly repeated several times and thus, SSRs are almost uniformly distributed throughout the plant genomes in abundant quantity. However, the evolution of SSRs mostly depends on their occurrence in the entire genome for instance, SSRs present in coding region of genes have high level of selection pressure hence, evolve slowly. However, the variation in coding SSRs emphasised more on the function of gene which ultimately decide the phenotype (Ellis and Burke 2007, Kulkarni *et al*. 2012), therefore, SSRs present in coding sequences have great importance for genomic research.

Molecular characterization of different species in the genus *Rosa* is a pre-requisite for crop improvement as well as to enhance evolutionary understanding (Rout *et al.* 1999, Gudin 2010). In this regard, development of molecular markers will help in several ways as they are being frequently used in plant biology for applications like diversity analysis, development of linkage maps, QTL mapping, and marker-assisted breeding (Kashyap *et al*. 2010, Channamallikarjuna *et al*. 2010, Sharma *et al*. 2010). In addition, several types of molecular marker techniques have been developed and compared for their efficiency and competency (Nybom 2004). Microsatellite (SSR) markers are one of the most efficient markers that are mostly preferred because of the

[1]Scientist (email: sapna.panwar8@gmail.com), [2]Principal Scientist (email: kanwar_iari@yahoo.co.in), [3]Scientist (email: namitabanyal@gmail.com), [4]Research Associate (email: biohuma@gmail.com), [5]Research Associate (email: rupesh0deshmukh@gmail.com), [6]Director (email: trsharma@nrcpb.org), National Research Centre on Plant Biotechnology, IARI, New Delhi 110 012

attributes like co-dominance, PCR based assay, easy, abundance availability with uniform distribution in genome etc. (Deshmukh *et al*. 2010). However, the cost of development for microsatellite markers is high and involves extensive laborious methods (McCouch *et al*. 1997). Hence, with the development in sequencing technology, several plant genomes have been sequenced with international collaborative efforts (www.phytozome.net). The technology makes possible to mine microsatellites in the genomic sequence and convert it in efficient markers. However, apart from the genomic sequences, gradually accumulating expressed sequence tags (ESTs) from several independent studies are available for relatively large number of horticultural crop species (Sonah *et al*. 2011b) which provides an alternative way for SSR marker development. Moreover, expressed sequence exhibited more conserved lineage among the related species (Bhati *et al*. 2010). It facilitated transferability of EST derived SSR markers (EST-SSRs) developed in a species to other related species (Bhati *et al.* 2010). Zhang *et al*. (2014) developed EST derived SSR markers for pear and also reported cross-species transferability in Rosaceae. The information will be helpful in better understanding of genetic relationship, evolution, comparative genomics and gene introgressions. Banemann *et al*. (2012) identified, characterized and finally validated the EST-SSR markers derived from gerbera EST database. Similarly, Shirasawa *et al.* (2013) developed EST derived SSR markers from publicly available EST sequences which serve as an information tool in revealing the relationships between capsicum lines. In present investigations, microsatellites were identified in ESTs and other sequence fragments available in public domain for the different *Rosa* species. Subsequently, the information was used for the development of EST-SSR markers and made available for research community through online database.

MATERIALS AND METHODS

The present study included different *Rosa* species such as *Rosa chinensis*, *Rosa* × *hybrida* and *Rosa lucieae*.

Entire EST sequences for the *Rosa* species including *R. chinensis* (1 792 ESTs), *R.* × *hybrida* (5 495 ESTs) and *R. lucieae* (1 849 ESTs) were retrieved in Fasta format from the NCBI database (www.ncbi.nlm.nih.gov, 1st May 2012). Most of the ESTs present in NCBI database are not well processed and contains poly (A/T) signals, short reads and ambiguous sequence calls (Ns). Therefore, cl EST sequence cleaning was done using the script trim_fasta.pl written by Jennifer Meneghin (http://alrlab.research.pdx.edu). For the EST cleaning, sequences less than 50 nucleotide length were removed and Poly (A/T) sequence greater than eight, present at the 5' or 3' end were trimmed. Moreover, sequence fragments other than the ESTs available in NCBI database for genus *Rosa* were retrieved in Fasta format. These sequences mostly contained genomic survey sequences and cDNA sequences. Therefore, cleaning process as like ESTs was also performed for these sequences.

EST database is a primary database which gathered sequences from several independent experiments. Therefore, several copies of same EST sequence may be present. Hence, to make it unique non-redundant, ESTs were assembled using SeqMan Pro (version 7.1.0) sequence assembly tool available in DNASTAR software package (DNASTAR Inc., Madison, WI). The ESTs from *R. chinensis, R.* × *hybrida* and *R. lucieae* and batch of other non-EST sequences assembled independently to form primary sequence contigs.

A script MIcroSAtelitte (MISA) written in perl language was used to identify microsatellites in all primary sequence contigs (http://pgrc.ipk-gatersleben.de/misa). To claim the presence of microsatellite, only one to six nucleotides motifs were considered. The minimum repeat unit for each microsatellite was defined as 14 for mono-, six for di-, and five for tri-, tetra-, penta- and hexa-nucleotides. Compound SSRs were defined when two SSRs in the same sequence contig was interrupted by 100 nucleotide bases. All primary contigs from ESTs and other sequences were re-assembled with SeqMan tool to form non-redundant (nr) sequence data for entire *Rosa* genus. Further, the nr data were subjected to microsatellite identification with MISA.

A previously demonstrated pipeline of primer designing for SSRs were used for nr sequence data of *Rosa* (Sonah *et al.* 2010). In this pipeline, positional information of SSR generated by MISA was used for designing repeats flanking primers. All the primer pairs were designed from the flanking sequences of the Simple Sequence Repeats (SSRs) using primer3_core (www.broadinstitute.org/genome_software/other/primer3.html). For batch mode operation of primer3_core software, two perl scripts p3_in.pl and p3_out.pl were used as interface modules for the program-to-program data interchange between MISA and the primer designing software Primer3. The optimal primer designing parameters, viz. 60ºC annealing temperature, 20 bp primer length and 50% GC content, were kept to ensure 100-280 bp amplicon size.

RESULTS AND DISCUSSION

ESTs and assembled unique contigs in Rosa species

A total of 9 136 ESTs for genus *Rosa* was available in NCBI. Although, hundreds of species belongs to genus *Rosa* but only three species have EST sequence resources. Highest number of ESTs (5 495) was available for *R.* × *hybrida* that comprised 2 712 Kbp sequence length followed by *R. lucieae* and *R. chinensis* having 1 849 and 1 792 ESTs that comprised 742 and 1 060 Kbp of sequence length, respectively (Table 1). However, the set of EST sequences was redundant and having duplicated or overlapping sequences. Therefore, the ESTs were assembled in contigs for meaningful representation of data, to reduce ambiguity and to increase collective length of single stretch of a sequence.

A total of 830 unique sequence contigs were formed from the 1 792 *R. chinensis* ESTs with an average 593.4 nucleotides, ranges from 103 to 1 461 nucleotide length. For *R.* × *hybrida*, 3 210 unique sequence contigs were

Table 1 Details of contigs assembled from the expressed sequence tags and other nucleotide sequences for plant genus *Rosa* using SeqMan assembly programme

Nucleotide data	Sequence		Contig			
	Total	Length	Total	Average length	Minimum	Maximum
R. chinensis ESTs	1 792	1 060 530	830	593.4	103	1 461
R.× hybrida ESTs	5 495	2 712 783	3 210	544.9	101	2 640
R. lucieae ESTs	1 849	742 032	1 232	430.8	100	1 362
Rosa nucleotides	2 286	2 169 484	365	1,672.9	161	265 481
Super contig (nr)	5 637	3 385 567	4 643	631.0	100	265 501

formed from 5 495 ESTs with an average of 544.9 nucleotide lengths. In *R. lucieae*, 1 849 ESTs were assembled in 1 232 contigs with an average length of 430.8 nucleotides per contig (Table 1). Assembling of ESTs is crucial step for their effective utilisation in functional annotation, gene discovery and marker development. Designing primer for ESTs shorter than 100 nucleotides is difficult even if containing SSRs. Therefore, assembled longer ESTs are helpful for the EST-SSR marker development. Previously, several EST assembling programmes have been used that mostly included CAP3 program and Seqman (Yu and Li 2008). SeqMan provides rapid assembling with minimum computational resources. Moreover, it facilitated removal of probable contamination of vector sequences prior to assembling. Detection of SNPs and variation present among the assembled sequences is possible in SeqMan, however, the number of ESTs available for *Rosa* species was not enough to identify significant SNPs.

Availability of non-EST sequence resource for Rosa species

A total of 2 286 nucleotide sequences other than the ESTs, representing about 45 species belonging to genus *Rosa* were retrieved from NCBI. The sequence comprises 2 169 Kbp sequence length and these sequences included mostly partial coding DNA sequences (CDS), gene sequences and sequenced genetic loci. The longest sequence with 265 477 nucleotides present in *Rosa* was black spot resistance *muRdr1* gene locus and gene has been sequenced from *Rosa multiflora* breeding line 88/124-46 (gi 332330338, gb HQ455834.1).This is followed by a gene involved in four season-flowering of angiosperm was having 8 925 nucleotides length (gi 119848698, dbj DD367100.1). Moreover, some longer CDS of genes were also present for instance a *SOC1*-like protein (*SOC1*) gene with complete CDS of length 8,200 bp sequenced from *R. × hybrida* cultivar is present in NCBI (gi 346214852, gbJF806633.1) and among these, most of the genes have been sequenced in many genotypes. Therefore, assembling of sequence data was performed to make non-redundant sequence and was subsequently used for the identification of microsatellite and for formation of super contigs along with EST contigs.

Distribution of microsatellites in Rosa ESTs

A total of 668 SSRs in 5 272 EST contigs representing *R. chinensis*, *R. × hybrida* and *R. lucieae* ESTs were identified. The frequency of ESTs containing SSR was near about 11% in all three species (Table 2). Moreover, there was no significant difference in SSRs identified per Kbp of sequence observed among the species. Overall a SSR per 4 Kbp sequence were present in *Rosa* species which suggests that *Rosa* species have relatively higher number of ESTs containing SSRs with higher frequency per Kbp of sequence. Previously, wide range of the frequency of SSR in plant species have been observed for instance in *Papaver somniferum* (4.5%), *Phaseolus vulgaris* (10%), *Coptis japonica* (10.8%) as reported by Tripathi *et al.* (2008). However, the EST data is not normalised and depends on the type of experiment conducted to generate ESTs.

Table 2 Microsatellites identified in EST sequences of three *Rosa* species

Species	ESTs contigs	Total SSRs	Contigs with SSR	SSR per Kb	Compound SSR
R. chinensis	830	105	91	4 681	9
R. × hybrida	3 210	421	359	4 151	35
R. lucieae	1 232	142	129	3 736	5
Total	5 272	668	579	4 189	49

Frequency of microsatellite motifs in Rosa ESTs

Di-nucleotide repeats were observed as the most frequent microsatellite motif type in all the three *Rosa* species (Fig 1). The highest 55.2 percentage of di-nucleotide was present in *R. chinensis* followed by *R. × hybrida* (50.1%) and *R. lucieae* (44.3%). Next frequent class of SSR motif was tri-nucleotides and the highest frequency of tri-nucleotide repeats was observed in *R. lucieae* (42.3%)

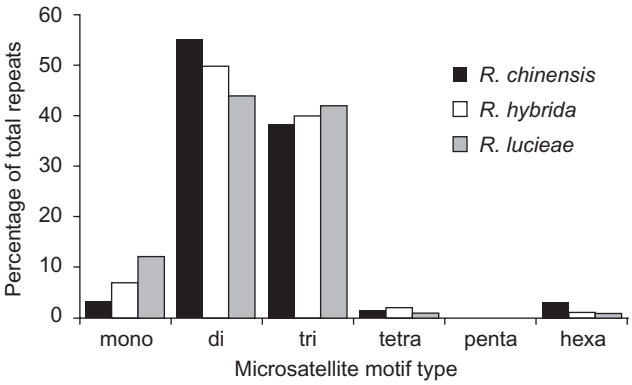

Fig 1 Frequency of microsatellite motif types identified in non redundant expressed sequence tags (ESTs) sequences of three *Rosa* species

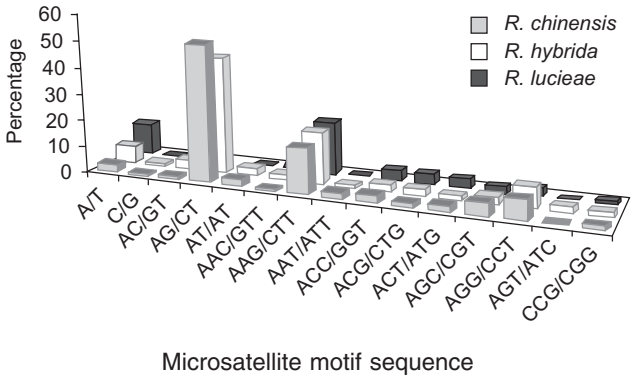

Fig 2 Frequency of microsatellite with particular motif sequence identified in EST sequences of three *Rosa* species. Percentages were calculated on the basis of total microsatellite identified with motif length 1-6 bases

followed by *R.* × *hybrida* (40.1%) and *R. chinensis* (38.1%). Compare to the di- and tri-nucleotide repeats, other motif types were rarely present in *Rosa* ESTs, therefore, comparative analysis of these repeats was not performed.

While analysing motif of particular sequence, microsatellite with AG/CT repeats were identified as most frequent class in all the three *Rosa* species (Fig 2) followed by tri-nucleotide repeats with AAG/CTT sequence. The frequency of these two motifs was similar to the previously observed frequency in the coding DNA sequences of *Brachypodium distachyon*, *Sorghum bicolor*, *Oryza sativa*, *Arabidopsis thaliana*, *Medicago truncatula*, and *Populus trichocarpa* (Sonah *et al.* 2011). Previous report suggests that the repeat AG/CT and AAG/CTT are favoured by both monocot as well as dicot species. However, CG/CG repeats have been observed only in monocots (Sonah *et al*. 2011). Likewise, in the present study no repeats with CG/CG sequence were observed in all three *Rosa* species. Whereas, the number of tetra-, penta- and hexa-nucleotide repeats identified were much smaller to compared frequency on the basis of sequence. This is similar with previous studies in which very low frequency of these repeats has been reported (Bhati *et al*. 2010, Sonah *et al.* 2011).

Frequency of microsatellites in non-EST sequences

A total of 130 microsatellites includes 10 compounds SSR were identified in 365 unique sequence contigs of assembled *Rosa* non-EST nucleotide sequences. Microsatellites were identified only in 69 out of 365 contigs. In these sequences, frequency of di-nucleotide repeats was much higher than the frequency observed in EST sequences. However, repeats with AG/CT sequence were observed with the highest (39.2%) frequency as found in ESTs (Fig 3) followed by repeats with AC/GT motif which is identified as major class with 19.2 percentage of total microsatellites. The non-EST sequences included genetic loci, promoters, introns and un-transcribed sequences along with the coding genes sequences, this might be the probable reason for the much higher frequency of mono and di-nucleotide repeats in non-EST sequences compared to the EST sequences.

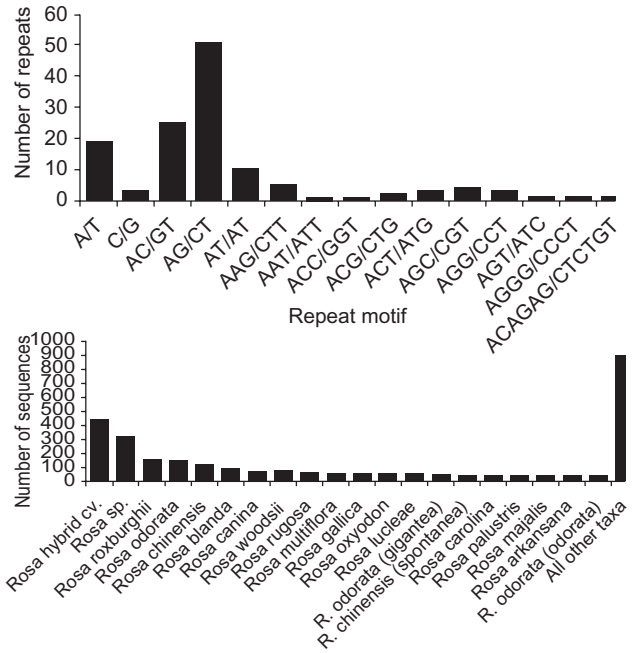

Fig 3 (a) Frequency of microsatellites with different motifs identified in (b) nucleotide sequences for different *Rosa* retrieved from NCBI (www.ncbi.nlm.nih.gov)

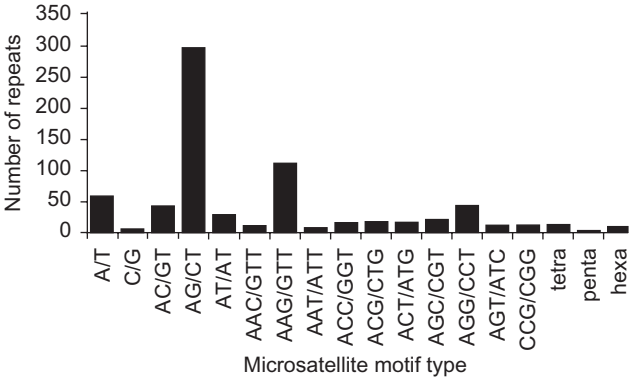

Fig 4 Frequency of microsatellite motif types identified in non redundant nucleotide sequence data including ESTs and other nucleotide sequences from plant genus *Rosa*

Microsatellite marker database for Rosa species

A local database of non redundant *Rosa* sequence were developed and subsequently used for microsatellite marker development. All the unique EST and non-EST sequence contigs were assembled into 4 643 super contigs (nr sequence data). These nr sequence comprised total of 29 28 933 nucleotide length, in which 681 SSRs were identified (Fig 4). The average length of nr contigs was 631 that facilitate designing of PCR primers in flanking region. Finally, a set of 293 primers was successfully designed (supplementary Table 1). A total of three primer pairs for each SSR were designed to provide alternative in case amplification with first set is unsuccessful. The primers were designed using the previously demonstrated pipeline with same optional settings and demonstrated successful amplification of primers with PCR in wet lab experiment (Sonah *et al.* 2011). The set of 293 EST-SSR markers developed in the

present study will be helpful for the gene mapping and tagging experiments, development of genetic linkage map, marker-assisted breeding in *Rosa* species. Moreover, EST-SSR markers are also useful to understand population structure and evaluation of genetic diversity which is pre-requisite for the effective utilization of available genetic resources in *Rosa* species.

REFERENCES

Benemann D P, Machado L N, Arge L W P , Bianchi V J, Oliveira A C, Maia L C and Peters J A. 2012. Identification, characterization and validation of SSR markers from the gerbera EST database. *Plant Omics Journal* **5**(2):159–66.

Bhati J, Sonah H, Jhang T, Singh N K and Sharma T R. 2010. Comparative analysis and EST mining reveals high degree of conservation among five Brassicaceae species. *Comparative and Functional Genomics:* 5202–38.

Buckler E S and Thornsberry J M. 2002. Plant molecular diversity and applications to genomics. *Current Opinion in Plant Biology* **5**(2): 107–11.

Channamallikarjuna V, Sonah H, Prasad M, Rao G J N, Chand S, Upreti H C, Singh N K and Sharma T R. 2010. Identification of major quantitative trait loci qSBR11-1 for sheath blight resistance in rice. *Molecular Breeding* **25**:155–66.

Deshmukh R K, Singh A, Jain N, Anand S, Gacche R N, Singh A K, Gaikawad K, Sharma T R, Mohapatra T and Singh N K. 2010. Identification of candidate genes for grain number in rice (*Oryza sativa* L.). *Functional and Integrative Genomics* **10**: 339–47.

Ellis J R and Burke J M. 2007. EST-SSRs as a resource for population genetic analyses. *Heredity* **99**: 125–32.

Foote M. 1997. The evolution of morphological diversity. *Annual Review of Ecology and Systematics* **28**: 129–52.

Gudin S. 2010. Rose: Genetics and breeding. *Plant Breeding Reviews.* Janick J (Ed). John Wiley and Sons, Inc., Oxford, UK.

Kashyap P, Singh A K, Singh S K and Deshmukh R. 2010. Genetic diversity analysis of indigenous and exotic apple genotypes using inter simple sequence repeat markers. *Indian Journal of Horticulture* **67**:15–20.

Kulkarni K P, Kulkarni S S, Gedda M, Bandevar M, Sonah H, Gacche R N, Deshmukh N K and Deshmukh R K. 2012. *In-silico* identification of rice gene homologues in brachypodium, sorghum and maize: insight into development of gene specific markers. *Webmed Central Bioinformatics* **3**(4): Wmc003258.

McCouch S R, Chen X, Panaud O, Temnykh S, Xu Y, Cho Y G, Huang N, Ishii T and Blair M. 1997. Microsatellite marker development, mapping and applications in rice genetics and breeding. *Plant Molecular Biology* **35**(1-2): 89–99.

Namita, Singh K P, Bharadwaj C, Sharma T R, Sonah H, Raju D V S and Deshmukh R K. 2011. Gene action and combining ability analysis for flower yield and its component traits in interspecific hybrids of marigold (*Tagetes* spp.). *Indian Journal of Agricultural Sciences* **81**(9): 807–11.

Nybom H. 2004. Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Molecular Ecology* **13**(5):1 143–55.

Panwar S, Singh K P, Namita and Sonah H. 2010. Genetic divergence analysis in rose (*Rosa×hybrida*) using morphological markers. *Journal of Ornamental Horticulture* **13**: 122–6.

Rout G R, Samantaray S, Mottley J and Das P. 1999. Biotechnology of the rose: A review of recent progress. *Scientia Horticulturae* **81**: 201–28.

Sharma J, Singh A, Pallavi J K, Sonah H and Gupta P. 2010. Assessment of genetic relationship among bread wheat (*Triticum aestivum* L. em. Thell.) genotypes using microsatellite markers. *International Journal of Applied Agricultural Research* **5**: 575–82.

Shirasawa K, Ishii, K, Kim C, Ban T, Suzuki M, Ito T, Muranaka T, Kobayashi, M, Nagata N, Isobe S and Tabata S. 2013. Development of capsicum EST–SSR markers for species identification and *in silico* mapping onto the tomato genome sequence. *Molecular Breeding* **31**:101–10.

Sonah H, Deshmukh R K, Sharma A, Singh V P, Gupta D K, Gacche R, Rana J C, Singh N K and Sharma T R. 2011. Genome-wide distribution and organization of microsatellite in plants: An insight of microsatellite based marker development in *Brachypodium. PLoSONE* **6**(6): e21298.

Sonah H, Deshmukh R K, Singh V P, Gupta D K, Singh N K and Sharma T R. 2011b. Genomic resources in horticultural crops: Status, utility and challenges. *Biotechnology Advances* **29**:199–209.

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 Bioinforma* **9**:78–91.

Yu H and Li Q. 2008. Exploiting EST databases for the development and characterization of EST–SSRs in the pacific oyster (*Crassostrea gigas*). *Journal of Heredity* **99**(2): 208–14.

Zhang M , Fan L, Liu Q, Song Y, Wei S, Zhang Sand Wu J. 2014. A novel set of EST-derived SSR markers for pear and cross-species transferability in Rosaceae. *Plant Molecular Biology Reporter* **32**: 290–302.

Zhu Y, Queller D C and Strassmann J E. 2000. A phylogenetic perspective on sequence evolution in microsatellite loci. *Journal of Molecular Evolution* **50**(4): 324–38.