The Ecoscan 11(1&2): 045-049, 2017
AN INTERNATIONAL QUARTERLY JOURNAL OF ENVIRONMENTAL SCIENCES
www.theecoscan.com

EST LINKED MARKERS IN *SOLANUM MELONGENA* RESPONSIVE TO ABIOTIC STRESS USING COMPUTATIONAL METHODS

REENA ROSY THOMAS[1]*, M. K. CHANDRA PRAKASH[1], M. KRISHNA REDDY[2], RIAZ MAHMOOD[3] AND PAPIYA MONDAL[1]

[1]Division of Social Sciences, Indian Institute of Horticultural Research, Bengaluru, Karnataka - 560 089, INDIA
[2]Division of Plant Pathology, Indian Institute of Horticultural Research, Bengaluru, Karnataka - 560 089, INDIA
[3]Department of Biotechnology and Bioinformatics, Kuvempu University, Shivamogga, Karnataka - 577 451, INDIA
e-mail: reenart@hotmail.com

ABSTRACT

Plants are exposed to a wide range of stresses and they respond to these stressed conditions by the activation of stress responsive genes. Microsatellites occur at thousands of locations in expressed regions, additionally, they have a higher mutation rate than other areas of DNA leading to genetic markers. A computational method was developed to capture exact repeat motifs in EST sequences of *Solanum melongena* and extract the sequences. This mathematical algorithm based repeat detection program searches for perfect repeats in the EST sequences without the need to specify the pre-defined repeat pattern for di, tri, tetra-, upto octa repeat with the exception of mono nucleotides. Hexa repeat motifs were abundantly available followed by tetra and di repeats. The identified EST derived markers with repeat motif can be used to identify the precise inheritance pattern of the expressed region. The redundant ESTs were filtered and 121 unique ESTs in *S. melongena* were found to have potential markers. These were spanned across 19 major proteins superfamilies having conserved domains associated with abiotic stress tolerance. These robust and potential EST derived markers are linked to genes that encode for specific traits providing tolerance against abiotic stresses which is useful in marker-assisted breeding programs.

KEY WORDS
Abiotic stress
Computational methods
EST sequences
Markers, *Solanum melongena*

Received : 00.00.2016
Revised : 00.00.2017
Accepted : 00.00.2017

***Corresponding author**

INTRODUCTION

Abiotic stress factors such as drought, elevated temperatures and high salinity affect plant growth which in turn results in reduced yield (Cushman and Bohnert, 2000). Current advances in genomic technologies provide effective methods for identifying stress-related genes. Brinjal (*Solanum melongena* L.) is an important solanaceous crop of sub-tropics and tropics and one of most common vegetable crops grown extensively throughout the year. A substantial proportion of brinjal yield is lost due to various stresses as they are sensitive to high heat, cold temperatures and water deficit. It is predicted that there could be more number of genes in brinjal genome in comparison with any other solanaceae crops (Hirakawa *et al*., 2014). Mining of EST sequence data for the presence of microsatellites is a productive way of identifying gene-associated markers (Tobias *et al*, 2005). EST derived markers are from the coding regions of the genome, they are genetically associated with a trait of interest (Yu *et al*., 2004). These markers are expected to have greater transferability between species (Scott *et al*., 2000; Thiel *et al*., 2003) since gene-coding regions are more likely to be conserved among related species. About 2-5% of the ESTs in different plant species are reported to contain simple sequence repeats (SSR) suitable for marker development (Kantety *et al*., 2002) which will be a valuable resource for tagging and mapping of genes related to agronomic and stress-resistant traits of interest (Zhang *et al*., 2014).

The markers tightly linked with gene are used for selection of parental lines for high selection efficiency (Yunbi, 2010) to develop crops conferring resistance to abiotic stresses (Panigrahi *et al*., 2013) and overcome the adverse situations in present climate changing scenario. The objective of the present study focus on finding the repeat motif in EST collections of *Solanum melongena* for identifying marker that points to a trait specific gene of interest for abiotic stress tolerance using computational methods.

MATERIALS AND METHODS

A total of 101270 EST sequences of *Solanum melongena* available in Sol Genomics Network (SGN) were used in this research study. Among these comprehensive collections, the sequences having repeat motifs were identified using the computer program developed under this work. Though a number of computational tools are available which either directly or indirectly detect repeats in the sequences, there are significant limitations such as inability to read large sequence files or multiple sequences at a time and to extract those sequences.

In this context, a computational method was developed to capture exact repeat motifs in EST sequences and extract the sequence. The repeat detection program, is a mathematical algorithm based (non-probabilistic) program for finding tandem

repeats specifically for perfect repeats in the EST sequences without the need to specify the pre-defined repeat pattern.

Computer program

The algorithm detects all significant repeat motifs where significance is assessed based on the exact match which uses one of the algorithms of Tandem Repeats Analyzer (TRA) described in Bilgen *et al*. (2004). The program overrides the inability of reading large sequence file of multiple EST sequences, which analyses the whole data set provided in a data folder (Karaca *et al.*, 2005). The program contains two modules: (i) locating exact motif in a sequence, (ii) extraction of EST sequences flanking the exact motif. The fist module searches for exact motifs, a string of repeated units, with all possible combinations of motif occurrences (4^m - 4)in the EST collections, where 'm' is the repeat motif length and 4 being the four bases (ATCG) of nucleotides. The program locates all the possible combinations of di, tri, tetra, etc. repeats by encapsulating all the probable motifs upto 8 nucleotides long in the complete EST collections with the exception of mono nucleotides. It searches for S_n, a string of repeated units in a EST sequence. $S_n = E_n [i_n, j_n]$ symbolizes the S_n starting with the i_n^{th} and ending with the j_n^{th} bases of the EST sequence (E_n). The distance between i_n and j_n will therefore be equal to $m_n \times r_n$ where m_n and r_n refer to a type of repeat motif length and the number of repeats in S_n string of each E of a fixed length, respectively. The second module finds the position of the repeat motif and locates the start and end coordinates of the EST sequence and extracts those sequences where the exact repeat motifs are found.

The identified ESTs are concatenated with a blank line in between and generated as FASTA file for batch processing. Further, Tandem Repeats Database were used which is a public repository of information on tandem repeats in DNA.to cross compare the identified EST sequences for the presence of repeat motifs. Based on the output result, very short and redundant ESTs were discarded and only non-redundant ESTs, having repeat sequences were selected.

The concatenated EST sequences were then explored using *in silico* approaches to detect the presence of conserved protein domains across several species associated with abiotic stress tolerance coming from different public resources and previous studies. The obtained sequences were looked for exclusive conserved protein domains in the NCBI conserved Domain Database (Marchler-Bauer *et al*., 2000) that were more relevant to abiotic stress and were classified that are evolutionarily conserved and belongs to their respective protein superfamily.

RESULTS AND DISCUSSION

From the results obtained from computer program by exploring the EST collections of *S. melongena,* hexa motif was dominant (35%) followed by tri, penta, tetra, etc. among the identified motifs (Fig. 1). The repeat motifs identified in brinjal were more, the reason being, the predicted genes in brinjal is very high though the genome size of brinjal is smaller among solanaceae crops (1.127 Gb).

The graphical representation of potential ESTs identified by the program with perfect repeat motifs is shown in Fig.2.
From the results obtained from Conserved Domain Database search, the ESTs showing significant similarity to known stress tolerant proteins are given in Fig. 3. The redundant ESTs were filtered and 121 unique ESTs in *S. melongena* were found to have potential markers which were spanned across 19 major proteins superfamilies responsive to abiotic stress.

***Solanum melongena* ESTs and its protein superfamilies association**

The graphical representation of ESTs having potential markers and its superfamily association is depicted (Fig. 3) using CIRCOS diagram (Krzywinski *et al*., 2009). The relationships between ESTs and superfamilies and vice versa are indicated by colour bands according to their volumes. ESTs are shown anticlockwise from top while protein superfamilies are shown clockwise from top. Among the 121 identified EST sequences of *S. melongena*, majority of them are associated with H2A, HSP90, MFS, MIP, HMG-box, S1-like and Alpha Crystallin

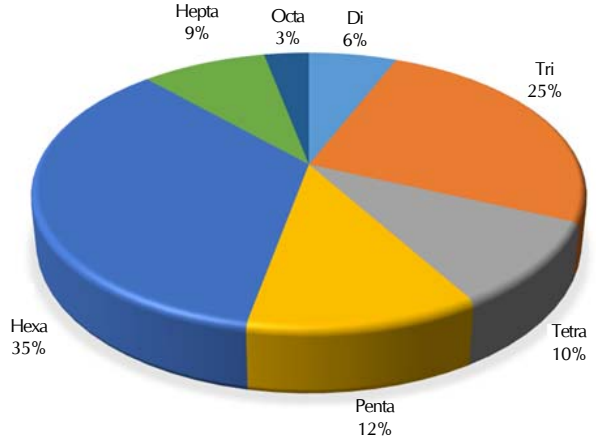

Figure 1: The distribution of di, tri, tetra, penta, hexa, hepta and octa repeat motifs in ESTs of brinjal.

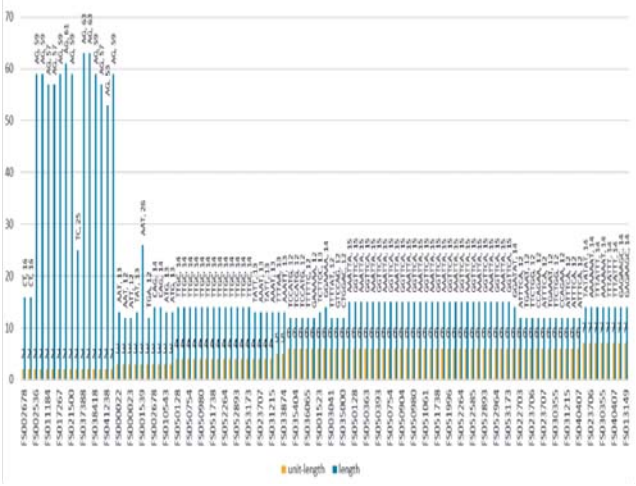

Figure 2: The above graph represents potential ESTs in brinjal with perfect repeat motifs

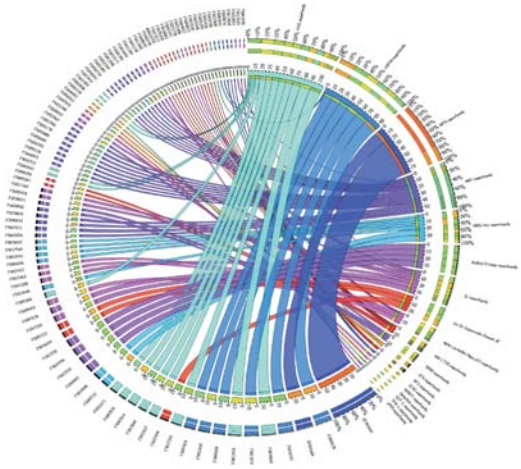

Figure 3: The ideogram representing *S. melongena* ESTs having potential markers and its respective superfamilies implicated in stress tolerance

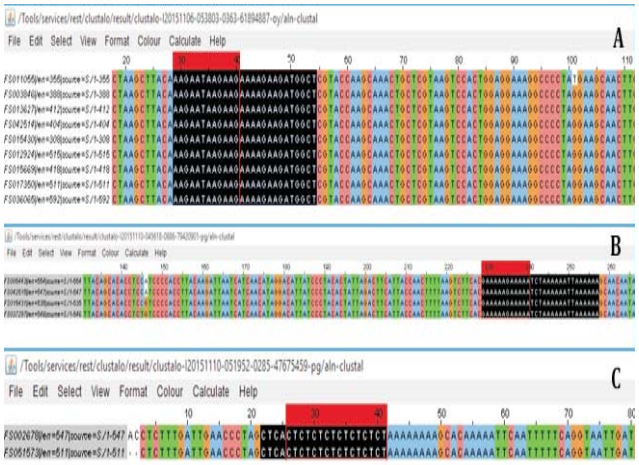

Figure 4: The graphical representation of MSA alignment of EST sequence with H2A markers in conserved region highlighted in black colour and the repeat motif in red colour

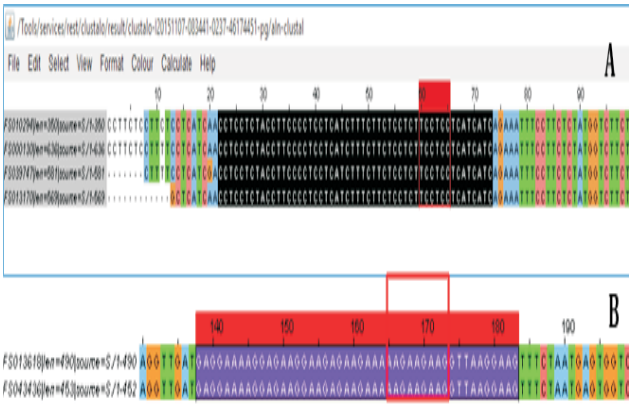

Figure 5: The graphical representation of MSA alignment of EST sequence with HSP90 markers in conserved region highlighted in black colour and the repeat motif in red colour

Figure 6: The graphical representation of MSA alignment of EST sequence with MIP markers in conserved region highlighted in black colour and the repeat motif in red colour

Table 1: The EST id and tri-repeat motif (TGT) along with MFS marker sequence

S. No	EST ID	Repeat Motif	Marker
1	EF091664	$(TGT)_2$	TCCTTGGTGTTGTCAACTTCCTTGGTATGATGTTCACATTCTTGGT
2	EF091666	$(TGT)_3$	TTGGTTCTCATGGTTGTGTGTTCTGTTGCTTCTG

domain superfamilies. These are associated with abiotic stress tolerance and exhibits significant similarity to a large number of published proteins. The multiple sequence alignment using Clustal W2 shows that the markers are highly conserved among the identified sequences. The markers having high similarity with published proteins responsive to abiotic stress are given below.

H2A superfamily

Having histone variant H3.3 protein performs essential roles in maintaining structural integrity of the nucleosome, chromatin condensation and binding of specific chromatin-associated proteins (Talbert *et al*., 2012). Histone modifications along with DNA methylation can be correlated with gene expression in response to abiotic stresses, such as water deficit, high-salinity and temperature shifts (Kim *et al*., 2008; Luo *et al*., 2012). The identified markers belongs to H2A protein superfamily are associated with salt and drought stress is shown in Fig. 4.

Figure 4A shows the Multiple Sequence alignment (MSA) of nine sequences having a 21 bp conserved marker region with tri nucleotide motif (AAG) repeated four times with an interruption of one nucleotide. Figure 4B shows four sequences having a 29 bp of conserved marker region with perfect penta nucleotide motif (GAAAAA) which is repeated twice and Figure 4C shows another two sequences having a 20 bp of conserved region with perfect di nucleotide motif (CT) which is repeated ten times in the conserved region with an interruption of one nucleotide. The above MSA revealed highly conserved marker

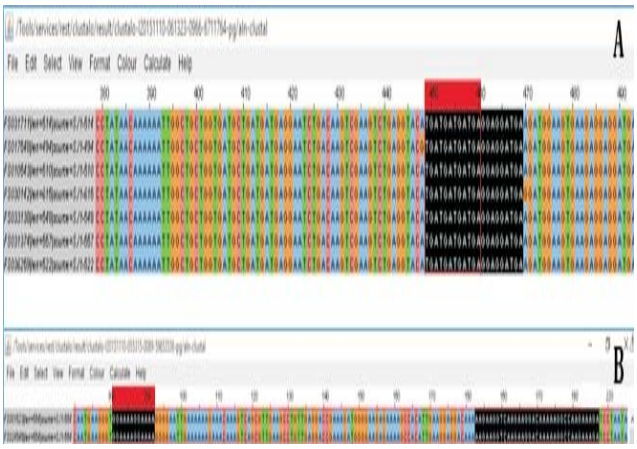

Figure 7: The graphical representation of MSA alignment of EST sequence with HMG box markers in conserved region highlighted in black colour and the repeat motif in red colour

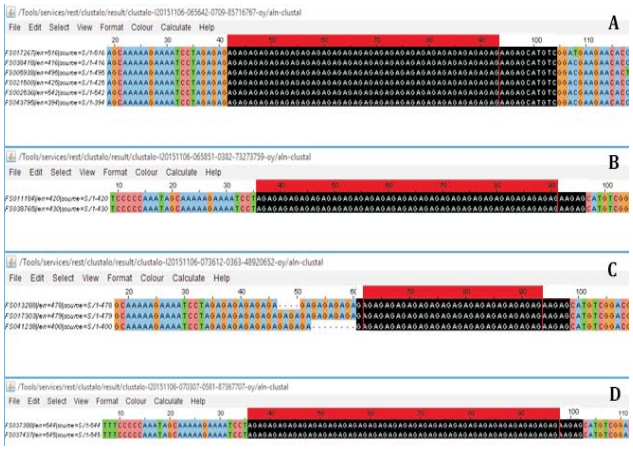

Figure 8: The graphical representation of MSA alignment of EST sequence with S1 like markers in conserved region highlighted in black colour and the repeat motif in red colour

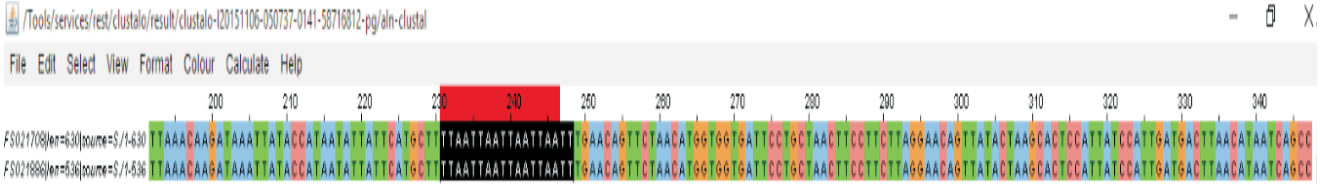

Figure 9: The graphical representation of MSA alignment of EST sequence with Alpha-crystallin markers in conserved region highlighted in black colour and the repeat motif in red colour

sequences from EST collections of *S. melongena*.

HSP90 superfamily

Having heat shock protein 90 is a chaperone protein which plays an important role in protein refolding and stabilizes proteins against heat stress. It is also involved in cell signalling, intracellular transport and preventing aggregation of the denatured proteins. These identified markers belongs to HSP90 protein superfamily are associated with heat, salt, osmotic and drought tolerance is shown in Fig. 5.

Figure 5A shows the multiple alignment of four sequences having a 52 bp conserved marker region with tri nucleotide motif (TCC) repeated twice. Figure 5B shows two sequences having a 45 bp of conserved marker region with perfect tri nucleotide motif (AAG) repeated three times.

Major Facilitator Superfamily

(MFS) is having a phosphate transporter protein that regulates stomatal movements in drought stressed condition. The phosphate transporter protein that belongs to MFS superfamily was having identities with only two EST sequences. However, the marker present in these two sequences exhibit very high similarity to a large number of published proteins. The EST ID, EF091664 had strong similarity with 57 published phosphate proteins and another EST ID, EF091666 with 19 published phosphate proteins across several crops. These markers with a repeat motif (TGT) are shown in Table 1, which are associated with Major Facilitator Superfamily.

Major intrinsic proteins (MIP) superfamily

Comprises of Aquaporin protein, which are water channel proteins that regulate the movement of water and other small molecules across plant vacuolar and plasma membranes; they are associated with plant tolerance to abiotic stresses. A total of 20 unique EST sequences exhibits a high similarity with Aquaporin proteins. The marker belongs to MIP protein superfamily associated with water stress tolerance are shown in Fig. 6.

Figure 6A shows a multiple alignment of four sequences having a 55bp conserved marker region with perfect hexa nucleotide motif (CTGGAC) repeated twice and Fig. 6B shows sixteen sequences having a 29 bp of conserved marker region with perfect hexa nucleotide motif (AAATTA) repeated twice.

HMG-box superfamily having HMG-box domains are found in high mobility group proteins, which are involved in the regulation of DNA-dependent processes such as transcription, replication and DNA repair. In *Arabidopsis*, HMGB proteins are involved in plant tolerance to different stress conditions (Pedersen and Grasser, 2010). The markers belongs to HMG box protein superfamily associated with with oxidative, cold and salt stressare shown in Fig. 7.

Figure 7A shows a multiple alignment of seven sequences having a 21 bp conserved marker region with tri nucleotide motif (TGA) repeated twice and Fig. 7B shows two sequences having a 35 bp of conserved marker region preceded with perfect hexa nucleotide motif (GGAAAA) repeated twice.

S1_like superfamily

Is found in a wide variety of RNA-associated proteins and contains RNA binding domain which also contains the Cold

Shock Domain tolerance to cold stress. Figure 8 shows that the di nucleotide motif, AG is longest motif and repeated several times in the marker sequences.

In Fig. 8A, six sequences aligned with MSA is having a 58bp conserved marker region of AG motif repeated 26 times. Fig. 8B shows a multiple alignment of two sequences having a 52 bp of marker region with AG motif repeated 28 times. Figure 8C shows that three sequences is having a 38 bp of marker region with AG motif repeated 16 times and Fig. 8D shows two sequences having a 62 bp of conserved marker region with the same AG motif repeated 31 times. This motif AG is surprisingly the longest among all repeat motif sequences in *S. melongena*. These markers belongs to S1 like protein superfamily and are linked with drought, salt, heat and cold stress.

Alpha-crystallin-Hsps_p23 Superfamily

Associated with small heat shock proteins (sHSPs) are ATP-independent chaperones that prevent aggregation at high temperatures and are important in refolding in combination with other HSPs and protect the enzyme activity during thermal stress (Jakob *et al*., 1993).

Fig. 9 shows that multiple alignment of two sequences is having a 18 bp conserved marker region with tetra nucleotide motif (TTAA) repeated four times.

Further, four more sequences viz., FS002297, FS009429, FS023416 and FS023417 were having repeat motifs $(CT)_{7,}$ $(AG)_{4,}$ $(TAT)_{6}$ and $(GAGGA)_{2}$ in their respective EST ids. These markers are associated with heat stress that belongs to Alpha-crystallin protein superfamily are shown in CIRCOS ideogram (Fig. 3).

The microsatellite markers identified using computational methods are associated with functional genes being in expressed region. These markers were found to be more transferable in solanaceous plants (Haq *et al*., 2014) and highly useful for molecular breeding program in *S. melongena* for identifying parental lines for high selection efficiency and also for imparting tolerance to abiotic stress in changing climate scenario.

ACKNOWLEDGEMENT

The authors wish to thank Centre for Agricultural Bioinformatics, the PI of Network project on Agricultural Bioinformatics and Computational Biology for funding this research work, also thankful for ICAR - Indian Institute of Horticultural Research and IASRI for technical support.

REFERENCES

Bilgen, M., Karaca, M., Onus, A. N. and Ince, A. G. 2004. A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences. *Bioinformatics*. **20:** 3379-3386.

Cushman, J. C. and Bohnert, H. J. 2000. Genomic approaches to plant stress tolerance. *Current Opin. Plant Biol.* **3:** 117-124.

Haq, S. U., Jain, R., Sharma, M., Kachhwaha, S. and Kothari, S. L. 2014. Identification and Characterization of microsatellites in Expressed Sequence Tags and their cross transferability in different plants, *International J. Genomics*. pp. 1-12.

Hirakawa, H., Shirasawa K., Miyatake K., Nunome T., Negoro S., Ohyama A. and Fukuoka, H. 2014. Draft Genome Sequence of Eggplant (*Solanum melongena* L.): the Representative Solanum Species Indigenous to the Old World. *DNA Research*. **21(6):** 649-60.

Jakob, U., Gaeste, l. M., Engel, K. and Buchner, J. 1993. Small heat shock proteins are molecular chaperones. *J. Biol. Chem*. **268:** 1517-1520.

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

Karaca, M., Bilgen, M., Onus, A. N., Ince, A. G. and Elmasulu, S. Y. 2005. Exact Tandem Repeats Analyzer (E-TRA): A new program for DNA sequence mining. *J. Genet.* **84:** 49-54.

Kim, J. M., To, T. K., Ishida, J., Morosawa, T., Kawashima, M. and Matsui, A. 2008. Alterations of lysine modifications on the histone H3 N-tail under drought stress conditions in *Arabidopsis thaliana*. *Plant Cell Physiol.* **49:** 1580-1588.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. and Marra, M. A. 2009. Circos: an Information Aesthetic for Comparative Genomics. *Genome Res.* **19:** 1639-1645.

Luo, M., Liu, X., Singh, P., Cui, Y., Zimmerli, L. and Wu, K. 2012. Chromatin modifications and remodelling in plant abiotic stress responses. *Biochem. Biophys. Acta.* **1819:** 129-136.

Marchler–Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y. and Bryant, S. H. 2000. CDD: a database of conserved domain alignments with links to domain three dimensional structure. *Nucleic Acids Res*. **30:** 281-283.

Panigrahi, J., Mishra, R. R., Sahu, A. R., Rath, S. C., Seth, S. and Mishra, S. P. 2013. Marker-assisted breeding for simple inherited traits conferring stress resistance in crop plants. *The Ecoscan*. **3:** 217-233.

Pedersen, D. S. and Grasser, K. D. 2010. The role of chromosomal HMGB proteins in plants. *Biochem. Biophys Acta*. **1799(1-2):** 171-174.

Scott, K. D., Eggler, P., Seaton, G., Rossetto, M., Ablett, E. M., Lee, L. S. and Henry, R. J. 2000. Analysis of SSRs derived from grape ESTs. *Theor. Appl. Genet.***100(5):** 723-726.

Talbert, P. B., Ahmad, K, Almouzni, G, Ausió, J, Berger, F., Bhalla, P. L., Bonner, W. M., Cande, W. Z., Chadwick, B. P., Chan, S. W., *et al*. 2012. A unified phylogeny-based nomenclature for histone variants. *Epigenetics Chromatin*. **5:** 7.

Thiel, T., Michalek, W., Varshney, R. K. and 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.

Tobias, C. M., Twigg, P., Hayden, D. M., Vogel, K. P., Mitchell, R., Lazo, G. R., Chow, E. K. and Sarath, G. 2005. Analysis of expressed sequence tags and the identification of associated short tandem repeats in switchgrass. *Theor. Appl. Genet.***111:** 956-964.

Yu, J. K., Dake, T. M., Singh, S., Benscher, D., Li, W., Gill, B. and Sorrells, M. E. 2004. Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. *Genome*. **47:** 805-818.

Yunbi, Xu 2010. Molecular Plant Breeding. Wallingford, UK, CAB International. p. 734.

Zhang, M., Mao, W., Zhang, G. and Wu, F. 2014. Development and Characterization of Polymorphic EST-SSR and Genomic SSR Markers for Tibetan Annual Wild Barley. *PLoS ONE*. **9(4):** e94881.