Genetics and Molecular Research

Transcriptome analysis of potential simple sequence repeat markers in *Ammopiptanthus mongolicus*

M. Jin[1,2], M.Y. Guo[1,2], L. Han[1], J.L. Li[3], S.Y Yang[1] and Y.H. Su[1]

[1]State Key Laboratory of Soil and Sustainable Agriculture/Institute of Soil Science Chinese Academy of Sciences, Nanjing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Forestry College, Nanjing Forestry University, Nanjing, China

Corresponding author: Y.H. Su
E-mail: yhsu@issas.ac.cn

Genet. Mol. Res. 15 (3): gmr.15038581
Received February 29, 2016
Accepted April 11, 2016
Published August 29, 2016
DOI http://dx.doi.org/10.4238/gmr.15038581

ABSTRACT. *Ammopiptanthus mongolicus*, an evergreen broadleaf legume shrub, can survive under conditions of high and low temperature, extreme salinity, and drought. This attribute makes it an ideal model for studying mechanisms of stress tolerance in plants. However, simple sequence repeat (SSR) resources for this species are insufficient in public databases. In this study, a total of 44,959 unigenes identified from the *A. mongolicus* transcriptome were used for SSR analysis by MIcroSAtellite (MISA). A total of 13,859 SSRs were found to be distributed within 10,409 unigenes, with an average length of 15 bp and an average density of one SSR per 4.4 kb. There were 222 different motif types in the *A. mongolicus* transcriptome, and mononucleotide repeats represented the main type, accounting for 44.2% of all SSRs. The (A/T)n repeat was the most frequent motif, accounting for 42.37%

of all SSRs. We also performed Gene Ontology functional analysis, Kyoto Encyclopedia of Genes and Genomes database pathway analysis, and eggNOG analysis, and identified 6157, 2301, and 9845 unigenes containing SSRs in these three databases, respectively. The functional categorization of *A. mongolicus* unigenes containing SSRs revealed that these unigenes represent many transcribed genes with different functions. These data provide sequence information that may be used to improve molecular-assisted markers for the study *A. mongolicus* genetic diversity.

Key words: *Ammopiptanthus mongolicus*; *De novo* transcriptome; SSR markers; GO functional analysis; KEGG pathway analysis; eggNOG analysis

INTRODUCTION

Ammopiptanthus mongolicus is an evergreen xerophyte of Mesquite found in the mid-Asian desert, and is an ancient relic from the Tertiary period. This species shows strong tolerance to high and low temperature, extreme salinity, and drought. It can survive under annual precipitation conditions of less than 50 mm and annual evaporation over 3000 mm (Liu, 1998). In addition, it helps to prevent sand movement, thus delaying further desertification (Ge et al., 2005). Due to its ecological importance and high academic value, *A. mongolicus* has become a good model for studying plants under abiotic stress conditions. Most studies have focused on morphological characteristics, physiological responses under abiotic stress conditions (Xu et al., 2002), and on several candidate genes involved in various stress responses (Wei et al., 2011a; Sun et al., 2013; Gu and Cheng, 2014). However, there are very few reports on the diversity of germplasm resources and molecular markers of *A. mongolicus*.

Simple sequence repeats (SSR) are arrays of short motifs that are 1-6 base pairs in length (Gupta et al., 1996). These single-locus markers are characterized by their hypervariability, abundance, reproducibility, Mendelian inheritance, and codominant nature (Scott et al., 2000). Based on the original sequences used to identify the simple repeats, SSRs can be divided into genomic SSRs and EST-SSRs. EST-SSRs are derived from expressed sequences, which are more evolutionary conserved than noncoding sequences. Therefore, EST-SSRs have a relatively high transferability compared with genomic SSRs (Wei et al., 2011b). With the development of next-generation sequencing technologies, many EST-SSRs have been found and evaluated in sweet potato (Wang et al., 2010), chickpea (Garg et al., 2011), *Epimedium sagittatum* (Zeng et al., 2010), Siberian wildrye (Zhou et al, 2016), and multiple other species. However, the results of these studies have shown that SSRs vary in different plant species.

In the present study, we utilized *A. mongolicus* transcriptome data obtained from Illumina paired-end sequencing to analyze SSRs. The aim of this study was to characterize genic markers, EST-SSRs, in order to evaluate and compare the frequency and distribution of various types of EST-SSRs in genic sequences, and to analyze the function of these unigenes containing SSRs. Our results provide a very useful genomic resource for future studies on *A. mongolicus*.

MATERIAL AND METHODS

Data on the *A. mongolicus* transcriptome was obtained from a previous study (has not been published). *A. mongolicus* seeds were soaked in water for 2 days at 26°C and then germinated on moist filter paper placed on sterile Petri dishes. Seedlings were transferred to half-strength Hoagland solution in a greenhouse at approximately 26°C with a photoperiod of 16-h light/8-h dark for 4 weeks. Next, total RNA was extracted and a paired-end cDNA library was constructed. Constructed paired-end libraries were sequenced using an Illumina HiSeq™ 2000. Following a quality check and *de novo* assembly, we obtained 44,959 unigenes.

Next, we used the MIcroSAtellite (MISA) identification tool to identify potent SSRs in all of the unigenes. The criteria for SSRs in the MISA script were mono-nucleotide repeats occurring more than 10 times, di-nucleotide repeats occurring more than six times, tri-, tetra-, penta-, and hexa-nucleotide repeats occurring more than five times.

By performing BLASTx (E value <0.00001) alignment between unigenes containing SSRs and the following databases, Nr (non-redundant protein sequences in NCBI), COG (Cluster of Orthologous Groups of proteins), and KEGG (Kyoto Encyclopedia of Genes and Genomes database), we obtained proteins with the highest similarity to the given unigenes, as well as the functional annotations. To obtain the Gene Ontology (GO) functional annotations, we used the blast2go (http://www.blast2go.com/b2ghome) (Conesa et al., 2005) and Map2Slim (http://www.geneontology.org/GO.slims.shtml#script). After aligning unigenes to the COG database, we determined the COG functional annotations. Metabolic pathway annotations were identified according to the KEGG database.

RESULTS

Frequency and distribution of SSRs in the *A. mongolicus* transcriptome

Screening 61,212,624 bp using the MISA software identified a total of 44,959 unigenes in the *A. mongolicus* transcriptome. We identified 13,859 SSRs distributed in 10,409 unigenes, accounting for 23.15% of the total unigenes. The frequency of SSRs in the *A. mongolicus* transcriptome was 30.82%, and the distribution density was 4.4 per kb. A total of 2647 unigenes contained more than one SSR and 683 SSRs were present in compound formation, accounting for 5.88 and 1.52% of the total unigenes, respectively (Table 1).

Table 1. SSRs in the *Ammopiptanthus mongolicus* transcriptome.

Items	Number
Total number of sequences examined	44,959
Total size of examined sequences (bp)	61,212,624
Total number of identified SSRs	13,859
Number of SSR containing sequences	10,409
Number of sequences containing more than one SSR	2,647
Number of SSRs present in compound formation	683

The length of the SSRs identified from the *A. mongolicus* transcriptome ranged from 10 to 25 bp, and the average length was 15 bp. As shown in Figure 1, the most frequent repeat length was 15 bp, possessed by 2295 SSRs, which was followed by 10, 18, and 12 bp, each of which possessed 2182, 1958, and 1670 SSRs, respectively. The highest repetition frequency

was 10, which was possessed by 2637 SSRs and accounted for 19.03% of the total SSRs. This was followed by 5, 7, and 11 repetitions, which occurred at frequencies between 1200 and 2400. A total of 6555 SSRs were repeated 5-8 times, accounting for 47.30% of the total; 5148 SSRs were repeated 9-12 times, accounting for 37.14% of the total; 1346 SSRs were repeated 13-16 times, accounting for 9.72% of the total; 668 SSRs were repeated 17-20 times, accounting for 4.82% of the total; and 142 SSRs were repeated 21-23 times, accounting for 1.02% of the total (Figure 2).

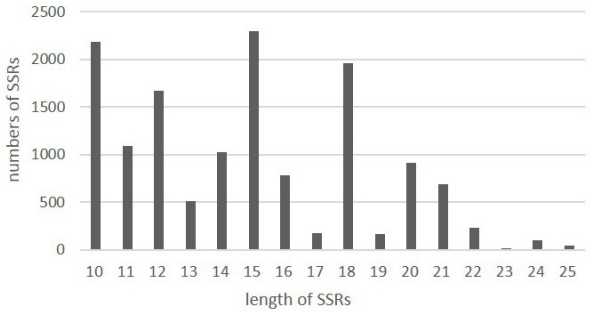

Figure 1. Length distribution of SSRs in the *Ammopiptanthus mongolicus* transcriptome.

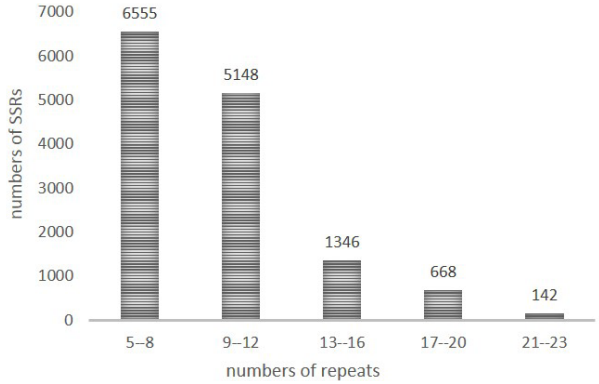

Figure 2. Distribution of the number of SSRs in the *Ammopiptanthus mongolicus* transcriptome.

Different types of SSRs in the *A. mongolicus* transcriptome

Based on the repeat motifs, all SSRs were divided into mono-, di-, tri-, tetra-, and penta-nucleotide repeats. Of all the types of repeat motifs in the *A. mongolicus* transcriptome, mono-nucleotide repeats were the most abundant (6127, 44.21%), followed by tri-nucleotide (3917, 28.26%), di-nucleotide (3403, 24.55%), tetra-nucleotide (364, 2.63%), and penta-nucleotide repeats (48, 0.35%) (Table 2). The total length of SSRs in the *A. mongolicus* transcriptome was 205,071 bp. The length of mono-, di-, tri-, tetra-, and penta-nucleotides was 76,546, 53,266, 66,591, 7468, and 1200 bp, respectively. The average length of each type of motif was 12, 16, 17, 21, and 25 bp, respectively (Table 2).

Table 2. Occurrence of SSRs in the *Ammopiptanthus mongolicus* transcriptome.

Repeat type	Number	Proportion (%)	Frequency (%)	Average distance (kb)	Total length (bp)	Average length (bp)
Mononucleotide	6,127	44.21	13.63	9.99	76,546	12
Dinucleotide	3,403	24.55	7.57	17.98	53,266	16
Trinucleotide	3,917	28.26	8.71	15.63	66,591	17
Tetranucleotide	364	2.63	0.81	16.82	7,468	21
Pentanucleotide	48	0.35	0.11	12.75	1,200	25
Total	13,859	100	30.82	4.4	205,071	15

We found 222 types of repeat units in 13,859 SSR from the *A. mongolicus* transcriptome. Mono-, di-, tri-, tetra-, and penta-nucleotide repeats accounted for 4, 12, 60, 109, and 37 types, respectively. Among these repeat units, (A/T)n was the dominant (42.37%), followed by (AG/CT)n (15.51%), (AAG/CTT)n (7.58%), and (AT/AT)n (5.21%) (Table 3). The distribution of tetra- and penta-nucleotides was scattered and 146 types of repeat units occupied only 2.73% of all SSRs. Furthermore, very few CG/CG (0.02%) repeats were identified in the database.

Table 3. Type and number of SSRs in the *Ammopiptanthus mongolicus* transcriptome.

Repeat motif	Repeat numbers					Total	Percentage (%)
	5	6	7	8	>8		
A/T	-	-	-	-	5872	5,872	42.37
C/G	-	-	-	-	255	255	1.84
AC/GT	-	178	119	69	162	528	3.81
AG/CT	-	589	402	428	731	2,150	15.51
AT/AT	-	205	132	101	284	722	5.21
CG/CG	-	3	-	-	-	3	0.02
AAC/GTT	258	188	81	4	-	531	3.83
AAG/CTT	495	387	164	5	-	1,051	7.58
AAT/ATT	212	172	122	4	-	510	3.68
ACC/GGT	175	97	35	5	-	312	2.25
ACG/CGT	35	13	4	1	-	53	0.38
ACT/AGT	59	34	27	7	-	127	0.92
AGC/CTG	178	79	61	6	-	324	2.34
AGG/CCT	209	92	36	5	-	342	2.47
ATC/ATG	297	148	80	5	-	530	3.82
CCG/CGG	92	36	7	2	-	137	0.99
Total	2,010	2,221	1,270	642	7,304	13,447	
Percentage (%)	14.50	16.03	9.16	4.63	52.69		97.03

Functional analysis of unigenes containing SSRs

By performing BLAST analysis of the 10,409 unigenes containing SSRs against the Nr, KEGG, and COG databases, we obtained 10,050 (96.55%) unigenes with homologous sequences in at least one of the above databases. Among them, 6157, 9845, and 2301 unigenes were found in the Nr, COG, and KEGG database, respectively. Overall, 1763 unigenes were found in all three databases, while 359 were not identified in any (Figure 3).

Based on Nr annotations, the GO classification system was used to classify possible functions of the unigenes. The unigenes were then classified into three main categories: biological processes, cellular components, and molecular function. For biological processes, the largest category was "metabolic process" (4237, 18.44%), followed by "cellular process" (3443, 14.99%). For cellular component, the top four categories were as follows: "cell" (2093, 9.44%), "intracellular" (1930, 8.40%), "membrane" (1239, 5.39%), and "cytoplasm" (997, 4.34%). For molecular function, the "binding" (3734, 16.25%) category was the most prominent (Figure 4).

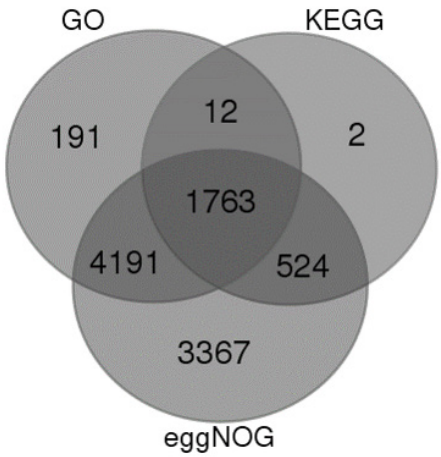

Figure 3. Numbers of unigenes containing SSRs blasted to the Nr, KEGG, and COG databases.

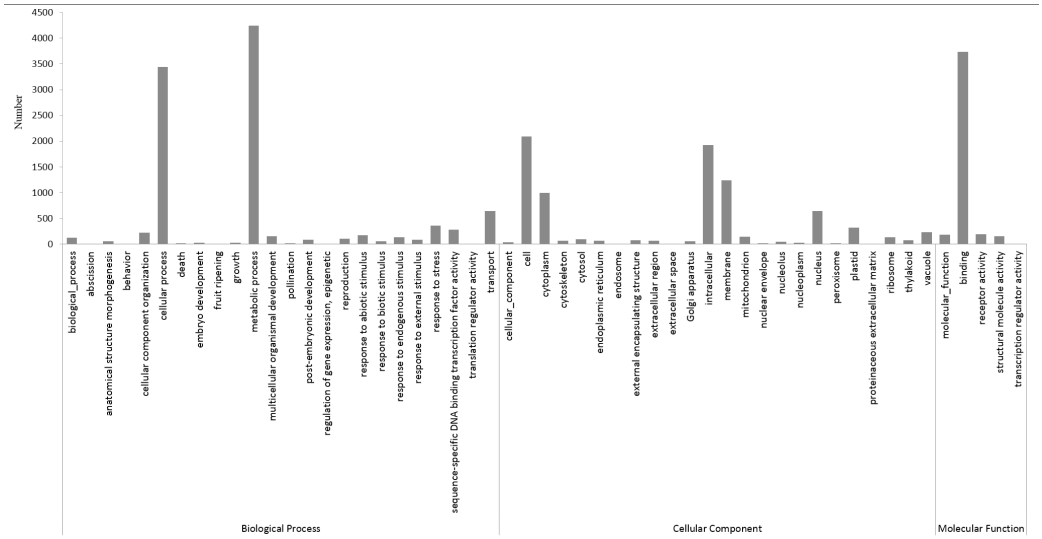

Figure 4. GO analysis of 6157 unigenes containing SSRs.

To better understand the characteristics of functional distribution of genes in *A. mongolicus*, we used eggNOG annotation, which is based on the COG database. A total of 6690 unigenes were classified into 26 groups according to their possible functions. The category "General function prediction only" accounted for a large part (1221, 18.25%), followed by "Signal transduction mechanisms" (633, 9.46%), "Posttranslational modification, protein turnover, chaperones" (522, 7.80%), and "Transcription" (453, 6.77%). In addition, 2012 (30.07%) unigenes were annotated as "Function unknown" (Figure 5).

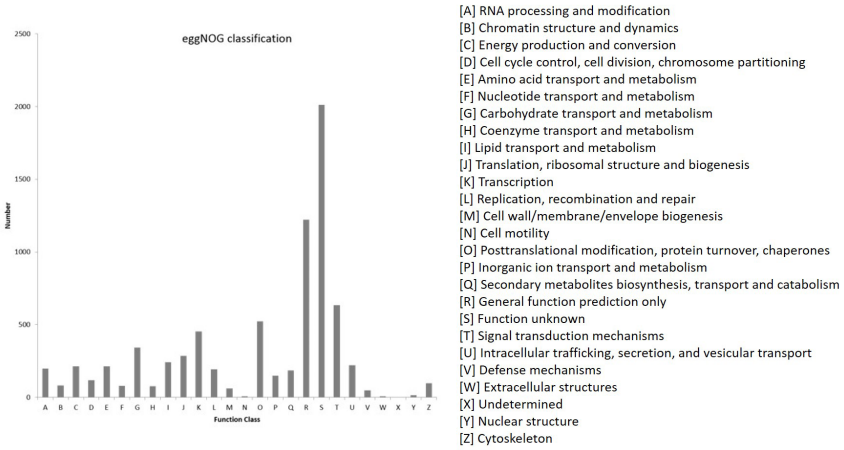

Figure 5. eggNOG analysis of 9845 unigenes containing SSRs.

DISCUSSION

A. mongolicus is the only genus with an evergreen broadleaf habit in the desert of eastern central Asia. Because of the ecological importance and the high academic value of *A. mongolicus*, some studies have been conducted on this species. However, there are only a few reports related to SSRs of *A. mongolicus*. Chen, et al. (2007) developed 11 polymorphic microsatellite markers for *A. mongolicus* using the fast isolation by amplified fragment length polymorphism of sequences containing repeats protocol, and Zhou et al. (2012) identified 1827 SSRs using 454 pyrosequencing, which is markedly lower than the 13,859 SSRs identified utilizing Illumina Hiseq2000™ sequencing in the present study. The 13,859 SSRs detected in the *A. mongolicus* transcriptome were distributed in 10,409 unigenes and the distribution density was 4.4 per kb. These SSRs appeared at a much higher frequency than observed in wheat (15.6 kb), barley (6.3 kb) (Kantety et al., 2002), maize (8.1 kb), soybean (7.4 kb), potato (11.1 kb), and cotton (20.0 kb) (Cardle et al., 2000). This means that SSRs are highly abundant in *A. mongolicus*.

In the present study, we found that the most common repeat units were mononucleotide repeats (A/T)n, which comprised 42.37% of the total SSRs in *A. mongolicus*. In many plant species such as onion, flax, apricot, and Para rubber tree, mononucleotide repeats were also the most abundant class of SSRs. However, in *Arabidopsis*, peanut, canola, and sugar beet, di-nucleotide repeats were the most frequent motifs (Kumpatla and Mukhopadhyay, 2005). In *Brassica napus* L., di- and tri-nucleotide repeat motifs were found to be the most abundant (Wang et al, 2015). Among the di-nucleotide repeats, (AG/CT)n (15.51%) was the most frequent motif identified in our study. (AAG/CTT)n (7.58%) was the dominant motif among tri- nucleotide repeats. However, the GC/GC motif is very rare in the *A. mongolicus* transcriptome, and appeared only three times in the present study. This finding is consistent with results reported for rice, maize, soybean, and wheat (Gao et al., 2003).

Putative functional annotation and categorization of unigenes containing SSRs in this study revealed that these unigenes are involved in various aspects of biological processes and pathways in *A. mongolicus*. Based on the results of the GO functional analysis, the majority of transcripts were assigned “cell” in the cellular component category, involved in “binding” in

the molecular function category, and involved in "metabolic activity" in the biological process in *A. mongolicus*. Similar results were reported in date palm (Zhao et al., 2012) and citrus (Palmieri et al., 2007).

In conclusion, *A. mongolicus* is an ecologically important plant species found in the mid-Asian desert, which shows strong tolerance to abiotic stress. However, information on SSRs in the *A. mongolicus* genome is insufficient. In this study, we identified 13,859 SSRs distributed in 10,409 unigenes in the *A. mongolicus* transcriptome. The SSRs identified and characterized in our study may provide a useful tool for research on genetic diversity, gene mapping, and marker-assisted selection in *A. mongolicus*. The functional categorization of *A. mongolicus* unigenes containing SSRs revealed that these ESTs represent many transcribed genes involved in different biologic processes and pathways.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the National Science Foundation of China (grant #91125028).

REFERENCES

Cardle L, Ramsay L, Milbourne D, Macaulay M, et al. (2000). Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156: 847-854.

Chen GQ, Huang HW, Kang M and Ge XJ (2007). Development and characterization of microsatellite markers for an endangered shrub, *Ammopiptanthus mongolicus* (Leguminosae) and cross-species amplification in *Ammopiptanthus nanus*. *Conserv. Genet.* 8: 1495-1497. http://dx.doi.org/10.1007/s10592-007-9306-2

Conesa A, Götz S, García-Gómez JM, Terol J, et al. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676. http://dx.doi.org/10.1093/bioinformatics/bti610

Gao LF, Tang JF, Li HW and Jia JZ (2003). Analysis of microsatellites in major crops assessed by computational and experimental approaches. *Mol. Breed.* 12: 245-261. http://dx.doi.org/10.1023/A:1026346121217

Garg R, Patel RK, Tyagi AK and Jain M (2011). *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res.* 18: 53-63. http://dx.doi.org/10.1093/dnares/dsq028

Ge XJ, Yu Y, Yuan YM, Huang HW, et al. (2005). Genetic diversity and geographic differentiation in endangered *Ammopiptanthus* (Leguminosae) populations in desert regions of northwest China as revealed by ISSR analysis. *Ann. Bot. (Lond.)* 95: 843-851. http://dx.doi.org/10.1093/aob/mci089

Gu LJ and Cheng HM (2014). Isolation, molecular cloning and characterization of a cold-responsive gene, AmDUF1517, from *Ammopiptanthus mongolicus*. *Plant Cell Tissue Organ Cult.* 117: 201-211. http://dx.doi.org/10.1007/s11240-014-0433-4

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

Kantety RV, La Rota M, Matthews DE and 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. http://dx.doi.org/10.1023/A:1014875206165

Kumpatla SP and Mukhopadhyay S (2005). Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome* 48: 985-998. http://dx.doi.org/10.1139/g05-060

Liu GH (1998). Study on the endangered reasons of *Ammopiptanthus mongolicus* in the desert of Alashan. *Bull. Bot. Res.* 3: 341-345.

Palmieri DA, Novelli VM, Bastianel M, Cristofani-Yaly M, et al. (2007). Frequency and distribution of microsatellites from ESTs of citrus. *Genet. Mol. Biol.* 30: 1009-1018. http://dx.doi.org/10.1590/S1415-47572007000500029

Scott KD, Eggler P, Seaton G, Rossetto M, et al. (2000). Analysis of SSRs derived from grape ESTs. *Theor. Appl. Genet.* 100: 723-726. http://dx.doi.org/10.1007/s001220051344

Sun J, Nie L, Sun G, Guo J, et al. (2013). Cloning and characterization of dehydrin gene from *Ammopiptanthus mongolicus*. *Mol. Biol. Rep.* 40: 2281-2291. http://dx.doi.org/10.1007/s11033-012-2291-7

Wang DJ, Yang CL, Dong L, Zhu JC, et al. (2015). Comparative transcriptome analyses of drought-resistant and -susceptible *Brassica napus* L. and development of EST-SSR markers by RNA-Seq. *J. Plant Biol.* 58: 259-269. http://dx.doi.org/10.1007/s12374-015-0113-x

Wang Z, Fang B, Chen J, Zhang X, et al. (2010). *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (*Ipomoea batatas*). *BMC Genomics* 11: 726-739. http://dx.doi.org/10.1186/1471-2164-11-726

Wei Q, Guo YJ, Cao HM and Kuai BK (2011a). Cloning and characterization of an AtNHX2-like Na+/H+ antiporter gene from *Ammopiptanthus mongolicus* (Leguminosae) and its ectopic expression enhanced drought and salt tolerance in *Arabidopsis thaliana*. *Plant Cell Tissue Organ Cult.* 105: 309-316. http://dx.doi.org/10.1007/s11240-010-9869-3

Wei W, Qi X, Wang L, Zhang Y, et al. (2011b). Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC Genomics* 12: 451-463. http://dx.doi.org/10.1186/1471-2164-12-451

Xu SJ, An LZ, Feng HY, Wang XL, et al. (2002). The seasonal effects of water stress on Ammopiptanthus mongolicus in a desert environment. *J. Arid Environ.* 51: 437-447. http://dx.doi.org/10.1006/jare.2001.0949

Zeng S, Xiao G, Guo J, Fei Z, et al. (2010). Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. *BMC Genomics* 11: 94-104. http://dx.doi.org/10.1186/1471-2164-11-94

Zhao Y, Williams R, Prakash CS and He G (2012). Identification and characterization of gene-based SSR markers in date palm (*Phoenix dactylifera* L.). *BMC Plant Biol.* 12: 237-244. http://dx.doi.org/10.1186/1471-2229-12-237

Zhou Q, Luo D, Ma L, Xie W, et al. (2016). Development and cross-species transferability of EST-SSR markers in Siberian wildrye (*Elymus sibiricus* L.) using Illumina sequencing. *Sci. Rep.* 6: 266-278. http://dx.doi.org/10.1038/srep20549

Zhou Y, Gao F, Liu R, Feng J, et al. (2012). *De novo* sequencing and analysis of root transcriptome using 454 pyrosequencing to discover putative genes associated with drought tolerance in *Ammopiptanthus mongolicus*. *BMC Genomics* 13: 20549-20558. http://dx.doi.org/10.1186/1471-2164-13-266